Penultimate draft

The original publication is available at www.springerlink.com

http://dx.doi.org/10.1007/s11023-009-9157-3

Where is cognitive science heading?

Paco Calvo Garzón[†], Ángel García Rodríguez

Departamento de Filosofía

Universidad de Murcia

Murcia, 30100 – Spain

[†]Corresponding author

Email: fjcalvo@um.es

Phone: +34 968363460

Fax: +34 968-36 39 67

Where is cognitive science heading?

William Ramsey (2007) claims that the move from classical to nonclassical cognitive science involves a shift, indeed a U-turn, in the status of cognitive science itself – namely, cognitive science is "moving away from representationalism" (235). This claim presupposes both that (i) the emergence of cognitivism, as a reaction to behaviourism, capitalized on the concept of representation, and that (ii) the materialization of nonclassical cognitive science since the 1990s involves a return to some form of pre-cognitivist behaviourism. As Ramsey puts it, after the cognitive revolution, a "revolution in reverse" (223) is now taking place. In this paper, we wish to contest Ramsey's analysis of where cognitive science is heading, and for that reason we will present a two-sided argument, to the effect that both (i) and (ii) may be called into question. But, first of all, let us take a closer look at Ramsey's argument in support of his views about current cognitive science.

Representation in cognitive explanation, according to Ramsey

Ramsey's thesis about the status of current cognitive science is a result of his views about the role of representations in cognitive explanations. In particular, Ramsey believes that there are different concepts of representation in contemporary cognitive explanations; and furthermore, that unlike the explanations offered by classical (computational) models of cognition, those offered by newer nonclassical models, such as connectionism and dynamicism, are not representational explanations at all (2007, xiii-xv).[1] Thus, his thesis about the status of cognitive science relies crucially on the

[1] Unless otherwise stated, page numbers refer to Ramsey's *Representation Reconsidered*.

(alleged) divide between the genuinely representational nature of classical cognitive explanations, and the nonrepresentational nature of nonclassical explanations. So, what are Ramsey's reasons for this divide?

Ramsey's reasons are directly related to the following claim: only classical cognitive explanations meet the *job description challenge*. This is, in Ramsey's own words, the challenge to provide "some sort of account of just how the structure's possession of intentional content is (in some way) relevant to what it does in the cognitive system. [...] We need, in other words, an account of how it actually *serves as* a representation in a physical system; of how it functions as a representation" (27). As a gloss, consider the familiar idea that cognitive science is founded on a hierarchy of three levels of analysis (computational, algorithmic and implementational), to the effect that the explanation of abilities characterized at the computational level must lie in states and processes found in the lower levels. Thus, in the context of providing a representational explanation for a given ability (at the computational level), the job description challenge is the challenge to show what it is for a system to employ representational explanatory structures at a lower level; more specifically, at the algorithmic level. Therefore, as stated, it is Ramsey's claim that only classical cognitive explanations meet the job description challenge, which in turn is due to the different notions of representation present in classical *versus* nonclassical accounts of cognition. So, what are these different concepts of representation?

As it is widely known, the main idea behind classical models of cognition is that cognitive tasks and abilities are explained in terms of symbol manipulation according to syntactic rules. But beyond this main idea, Ramsey distinguishes two notions of representation in classical models, in virtue of two different features of the explanations they offer. The first feature is that classical cognitive explanations are task-

decompositional, involving the division of a given cognitive task into further subtasks. There is here a notion of representation, for the inputs/outputs of the subtasks must represent aspects of the larger task, for them to count as subtasks of the larger task. Therefore, if cognitive explanations are task-decompositional, they must be representational. As the key lies with the inputs/outputs interior to the subtasks, Ramsey calls the representational posits involved interior input/output representations; or for short, *IO-representations* (72f). The second feature of classical explanations is that they involve models or simulations of the cognitive task or ability to be explained. The idea of a model is that of a structural isomorphism between the target domain and the explanation, which in turn leads to the idea that the constituents of the model stand in for aspects of the target domain. Therefore, if cognitive explanations involve models or simulations, they must be representational. Ramsey labels the notion of representation involved *S-representation* (78ff).

By contrast, nonclassical approaches to cognition, including connectionist and dynamicist models, reject the main idea behind classical models – namely, that cognitive explanations are a matter of symbol manipulation according to syntactic rules. However, nonclassical approaches, particularly connectionist explanations, are often thought to involve representations, despite the fact that they involve neither task-decomposition nor simulations. So, what notion of representation is at work here? According to Ramsey, there are in fact two such notions. First, it is argued that the internal "hidden" units in connectionist networks represent, in virtue of the fact that they respond to, detect, or are receptive to, different aspects of the environment. This involves a particular notion of representation, called *the receptor notion*, characterized in terms of a "causal or nomic dependency relation" to some external condition (123). Second, it is said that, in accordance with the distributed processing of connectionist

networks, the representational status of such networks cannot be a matter of one-to-one correspondences between items. Instead, it is thought that representations are tacitly distributed over the entire system, in so far as the system as a whole embodies a practical know-how pertaining to the cognitive task at hand; in other words, the system has the "potential to generate the right sort of processes and input-output mappings" (156). The result is that a different notion of representation, *the tacit notion*, is associated with connectionist explanations.

On the basis of this taxonomy of notions of representation, Ramsey claims that only classical explanations involve representations because, in his own words, "the explanatory utility of a representational posit ... depends on the way it actually allows us to understand how cognition involves the employment of an internal state that serves to stand for (or stand in for) something else" (221). In other words, given that explanations in terms of IO- and S-representations involve structures or processes standing (in) for something else, it follows that those explanations are genuinely representational. But for Ramsey, things are very different with nonclassical accounts of cognition, as it has not been shown that explanations in terms of the receptor and tacit notions are genuinely representational. Thus, as Ramsey notes regarding the receptor notion of representation, the fact that certain states "have the function of causing something to happen in certain conditions ... alone gives us no reason to treat them as representations"; and should be treated instead as "reliable causal mediators or perhaps relay switches" (126). Similarly with the tacit notion: as "it is far from clear just how the dispositional nature of a system's internal structures bestows upon them a representational function" (167), tacit representational states should be viewed as "nothing other than straightforward, non-representational dispositional states" (152). Hence, Ramsey's divide between the

genuinely representational nature of classical cognitive explanations *versus* the nonrepresentational nature of nonclassical explanations.

This article will take issue with Ramsey's claim that there is such a neat divide between classical and nonclassical cognitive explanations. To this end, two different strategies will be offered. On the one hand, by pursuing a beefing-up strategy, it will be argued that nonclassical (both connectionist and dynamicist) explanations can be treated as representational explanations, in exactly the same sense as classical explanations; in particular, they can be seen as exploiting the notion of structural isomorphism associated with S-representations. On the other hand, by pursuing this time a deflationary strategy, it will be claimed that explanations in terms of S-representations fail the job description challenge for the same reason that nonclassical accounts, like those involving the receptor notion of representation, do.

It is important to emphasize that the gist of our critical argument is only that if either strategy is successful, no difference between classical and nonclassical models of cognition ensues. In particular, we are not committed to defending one notion of representation, for instance S-representations, as the relevant notion for cognitive science; and neither do we take sides in the controversy whether classical or nonclassical models provide the correct architecture of the brain. Following Ramsey's lead, all that interest us here is the notion of representation and its place in the different kinds of cognitive explanations currently on offer. Bearing this in mind, let us consider the aforementioned strategies in turn.

The beefing-up strategy

As we've seen, Ramsey cashes out the model-based approach in terms of an alleged structural isomorphism between features of the model and the target domain; a form of

isomorphism that grants the representational import of the model. Unfortunately, connectionist networks appear to operate in a manner that does not grant it. As Ramsey points out, although hidden unit activations co-vary with the input vectors the network is fed with, such correlations are not structure-preserving. Courtesy of the learning algorithm for connection weight adjustment, neural networks tune to their environments inasmuch as hidden units acquire values of activation that causally mediate between incoming patterns of activation and output responses. We may therefore say that hidden vectors depend nomically upon the input vectors. Put bluntly, if connectionist networks are function approximators, hidden units serve the purpose of relay switches for the cognitive function to be approximated.

First of all, we wish to call into question this appraisal of connectionist theory. It is certainly true that you can read the dynamics of a neural network in receptor terms. However, it is not clear that the internal states that develop as a result of training fail to be structure-preserving. Learning algorithms, such as backpropagation, allow the system to adjust its connection weights in order to tune to the statistical regularities contained in the data pool. As a result of training, hidden space gets partitioned in order to reflect in a systematic manner the structure of the corpus. Thus, if the network is trained to distinguish mines from rocks (121), it will develop the representational resources to separate metrically hidden patterns of activation that stand for particular mines from those that stand for particular rocks. Of course, this is not new. It's the old story of how a connectionist network can approximate a non-linearly separable function by recoding the similarity relations that obtain in the input space in terms of a different set of relations that obtain at the hidden layer (as the well-known XOR problem has taught us). Statistical techniques such as cluster analysis allow us then to grasp what the network is doing. If the network successfully tells mines apart from rocks it is because

the hidden n-dimensional hyperspace has been partitioned by an (n-1)-hyperplane into two different subregions. These volumes are occupied by more and less prototypical mine and rock hidden vectors, and their metric relations statistically reflect the relations that obtain between real mines and rocks out there. Therefore, the conclusion that connectionist hidden states are subject to an S-representational reading seems to follow.

Nevertheless, in order to better appraise how the beefing-up strategy is meant to deliver the goods, we may pay closer attention to the way in which, according to Ramsey, S-representations are meant to work. Ramsey asks the reader (81) to imagine someone who is asked to infer genealogical relationships out of a large set of individuals that belong to the same family. A solution is to write down the names of the family members as labels, and connect these with arrows that designate specific relations such as "married to", "son/daughter of", etc. Once the diagram is completed, this diagram can be exploited in order to make inferences that relate to already known links or to familial relations that the subject may not even be aware of (*if Bob is the F of Jim, who is the G of Mary, then Bob is the H of Mary*). Put bluntly, inferential coherence is granted because the network of labels and arrows *models* the actual genealogical tree.

Interestingly enough, during the re-emergence of connectionism in the mid 80s, the genealogical example was analyzed as an illustration of the manifest generalization capacities of connectionist networks. Hinton (1986) considered a genealogical tree problem and showed how a network that models in hidden space, as a result of training, the familial tree structure, could infer unknown familial relations. According to Ramsey, the classicist familial diagram clearly meets the job description challenge because it "generates a symbolic model of the target domain" (82). Instead of symbols, Hinton's networks employ subsymbols, which may equally be thought of as S-

representations "by serving as parts of the model" (83). Certainly, subsymbols are finer-grained entities than symbols, and may be more difficult to keep track of, but that is precisely the job that cluster analysis and other statistical dimension-reducing techniques are meant to do. In this way, inferential coherence can be equally granted in the case of connectionist theory, not because the network induces the labels and arrows of the classicist diagram, but rather because hidden partitions, as identified statistically, *model* subsymbolically the genealogical tree.

Ramsey may nonetheless disagree with our understanding of connectionist modelling. In fact, he considers cluster analysis: "To draw the conclusion that a vector analysis reveals a networks's representational scheme, you have to first assume the hidden units are, in fact, serving as representations. The mere clustering in the response profile doesn't show this" (145). However, we believe that this reading of the situation misses the target by focusing upon the networks' hidden unit activations; a reading that tips the balance in favour of the receptor interpretation: "some sort of internal state reliably responds to, is caused by, or in some way nomically depends upon some external condition. When this occurs, such a state, whether it be called a "representation" or a "detector" … is viewed as having the role of representing that external condition because of this causal or nomic dependency relation" (123). In our view, this nomic-based reading of the sort of explanation that connectionism offers is the result of focusing too narrowly upon connectionist *components*, that is, the hidden units themselves (e.g., *hidden pattern-Bob/Jim/Mary*), rather than upon *architectural* features that have to do with the patterns of connectivity that obtain among the processing units. Let us elaborate.

Considering the above quotes, the question is whether we need to assume beforehand that hidden units play a representational role. Well, not necessarily. That

would be tantamount to claiming, in the case of the classical framework, that what counts for the purpose of meeting the job description challenge is that there is a one-to-one representational relationship to be found between individual internal posits and their respective referents in the environment, taken in isolation from the role the former play within the model. But, as the family tree problem exemplifies, the virtue of S-representations was precisely that what counted was rather how the set of relations among constituents in the model matches the set of relations among constituents in the target domain, and not the one-to-one correspondences among components belonging to the two different sets, taken in isolation. Similarly, in connectionist networks, it is the structural isomorphism itself between real world relations among family members, on the one hand, and hidden relations among hidden patterns of activation, on the other, what made the model S-representational.

By the same token, in the case of connectionism, it is not the clustering, understood as the statistical localization of hidden patterns of activation (components), what counts, but rather the recoding itself, courtesy of the architecture of the network. Putting it in slightly different terms, it is the capacity to abstract away from physical details of the input signal what allows the network to build up an abstract space where distances between vectors reflect environmental dependencies. In fact, an increasing level of abstractness can be obtained by inserting hierarchies of hidden layers. Ryder (2004), for example, exploits these architectural constraints in order to put forward a cortical network model of mental representation known as SINBAD ("Set of INteracting BAckpropagating Dendrites"). SINBAD is organized hierarchically in such a way that higher levels of abstractness can be computed as the layers distance their recodings from the metric relations of the sensory input space. The result is that hidden partitions end up tuning the statistically relevant dependencies contained in the data

pool. Crucially, as Ryder pinpoints, SINBAD cortical networks are structurally isomorphic to the environment they are trained in.[2]

Certainly, as Ramsey points out, the real issue boils down to what sort of theories get developed in cognitive neuroscience, connectionist cognitive modelling, and cognitive ethology (118), among other disciplines. And in this regard, the notion of a detector (the computational counterpart of the biological neuron) plays a central role in cognitive neuroscience. It further goes without saying that biological detectors and philosophical receptors fit hand in glove. But the reason why Ramsey reads connectionist pattern of activation in terms of receptors may be simpler. When he considers connectionism he has in mind three-layer feedforward neural networks of the sort that can categorize mines and rocks, or solve family tree problems. In this sort of networks cognitive activity is accounted for in terms of the vectorial transformation of input patterns of activation into output ones via hidden vectors. The receptor picture it offers in terms of one-to-one vectorial correspondences cries out. But as we already mentioned, this result is achieved at the expense of focusing upon components (neurons) instead of architectural constraints.

---

[2] Incidentally, Ramsey himself acknowledges that Ryder's notion of cortical representation is model-based (80). He also mentions Grush's (2004) *emulation theory* of mental representation as a second nomic deserter. However, as Ramsey mulls over the implications of a non-representational psychology, he observes that these model-based nonclassical theories are the exception rather than the rule (223). Another model that would serve equally to illustrate the beefing up strategy is O'Brien and Opie's (2004) *structuralist theory* of mental representation. Their model is cashed out in terms of second-order resemblances; a category that insofar as it comprehends all forms of structural isomorphism would receive Ramsey's beneplacit. Fortunately, the degree of popularity of a theory sheds little light on the alleged representational status of the explanations it offers, so we need not consider further how exceptional model-based connectionist theories happen to be.

In this respect, in line with modelling results in computational and cognitive neuroscience (Rolls and Treves, 1998), we may consider nonclassical architectures that depart in critical respects from the sort of straw-man connectionism that Ramsey appears to have in mind. By non-classical forms of connectionism, we mean the class of models that have different combinations of pattern associator/autoassociative memory/competitive network topologies, with bidirectional connectivity and inhibitory competition, and that employ combined Hebbian and activation-phase learning algorithms.[3] The idea is that by employing bidirectional and recurrent forms of connectivity these architectures differ from the feedforward picture in such a way that relational information can be encoded. Going back to Ryder's SINBAD model for the sake of illustration, we can see why these architectural features matter. As Ryder puts it: "The reason that a cortical SINBAD network develops into a dynamic isomorphism is that cells' inputs are not only sensory, but are also (in fact primarily) derived from within the cortical network. A cell's tuning is guided, in part, by these intracortical connections" (2004, p. 221). In this way, nonfeedforward connectivity allows the network's internal space to reflect the relational information among hidden patterns as activations are transformed and connection weights altered by taking into account internal connectivity. In our view, once the emphasis is laid upon layered connectivity, both across layers and within layers, we can see that the frameworks that cognitive neuroscience and related disciplines put forward crucially exploit model-based resources. Although the nomic-based resources present in detectors and three-layer feedforward networks are not dismissed out of hand, the moral to be drawn from Hinton's and Ryder's nets is that the representational status of such connectionist

[3] See Calvo Garzón (2003a), and the references therein.

networks lies in the isomorphic relations between the components of the modelling nets and the constituents of the target domain.

Finally, what about dynamicist theories of cognition? Well, if connectionist hidden states can get beefed-up into full-blown representational states for the system, the same goes for the internal states of dynamical models. The reason is simply that if connectionism can be understood in S-representational terms because of the emphasis laid upon architectural constraints, we have stronger reasons to believe that the same holds for dynamicism insofar as it departs from the static structure of three layer feedforward networks more radically than nonclassical connectionist networks. In this way, and for the same reasons offered above, dynamicist internal states are also statistically isomorphic to the environment in a way that does not boil down to one-to-one correspondences.[4]

Summing up, connectionist and dynamicist models of cognition have the resources to exploit the notion of structural isomorphism associated with S-representations. In so far as this is a feature common to both classical and nonclassical explanations of cognition, Ramsey's claim that only classical explanations are genuinely representational must be resisted. But this is not the only way to oppose

[4] It must be said that the borderline between connectionist and dynamicist models of cognition cannot be drawn easily (Spencer and Thelen, 2003). Ryder's SINBAD model is a dynamic system after all. As Ryder (2006) notes elsewhere: "The type of models the cortex is designed to build are dynamic models. The elements of a static model and the isomorphic structure it represents are constants... By contrast, in a dynamic model the elements in the isomorphic structures are variables. Rather than mirroring spatial structure, a dynamic model mirrors covariational structure" (pp. 125-6). However the divide is drawn between connectionist and dynamicist networks, it's clear that the beefing-up strategy works in the latter case, if it does in the former class of models.

Ramsey's characterization of the different nature of clasical *versus* nonclassical explanations, as will be seen next.

The deflationary strategy

As noted above, Ramsey's test to determine whether a given cognitive explanation is genuinely representational is the job description challenge. To repeat, this is the challenge to show what function a system's structures and states have for the system. More precisely, what this means is that it must be shown that a given structure plays a representational role, without invoking or presupposing a cognitive agent. This is related to the fact that the explanations proper to cognitive science must be agent-free explanations, if the different levels of analysis (computational, algorithmic and implementational) are not to be conflated. The underlying idea here could be related to the personal/subpersonal distinction, as follows: if the aim of cognitive science is to provide an explanation of a given computational task or ability at the lower levels, and the former are typical personal-level tasks or abilities, then the explanatory aims of cognitive science must involve positing states or structures at a different (subpersonal) level. Consider, in this light, Ramsey's notion of a mindless system.

When introducing the family tree problem as an illustration of S-representations in the classical framework, Ramsey considers the complaint that we may be after all presupposing some sort of agency that allows the system to infer cognitively the genealogical relations. Thus, someone may argue that discovering the fact that Bob is the H of Mary is the result of the personal level processing of the fact that Bob is the F of Jim and the fact that Jim is the G of Mary. However, Ramsey is careful enough to differentiate the personal inference-making abilities from the subpersonal syntactic crunching of the family tree relations. It is by focusing upon the latter that the job

description challenge is allegedly met. In fact, the structural isomorphism between the labels and arrows in the diagram, on the one hand, and the relations that obtain among family members, on the other, would allow a mindless system to remain competent in the task (85). Mindlessness thus illustrates how to avoid the conflation between personal and subpersonal levels. As Ramsey points out, "the property of being nomically dependent upon some other condition and the property of sharing a structural isomorphism with other things are both properties that are sufficiently natural to suggest the possibility of constructing an account of representation that is properly mindless". However, he continues, "… the mindless strategy does not work equally well for both of these representational notions" (193). From here, the claim that "the conditions that underlie S-representation successfully answer the job description challenge ... while the conditions that typically underlie receptor notion do not" (189) is a simple step away. But is it truly the case that the mindless strategy works better in the case of S-representations?

Ramsey offers a further example to illustrate and defend the claim that only mindless systems using S-representations, as opposed to those using the receptor notion, meet the job description challenge. Consider car A (194), which successfully moves through a curved segment of a walled track, because it has two rods sticking out from the corners of the front bumper, so that if a rod is pushed inwards as the car approaches one of the walls, a servomechanism is activated, turning the car away from the wall and through the curve. According to Ramsey, car A is a receptor system, for the states of the car that explain its navigational success causally co-vary with external conditions. However, there is nothing representational about the system: all that is required to explain the car's success are a set of internal structures and states acting as causal mechanisms or relay switches.

Compare now the receptor car with car B, which successfully negotiates the same curved segment of the track, because it has a rudder that fits into a groove along the track, to the effect that because the groove is shaped as the track itself, changes in the rudder as the car moves along the groove bring about changes in the car's steering wheel and front wheels. According to Ramsey, car B is a modelling or S-representational system, because the car's internal states exploit the fact that "an area of the groove functions as a 'stand-in' for a segment of the track [which] is just to say that an area of the groove is playing a representational role" (199).

For Ramsey, the point of the contrast between both cars is that viewing a system as representational (and the explanation of some of the system's successful behaviour as involving representational internal states) is independent of the presence of cognitive agents (recall that both cars are mindless or agent-free systems). In fact, all that is required, as in the aforementioned family tree example, is that a structure or state of the system stand in for the real world; for "[a] mindless system can still take advantage of the structural isomorphism between internal structures and the world, and in so doing, employ elements of those internal structures as representations-qua-stand-ins" (200). The point is not that components of the system *can* be viewed or described as standing in for elements of the real world; after all, they can also be described as parts of an internal causal mechanism. Rather, the point is that it is more natural, intuitive, beneficial (196), and less contrived (199), to view the components of the system as standing in for elements in the real world; hence, as performing a representational role. Let us consider this a little further.

Ramsey's reason for claiming that it is more natural to view car B as a representational system is that the groove where the rudder fits is a model, a sort of map, of the track itself (200). Of course, any map can be used by a minded being to

guide its behaviour. But given that Ramsey is interested in mindless systems, the crucial underlying thought must be that maps are inherently representational; i.e., they are representational independently of the presence of a cognitive agent. More precisely, the underlying thought must be that the very existence of a map, as used by a mindless system, is sufficient to explain a given cognitive success in representational terms. So, if as claimed car B employs a sort of map, then car B is a mindless but representational system.

The underlying thought here seems to be that maps, as used by a mindless system, are representational, independently of cognitive agents. This idea can be glossed by saying that maps fit, or not, directly, simply, or *naturally*; or perhaps, that there is a *natural* isomorphism between the map, as used by the mindless system, and the target domain. Hence, explanation of cognitive success is a matter of processes that happen naturally, independently of cognitive agents. But the problem with this gloss is that this notion of a natural fit or isomorphism cannot do all the work it is expected to do, and this for two interrelated reasons.

On the one hand, one could separate the notion of fit from the explanation of cognitive success. In this respect, consider how a map could be broken up into squares and put back together with the squares arranged in a different (one may say, wrong) order. This would be a sort of coded map; that is, a map which would only serve as a representational device for those in possession of the code, i.e. for those who know (perhaps because they have memorized it) the way in which the different squares should relate to each other. A coded map like this one would be an example of a map that does not fit (in the intended direct, simple or natural way). In other words, it would not be the case that the map is (again, in the intended sense) inherently representational; that is, representational independently of any cognitive agent. But the map could explain a

system's cognitive success: consider, for instance, how a secret agent might benefit from the use of such a map. Thus, the notion of fit can be separated from the explanation of cognitive success. The important point here is *not* that a contrast must be drawn between coded and uncoded maps – namely, that unlike uncoded maps, coded maps need supplementing with the relevant code in order to perform a representational role. The important point is that coded and uncoded maps alike require the idea of a cognitive user, as opposed to that of (direct, simple or natural) fit, to explain any cognitive (e.g., navigational) success.

On the other hand, the idea of a cognitive user must be different from, and external to, that of the mindless system itself, or a part of it. In this regard, consider how the advocate of the idea that there are natural representational processes, independently of cognitive agents, might try to accommodate the example of the previous paragraph, arguing that mindless systems that successfully use maps to navigate through the world must have an extra component (a part of the system) playing the role of the code. It could even be suggested that this is what we *normally* take a map to be. In other words, that it is *normal* maps (those including the code) that fit by themselves, or are inherently representational. But this will not do. For the notion of a normal map, or that of a normal fit between a map and the world, is very closely related to what *people* do as a result of learning to use a map to navigate through the world. Therefore, there could not be such representational devices as maps in the absence of an established practice of map reading into which people are initiated.

The point of this criticism is not that isomorphisms are cheap, something Ramsey can deal with by arguing that there is only one isomorphism actually employed by the mindless system in question (in our current example, car B). This reply still makes room for the idea that there is a natural fit between the structure in question, as used by the

mindless system, and the target domain, which is sufficient to explain a cognitive success without invoking a cognitive agent. Rather, the gist of this criticism is that talk of a natural fit, or isomorphism, independently of cognitive agents themselves, is an illusion. This is not to say that there are no such things as (natural) isomorphisms; it is only to say that (natural) isomorphisms are not intelligible in the absence of cognitive agents.

This criticism is in fact an adaptation of Wittgenstein's rejection of the idea of an intrinsic fit between a mental picture and its application, in the absence of a communal practice of users of the picture (see *PI* §§139ss). Wittgenstein's primary target is his own picture theory of mental representation as advocated in the *Tractatus*, but his argument can be applied *mutatis mutandis* to Ramsey's conception of S-representation. To sum up, if this Wittgenstein-inspired criticism works, it will have been shown that S-representations do not meet the job description challenge, for the notion of a cognitive agent must be invoked to characterize the idea of a structural isomorphism associated with S-representations. Assuming as Ramsey does both that only those cognitive explanations that employ a notion of representation that meets the job description challenge are genuinely representational, and that classical models of cognition are based on the notion of S-representation, it follows that classical explanations of cognition are not genuinely representational. Therefore, Ramsey's claim that only classical explanations of cognition are genuinely representational can be resisted; and hence, his characterization of the different status of classical *versus* nonclassical explanations of cognition.

So far, it has been argued that if either the beefing-up or the deflationary strategies work, Ramsey's characterization of the divide between classical and nonclassical models of cognition as a divide between genuinely representational *versus*

nonrepresentational explanations can be resisted. In the last two sections, this critique has been pressed in connection with the notions of S-representation and the receptor notion, as present in classical and nonclassical models respectively. And although the contrast between these two notions looms large in Ramsey's dialectic, these are only two of the notions of representation he admits to be currently on offer in cognitive science. According to Ramsey, classical models also employ IO-representations; and nonclassical models, the tacit notion of representation. So, can these two different notions of representation carry the weight of Ramsey's claim that only classical cognitive explanations are genuinely representational? To anticipate a little, in the following section it will be argued both that, as was the case with the receptor notion, explanations in terms of the tacit notion of representation can be beefed-up; and that, like S-representations, explanations in terms of IO-representations can be deflated. As a result, assuming Ramsey's full taxonomy of notions of representation in current cognitive science, his characterization of the divide between classical and nonclassical cognitive explanations will be resisted.

Tacit and IO-representations

Let us consider the tacit notion first. Nonclassical approaches, to remind the reader, may help themselves to tacit representations that are distributed over the entire system. Hinton's (1986) family tree network, for example, succeeds because it employs distributed patterns of activation superposed on the same weight matrix. Being fed with the patterns *Bob* and *H of*, the network will output *Mary*. This, let's assume, correct response will be triggered even when the network has not been trained on that relationship, as long as it has been trained on others (*Bob is the F of Jim, who is the G of Mary*) that allow the network to infer the H-of link. In this way, the network's

generalization capabilities are explained by appealing to the encoding in the weight matrix of a practical know-how that pertains in our example to the cognitive task of uncovering genealogical relations in the family tree. Hinton's network is thus *disposed* to output correct responses tacitly.

Unfortunately, Ramsey questions the notion of superpositionality as what underlies our understanding of tacit representations. Once it is called into question, we are left with (mere) superposed dispositions, which are representationless. As we saw earlier, tacit representations reduce to nonrepresentational dispositional states of the sort that can be found in many reactive systems. Ramsey spells out in detail the reasons why he thinks that the dispositional properties of a neural network are not representational. For the purposes of this section, however, we shall consider a more specific aspect of Ramsey's criticism – namely, his questioning the very conceivability of superposed representations in connectionist theory. As he observes, it is "far from a given that the idea of superposed representations is even fully intelligible" (178). His concern is "how it is that truly distinct, specifiable items of content can be simultaneously represented by the *exact same* … computational conditions" (178).

Ramsey is asking how different contentful states can be represented by means of the same computational conditions. But we don't think it is obvious that the "*exact same* … computational conditions" are present across items of content. In our view that depends on how the vehicles of content are individuated in connectionist networks. Ramsey considers the weight matrix as a whole as the vehicle of content, and that is why he targets superpositionality. But we must drop the idea that the only kind of description available to the connectionist modeller is the one at the level of the weights matrix, as superposition, under Ramsey's lens, implies. How can we then individuate the vehicles of content in connectionist networks?

Although we may help ourselves to a number of candidates in the connectionist literature as the vehicles of content, for present purposes we shall consider what O'Brien and Opie (2006) call *fan-ins*.[5] A fan-in is the specific part of the overall weight matrix that transforms input patterns of activation for each single hidden layer unit. In O'Brien and Opie's words, and considering the sort of simple feedforward architectures that Ramsey has in mind, a fan-in "is the vector of weights modulating the effect of incoming activity on a particular hidden unit. Within any feedforward network there is one fan-in per hidden unit, each corresponding to a row of the network's hidden layer weight matrix" (2006, p. 38). But notice that it is only by taking the whole weight matrix as the vehicle of content that Ramsey's worry that "truly distinct, specifiable items of content can be simultaneously represented by the *exact same* … computational conditions" makes sense. The same computational conditions are those that we identify under the modulation of activation patterns that the weight matrix produces throughout the whole network. However, by individuating the vehicles of content as fan-ins the worry that tells against a superpositional concept of representation disappears. The reason is that different fan-ins correspond to different computational conditions, as the weight matrix gets divided up into rows.

Granting thus that superposition can be conceived of, why are fan-ins to be interpreted representationally? Fortunately, in order to answer this question, once we have a harmless form of superposition, we can rescue the beefing-up strategy already

[5] Other options include the well known microfeatural descriptions of Churchland (1989), and the more recent clustering approach of Shea (2007). Churchland's microfeatural rendering of the vehicles of content is fleshed out in terms of hidden patterns of activation, and therefore takes us back to the receptor notion of representation (for a criticism of Churchland's connectionist semantics, see Calvo Garzón, 2003b). On the other hand, although Shea's clustering approach escapes the receptor notion, and may be a more promising candidate in that sense, it is beyond the scope of this paper to analyze it in detail.

deployed in the case of the receptor notion. As O'Brien and Opie point out, "if we are to discover any structural resemblance between a network's connection weights and its task domain it is the fan-ins on which we should focus… Given the crucial role of fan-ins in network processing, we offer the following proposal: the fan-ins in the hidden layer of a successful connectionist network structurally resemble aspects of the network's task domain." (2006, pp. 37-8). The reader can see that the bearing of O'Brien and Opie's treatment of fan-ins is straightforward upon our current concerns. By typing fan-ins as the vehicles of content, we can apply to the case of superposition the structural isomorphic way of beefing-up the receptor notion of representation in connectionist networks, bypassing thus Ramsey's worries. In short, fan-ins don't superpose in Ramsey's sense, and are thus not subject to the dispositions-based criticism.

Finally, what about IO-representations? Ramsey chooses multiplication to illustrate IO-representations. We may define multiplication as the "task of transforming numerals of one sort (those standing for multiplicands) into numerals of another sort (those standing for products)" (69). In this way, a function that, computationally speaking, is characterized in terms of abstract entities such as numbers and products, gets algorithmically fleshed out in terms of numerals that the system can manipulate, and that stand for numbers and products.

It is noteworthy that the case of multiplication can be cashed out in S-representational terms. As Ramsey points out, "many would claim that *all* numerical computations are simulations of various mathematical functions. If this is true, then the IO notion could be reduced to a special type of S-representation for these types of computational operations" (104). But once IO-representations reduce to S-representations, the reader will have spotted an obvious opportunity for the sceptic.

Namely, to deflate IO-representations once again: If S-representations do deflate for the map-related reasons adduced in the previous section, and IO-representations, following Ramsey, reduce to S-representations, then IO-representations may get deflated, so to speak, for free. Nevertheless, as we shall see next, even if IO-representations do not reduce to S-representations, there are reasons to believe that the job description challenge still fails to be met.

What is it precisely about numerals that allow the system to meet the job description challenge? Ramsey notes that cognitive explanations are task-decompositional in such a way that inputs/outputs of the subtasks that obtain as a result of decomposition represent aspects of the larger task. In our example, since multiplication decomposes into (a series of) additions, the summands represent aspects of multiplication insofar as they are input/outputs of the addition subtask that the specific multiplicands involved in the larger task determine. As Ramsey notes, "many of the tasks performed by the inner sub-systems should be seen as natural 'parts' of the main computations that form the overall explanandum … Our ability to do multiplications … might be explained by appealing to a sub-process that repeatedly adds a number to itself" (72). So, the notion of task-decomposition associated with IO-representations crucially involves that of a natural task/subtask relation. However, this notion is not without problems.

Ramsey chooses a highly disembodied task to illustrate IO-representations, but there is an initial worry that more ecological tasks cannot be decomposed into such neat "natural parts", as apparently exemplified by the relation between addition and multiplication. After all, what would be the "natural parts" of more embodied tasks, such as face recognition, that would form Ramsey's "overall explanandum"? Assuming that face recognition can be decomposed into subtasks representing aspects of the larger

task, the problem is that the subtasks involved in face recognition may vary widely across agents and even for the same agent on different episodes of face recognition. It is then difficult to make sense of the equivalent of the adders in highly idiosyncratic embodied tasks. Specific subtasks will receive incoming inputs and will operate upon them. Likewise, specific subtasks will deliver outputs that may be used by other subtasks and that will result ultimately in the output performance of the overall face recognition system. However, the equivalent of the addition/multiplication "natural" relation that allows for the exploitation of those inputs and outputs internally is lost to sight.

But the problem with the notion a natural task/subtask relation is not just that there is no clear sense in which the notion can be transferred from highly disembodied to more ecological tasks. If this was all that is wrong, the very idea of a natural task/subtask relation would be left unscathed (albeit one that could not be easily applied to ecological cognitive abilities). And as a result, that very idea could be appealed to in support of the claim that mindless systems whose cognitive abilities are explained in task-decompositional terms are indeed representational systems, and the cognitive explanations involved genuinely representational. In other words, it could be claimed that a task-decompositional explanation involves explaining a personal level task in terms of subtasks, where aspects of the latter stand for aspects of the former *naturally*, that is, independently of the presence of a cognitive agent. Thus, it could be concluded that in so far as the representational status of the explanation does not invoke or presuppose cognitive agents, the job description challenge is met.

However, as with the notion of fit explored above, the notion of a natural task/subtask relation cannot do all this work. Consider again Ramsey's own example, where addition is a "natural part" (or subtask) of multiplication. There is no problem

with the idea that multiplication involves a series of additions. The problem lies with the idea that the relation in question is "natural" – that is, that addition and multiplication are so related independently of any cognitive agents. For in so far as addition and multiplication are the mathematical operations they are because of what *people* do, the relation between them is not independent of any cognitive agents. Only abstracting from this, could one say that addition is (in the intended sense) a natural subtask of multiplication. But if one did so abstract, one would be assuming, rather than showing, that the job description challenge is met. So, like S-representational explanations, IO-representational explanations can also be deflated.

Where is cognitive science heading?

The previous sections have presented a sustained two-sided argument against Ramsey's thesis that only classical accounts of cognition, as opposed to nonclassical (connectionist and dynamicist) accounts, are genuinely representational: on the one hand, connectionist and dynamicist accounts have the resources to exploit the crucial notion of a structural isomorphism, like classical accounts (the beefing-up strategy); on the other hand, IO- and S-representations refer to a cognitive agent, and therefore do not meet the job description challenge (the deflationary strategy). Both sides of this argument work independently of each other; and success with either of them will mean trouble for Ramsey's neat divide between classical and representational, *versus* nonclassical and nonrepresentational, models of cognition.

However, this two-sided argument appears easily to run into serious difficulties. First, if nonclassical explanations in terms of receptors or dispositional properties are beefed-up, then all sorts of systems, including thermostats (136), faucets (144) or rubber bands (178), to name a few examples considered by Ramsey, will have representational

properties; hence panrepresentationalism follows. Second, if classical explanations are deflated, and no (subpersonal) algorithmic system turns out to be genuinely representational, then fictionalism about computational processes ensues, either in Dennett's sense (i.e., we are just applying the intentional stance to real computational processes) or in Searle's (i.e., we are just treating the system's processes as computational, though they really are not). But these are unwelcome consequences, in as much as both panrepresentationalism and fictionalism are "dubious" (28) and "counter-intuitive" (102). Therefore, the difference between classical and nonclassical models should be upheld.[6]

But perhaps things are not as bad as they seem. For regarding the risk of panrepresentationalism, what the beefing-up strategy entails is not that all sorts of things have representational properties; but rather that certain types of connectionist and dynamical systems, perhaps amongst them the human brain, have the resources to exploit the notion of structural isomorphism associated with classical models of cognition. Therefore, the beefing-up strategy, if successful, is limited in scope, for it does not apply automatically to all systems involving receptors or dispositional properties. Hence, the risk of panrepresentationalism is diffused.

Regarding the risk of fictionalism, what the deflationary strategy means, if successful, is that there are no inherently representational algorithmic processes (i.e., representational in the absence of a cognitive agent). In so far as cognitive science is conceived of as a reductive project, where personal level abilities are explained in terms of what happens at a different subpersonal level, it follows from the deflationary

[6] A word of caution is needed here, regarding the dialectics of this paragraph, for Ramsey explicitly considers the dangers of panrepresentationalism to motivate his claim that nonclassical explanations in terms of receptors and dispositions are not genuinely representational. However, his rejection of fictionalism is not meant to directly support the claim that classical explanations in terms of IO- or S-representations are genuinely representational, but only to respond to a possible objection (98ff). Thanks to Bill Ramsey for bringing this point to our attention.

strategy that there is no room for representations within cognitive science. This may look bleak, but it is important to note that the deflationary strategy does not mean the elimination of representational explanations altogether (as personal level explanations), but only the elimination of representational explanations from such reductive undertakings as cognitive science. Now, if this is all that follows from the deflationary strategy, then fictionalism stops being a worrying issue. To begin with, the deflationary strategy must be distinguished from the thought, typically associated with Dennett's fictionalism, that the intentional stance is nothing more than a useful device in order to explain certain cognitive abilities; for the deflationary strategy does not deny that people, as opposed to subpersonal systems, have representational properties. In addition, it must also be distinguished from Searle's claim that computational processes are not real processes; for all the deflationary strategy entails is that processes at the algorithmic level are not inherently representational, not that there are no such processes. In a nutshell, perhaps there are good reasons for moving towards fictionalism in either Dennett's or Searle's senses (perhaps not), but the deflationary strategy does not itself promote such a move.

Summing up, then, the two-sided argument against Ramsey's neat divide between classical and representational, *versus* nonclassical and nonrepresentational, accounts of cognitive explanations can be defended from the double risk of panrepresentationalism and fictionalism. What this means is that if connectionist and dynamicist explanations can be beefed-up, then cognitive science is not on a fast road to a "dubious" panrepresentationalism; and if task-decompositional, model-based explanations can be deflated, then cognitive science is not flirting with a "counter-intuitive" form of fictionalism.

Where then is cognitive science heading? According to Ramsey, insofar as nonclassical models of cognition gain momentum, cognitive science research is "quietly and unwittingly moving away from representationalism" (235). This nonrepresentational direction presupposes, first, that (i) the emergence of cognitivism in the 1950s, as a reaction to behaviourism, capitalizes on the concept of representation, and second, that (ii) the materialization of nonclassical cognitive science since the 1990s involves a return to some form of neo-behaviourism, insofar as the receptor and the tacit notions of representation fail to meet the job description challenge. In this way, after the cognitive revolution, we seem to have a "revolution in reverse" (223). However, this need not be the case since, if the line of argument put forward in this paper is correct, both (i) and (ii) may be called into question. On the one hand, *contra* (i), if the deflationary strategy works, cognitivism, as classically understood, would fail to capitalize, after all, on the concept of representation. And the reason is that IO-representations and S-representations would fail to meet the job description challenge for the reasons offered earlier. On the other hand, *contra* (ii), in case the beefing-up strategy is sound, the return to a pre-cognitivist era cancels out, since nonclassical models of cognition would posit representational states that would meet the job description challenge. In this way, either cognitive science is still representational, and therefore cognitivist, or a cognitive revolution, properly speaking, never took place.

There is a further (third) possible scenario for cognitive science, if both strategies join forces to show that nonclassical and classical accounts face the job description challenge in equal conditions (by the beefing-up strategy), but equally fail to meet it (by the deflationary strategy). In this scenario, cognitive science does not offer representational explanations (in the relevant sense), as the advocate of beefed-up connectionism and dynamicism suggests. Hence, anti-representationalism in cognitive

science would ensue. But for the main objective of this paper – that is, to contest Ramsey's claim that cognitive science is in the middle of a U-turn – it is sufficient to consider the consequences of either strategy on its own; hence, for current purposes, we remain neutral on the dispute between representationalists and anti-representationalists in cognitive science.

Acknowledgements

We are grateful to Bill Ramsey for helpful comments and suggestions on a previous version of this paper. Preparation of the manuscript was supported by DGICYT Project HUM2006-11603-C02-01 (Spanish Ministry of Science and Education and Feder Funds).

References

Calvo Garzón, F. (2003a). Non-classical connectionism should enter the decathlon. *Behavioral and Brain Sciences*, 26, 603-604.

Calvo Garzón, F. (2003b). Connectionist semantics and the collateral information challenge. *Mind & Language*, 18, 77-94.

Churchland, P.M. (1989). *A Neurocomputational perspective: The nature of mind and the structure of science*. Cambridge, Mass.: MIT Press.

Grush, R. (2004). The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27, 377-442.

Hinton, G. (1986). *Learning distributed representations of concepts*. In Proceedings of the 8th Annual Conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.

O'Brien, G. & Opie, J. (2004). Notes towards a structuralist theory of mental representation. In H.Clapin, P.Staines & P.Slezak (eds.) *Representation in mind: New approaches to mental representation*. Elsevier.

O'Brien, G. & Opie, J. (2006). How do Connectionist Networks Compute? *Cognitive Processing*, 7(1), 30-41.

Ramsey, W. (2007). *Representation reconsidered*. New York: Cambridge University Press.

Rolls, E.T. & Treves, A. (1998). *Neural Networks and Brain Function*. Oxford: Oxford University Press.

Ryder, D. (2004). SINBAD neurosemantics: A theory of mental representation. *Mind & Language*, 19(2), 211-240.

Ryder, D. (2006). On thinking of kinds: a neuroscientific perspective. In Graham Macdonald and David Papineau (eds.) *Teleosemantics* (pp. 115-145). Oxford: Oxford University Press.

Shea, N. (2007). Content and its vehicles in connectionist systems. *Mind & Language*, 22, 246-269.

Spencer, J.P. & Thelen, E. (Eds.) (2003). [Special issue]. Connectionist and Dynamic Systems Approaches to Development. *Developmental Science*, 4(4).

Wittgenstein, L. (2001). *Philosophical Investigations*. Oxford: Blackwell (Third edition; first edition, 1953. Translated by G.E.M. Anscombe.)