

Emma Borg

## *If Mirror Neurons are the Answer, What was the Question?*

**Abstract:** *Mirror neurons are neurons which fire in two distinct conditions: (i) when an agent performs a specific action, like a precision grasp of an object using fingers, and (ii) when an agent observes that action performed by another. Some theorists have suggested that the existence of such neurons may lend support to the simulation approach to mindreading (e.g. Gallese and Goldman, 1998, 'Mirror neurons and the simulation theory of mind reading'). In this note I critically examine this suggestion, in both its original and a revised form (due to Iacoboni et al., 2005, 'Grasping the intentions of others with one's own mirror neuron system'), and argue that the existence of mirror neurons can in fact tell us very little about how intentional attribution actually proceeds.*

Recent neurological studies have revealed the existence of two different kinds of neuron, so-called 'mirror neurons' and 'canonical neurons', in certain regions of monkey and human brains. These neurons apparently play a dual role in the brain-based life of their hosts. Both kinds of neuron are active in the production of motor actions by the agent (with different patterns of firing producing different motor actions, such as precision grasping with fingers versus more brute grasping with the whole hand or the mouth), but, surprisingly, both sets of neurons also fire in certain other conditions. Thus mirror neurons are active in both conditions (i) and (ii):

Correspondence:

Prof. E. Borg, Philosophy Dept., University of Reading, Reading RG6 6AA, UK.  
e.g.n.borg@reading.ac.uk

- (i) the production of a specific motor action (e.g. grasping with fingers) by an agent
- (ii) the observation of a conspecific performing the motor action in (i).<sup>1</sup>

Canonical neurons, on the other hand, are active in both condition (i) and condition (iii):

- (iii) the observation of an object which provides affordance for the motor action in (i), when that object is not being acted on by a conspecific.

When mirror neurons (MNs) and canonical neurons (CNs) fire in conditions (ii) and (iii) their activity is somehow taken ‘off line’. That is to say, their firing does not result in any observable motor action. Whether this is due merely to the absence of the firing of the complete pattern responsible for motor action (i.e. the fact that MNs and CNs do not fire simultaneously in these conditions) or whether it is due to some subsequent inhibition mechanism is currently unclear.

This experimental finding that our brains are, to some extent at least, ‘doing the same thing’ when an agent  $\phi$ -s and when an agent witnesses another person  $\phi$ -ing has been taken by some theorists to lend support to the simulation theory of intentional attribution.<sup>2</sup> As is well-known, simulation theory stands in opposition to the more traditional theory-theory approach to mindreading. According to the latter account, grasping the mental states of others is a matter of applying one’s theory of common-sense, belief-desire psychology to that other person. We work out what someone will think or do via subsuming them under very general intentional laws, such as ‘if A wants p and believes that doing q will bring about p, then *ceteris paribus*, A will do q’. The theory-theory approach is, then, very different to the picture

[1] Here the range of ‘conspecific’ is somewhat vague: MNs fire when a monkey witnesses another monkey perform the action in question but also when they see humans perform the same act. MNs do not fire when the subject sees very different entities, like machines, perform the same physically described action. There is also some evidence that what an organism counts as performance of the action is relatively fluid, with one experiment from the Rizzolatti lab showing that when an action is performed using an intermediary object (e.g. an experimenter opening a peanut using tweezers) this initially causes no MN firing, but that after repeated exposure to the action, appropriate MN firing can be induced (discussed in Arbib *et al.*, 2005, p. 243).

[2] Gallese writes: ‘Whenever we are looking at someone performing an action, beside the activation of various visual areas, there is a concurrent activation of the motor circuits that are recruited when we ourselves perform that action. Although we do not overtly reproduce the observed action, nevertheless our motor system becomes active as if we were executing that very same action that we are observing. To spell it out in different words, action observation implies *action simulation*’ (Gallese, 2001, p. 37).

proffered by simulation theory, which we might (somewhat crudely and certainly with objections from some advocates of simulation) characterise as the method of assigning mental states to others via ‘putting oneself in their shoes’. According to simulation theory, understanding another’s mental states is a matter of using one’s own intentional mechanisms in a piece of ‘pretence’ reasoning to determine what state I would be in if I were in the other’s situation and were armed with their beliefs and desires, etc. This reasoning process must, of course, be ‘off line’ to prevent its outputs resulting in action: when I use my own intentional system to work out that you are rolling up into a ball because you believe you are being attacked by a bear it’s not going to be helpful if that entails me rolling up into a ball as well.

Clearly, then, this idea of utilising one’s own intentional systems in an off-line capacity when making an intentional attribution to another is very much in keeping with the apparent behaviour of MNs outlined above.<sup>3</sup> Patterns of MN stimulation occur both in producing a given act in an agent and in witnessing that action in another, with the difference in output being explained by the ‘off line’ nature of the firing in the latter case. Hence the idea that the discovery of MNs can lend support to simulation theory by providing a neurological basis for it — I’ll label this claim *the MN hypothesis*. However, in this note, I’d like to ask exactly how much support the discovery of MNs can really lend to any account of intentional attribution. The structure of the note is as follows: in the next section I’ll investigate the MN hypothesis in a bit more detail. As we will see, the claim as initially made seems to face an obvious worry. However, we will then turn (in §2) to a revised version of the hypothesis, which avoids the initial problem. I will suggest, however, that the revised version faces problems of its own and should also be rejected. I’ll thus conclude that the discovery of MNs does not support any particular theory of intentional attribution, and close with what I take to be the general message of this finding for cognitive science.

## 1. The MN Hypothesis

The idea that the discovery of MNs can be used to support simulation theory was (as far as I am aware) originally set out by Vittorio Gallese and Alvin Goldman in ‘Mirror neurons and the simulation theory of mind reading’. As they note a key idea in simulation theory is that the

---

[3] As Gallese puts it: ‘MN activity seems to be nature’s way of getting the observer into the same ‘mental shoes’ as the target, which is exactly what the conjectured simulation heuristic is all about’ (Gallese, 2000, p. 4).

agent engaging in intentional attribution tries to mimic or mirror the target of her attribution. They then point to MNs as providing initial evidence of such mimicry in the mindreading process. Furthermore, simulation theory allows not only the prediction of mental states and action for another, it also provides for the retrodiction of mental states as well, allowing the agent to address the question ‘what goals did the target have that led them to  $\phi$ ?’ They then propose:

it is conceivable that externally-generated MN activity serves the purpose of ‘retrodicting’ the target’s mental state, moving backwards from the observed action. Let us interpret internally generated activation in MNs as constituting a plan to execute a certain action, for example, the action of holding a certain object, grasping it or manipulating it. When the same MNs are externally activated — by observing a target agent execute the same action — MN activation still constitutes a plan to execute this action. But in the latter case the subject of MN activity knows (visually) that the observed target is concurrently performing this very action. So we assume that he ‘tags’ the plan in question as belonging to the target (Gallese and Goldman, 1998, p. 497).

The thought is that MN stimulation provides the neurological basis of plan formation. Thus where we see action by a conspecific we cannot help but view it as goal directed, intentional behaviour because merely witnessing the behaviour triggers the formation of the plan or intention (the pattern of MN stimulation) that would have accompanied the behaviour if we had intentionally engaged in it ourselves.

Notice, however, that even if this is the right way to view the behaviour of MNs, it seems to give us only an extremely limited picture of intentional attribution. For MN firing is triggered by gestures (like hand grasping) and thus, even if the current account is right, the dual function of MNs merely allows us to see similar gestures in conspecifics as intentional actions per se, it does not help us to determine which overarching intention caused that motor action. That is to say, when I see you grasp something it may be that (due to the firing of mirror neurons) I can’t help but see you as performing a self-motivated action (‘you intended to grasp’), but standardly this is not the notion of intention which is really at issue between theory-theory and simulation theory. In mindreading proper I want to be able to work out that you picked up the cup because you were thirsty, not merely that you grasped the cup because you intended to grasp the cup.<sup>4</sup> It seems

---

[4] Rizzolatti *et al.* (2001) are careful to respect this point when discussing experiments where the goal of a monkey’s action (e.g. a piece of food) is occluded from an observer monkey behind a screen: ‘It is important to stress that we are not claiming that, in the experiment of Umiltà *et al.*, the monkeys understood the intention of the agent (that is, why the observed

that we should hold apart here the recognition of an action as intentional (a self-motivated action performed by a creature with a mind like mine) and the recognition of an action as performed with a specific intention (e.g. picking up the cup with the intention of drinking).<sup>5</sup> At best, it seems MNs might help with the former task, with the firing of mirror neurons perhaps helping to determine the scope of the predicate ‘intentional agent’ for a given individual, yet it seems they cannot contribute to the latter task.<sup>6</sup>

Now Gallese and Goldman are careful not to overstate their case, writing:

It should be emphasized here that the hypothesis being advanced is not that MNs themselves constitute a full-scale realization of the simulation heuristic ... Our conjecture is only that MNs represent a primitive version, or possibly a precursor in phylogeny, of a simulation heuristic that might underlie mindreading (Gallese and Goldman, 1998, p. 498).

This aspect of their account has led Gallagher 2006 to characterise Gallese and Goldman’s view as a *hybrid* version of simulation theory: it appeals both to sub-personal, implicit simulation (in the form of MN activity) *and* to personal-level, explicit simulation (to achieve a full realisation of the mindreading system).<sup>7</sup> However, one worry with this kind of hybrid approach concerns the extent to which the existence of implicit simulation could lend support to the claim that explicit simulation is necessary in a full account of mindreading. The problem is that, even if there were the kind of precursor system Gallese and Goldman envisage (one which recognised a movement as an intentional movement via a form of implicit simulation), it would remain an

---

action was performed), but only that they understood the action meaning (that is, what the agent did)’ (p. 667). However, although Rizzolatti *et al.* do stress the role of mirror neurons in what I above call ‘intentional explanation per se’, still the authors conclude that ‘the mirror system could underlie other fundamental cognitive functions ... such as language understanding and mind reading. Although we still lack a satisfactory comprehension of those higher capacities, and the precise role of the mirror system in these functions remains unknown, we think that the mirror system offers a new and very promising heuristic tool for their empirical investigation’ (p. 669). It is this latter claim, a variety of the MN hypothesis, which is at issue here.

- [5] These two notions seem to be run together when Gallese writes: ‘We can immediately tell whether a given observed act or behaviour is the result of a purposeful attitude or rather the unpredicted consequence of some accidental event, totally unrelated to the agent’s will. In other words, we are able to understand the behaviour of others in terms of their mental states. I will designate this ability as *mind-reading*’ (Gallese, 2001, p. 33).
- [6] Note even here though that experimental evidence seems to show MN firing could be at best a sufficient, not a necessary, condition for seeing an act as intentional, since some stimuli for which one would not predict MN firing nevertheless seem to lead to intentional explanation. See the work of Heider and Simmel 1944, Gergely *et al.* 1995.
- [7] The hybrid view also seems to be endorsed in Williams *et al.* (2001).

entirely open question what form a full-blown system of intentional attribution might take. Specifically, while MNs might provide a precursor to a simulation mechanism for mindreading proper, they could just as well provide a precursor for a theory-theory mechanism, for if all they do is help determine which actions in the world count as intentional this still leaves it entirely open how one proceeds to attribute specific intentions to those actions.<sup>8</sup>

## 2. The Revised MN Hypothesis

This objection to the original, hybrid version of the MN hypothesis can, however, be avoided by stronger conceptions of the relationship between MN activity and simulation theory. Specifically, we could take the view that simulation should be understood in an entirely implicit, sub-personal way, maintaining that the behaviour of MNs is sufficient to show how full-blown intentional attribution takes place. On this view, the MN system does not provide a precursor to some other kind of mindreading system but itself constitutes the full process of intentional attribution. Hence, to the extent that MN activity can be seen as a form of simulation, the simulation theory of mindreading would be vindicated.<sup>9</sup>

Now, one obvious worry with this sort of move would seem to be the point raised above that MN firing can get us as far as seeing an action as intentional (i.e. self-motivated) but it cannot help in getting us to see an action as the result of an overarching intention (e.g. to drink). However, in an interesting and ambitious paper, Marcus Jacoboni, Vittorio Gallese and others, have tried to counter this worry. In ‘Grasping the intentions of others with one’s own mirror neuron system’ (Jacoboni *et al.*, 2005) the authors point to a more detailed analysis of the firing patterns of MNs to suggest that, in their words, ‘MNs may encode not just the what but the why of an action’. In a nice

---

[8] One reader for this journal objected that the above argument requires the premise that while implicit or sub-personal simulation may suffice for an observer to represent a target’s act as intentional or goal-directed, only explicit or personal usages of simulation could enable an observer to represent the specific content of a target’s intention. Yet this premise itself is open to question. However the argument above does not, I think, actually rely on this premise. As far as the argument above goes it could turn out that implicit or sub-personal simulation *is* sufficient for mindreading proper. The point is rather that implicit or sub-personal simulation *characterised as MN activity* cannot be sufficient for mindreading proper. At best *this kind of* implicit simulation could be sufficient for attribution of intentional status per se. This is just to reiterate that, as it stands, MN activity provides no more (or less) support for some kind of simulation theory of full-blown mindreading than for a theory-theory account.

[9] See Gallagher (2006) for objections to the view that MN activity should be thought of as a form of simulation at all.

set of experiments they recorded the level of neuronal activity in three different conditions:

- (1) **Background condition:** participants were exposed to one of two scenes, each composed of the same set of objects, differently arranged. One scene depicted a teapot and cups, etc, arranged to suggest 'before tea', while the other depicted the same objects differently arranged to suggest 'after tea'.
- (2) **Action condition:** participants were exposed to an action performed on an object without any context. In this action scene all participants saw was a hand grasping a cup, either using a precision finger grasp or using a whole hand pick up.
- (3) **Intention condition:** participants were exposed to a combination of scenes (1) and (2). That is to say, they saw one of the two actions (either a finger grasp or a whole hand pick up of a cup) set against one of the two background conditions (either a before tea or an after tea background).

The results showed a large increase in levels of MN stimulation in condition (3) as against condition (2). Furthermore, this additional stimulation could not be explained merely by noting the increased complexity of the scene in (3) (i.e. crudely, appealing to the idea that (3) is the result of 'adding' scenes (1) and (2)). It could not be explained in this way because the levels of stimulation differed depending on which of the background conditions was used: there was greater stimulation when viewing either kind of grasping in the before tea, 'drinking', context than in the after tea, 'clearing up', context. This was the case even though both scenes involved the same actions and the same set of objects, differing only in the arrangement of objects (and this difference in arrangement of objects caused no parallel increase in firing in condition (1) alone). Iacoboni et al conclude:

The data ... suggest that the role of the mirror neuron system in coding actions is more complex than previously shown and extends from action recognition to the coding of intentions. Experiments in monkeys demonstrated that frontal and parietal mirror neurons code the 'what' of the observed action (e.g. 'the hand grasps the cup'). They did not address, however, the issue of whether these neurons, or a subset of them, also code the 'why' of an action (e.g. 'the hand grasps the cup *in order to drink*')....Because 'drinking' and 'cleaning' contexts determined different activations in the Intention condition, it appears that

there are sets of neurons in human inferior frontal cortex that specifically code the ‘why’ of the action and respond differently to different intentions (Iacoboni *et al.*, 2005, pp. 4–5).

Their neurological explanation for this increased activity, and therefore the way in which MNs encode intentions, is:

that a subset of mirror neurons in the inferior frontal cortex discharge in response to the motor acts that are most likely to follow the observed one. In other words, in the Intention condition, there is activation of classical mirror neurons coding other potential actions sequentially related to the observed one. This interpretation of our findings implies that, in addition to the classically described mirror neurons that fire during the execution and observation of the same motor act... there are neurons that are visually triggered by a given motor act (e.g. grasping observation) but discharge during the execution not of the same motor act but of another act, functionally related to the observed act (e.g. bringing to the mouth). Neurons of this type have indeed been previously reported in [region] F5 and referred to as ‘logically related’ neurons (p. 5).

Let’s assume that this interpretation of the increased activity observed in Intention conditions is indeed correct. So, to summarise, what is happening in the two conditions is as follows:

(a) *‘Before Tea’ Condition:*

- (i) firing of MNs appropriate to observing a conspecific grasp a cup
- (ii) firing of logically related MNs appropriate to some functionally related action, e.g. moving the cup to one’s lips.

(b) *‘After Tea’ Condition:*

- (i) firing of MNs appropriate to observing a conspecific grasp a cup
- (ii) firing of logically related MNs appropriate to some functionally related action, e.g. moving the cup to one side.

It is the addition of (ii) which makes the difference between the two conditions and allows us to treat the action in (a) as a grasp-to-drink and the action in (b) as a grasp-to-clear-up.<sup>10</sup> This gives us our *revised version of the MN hypothesis* : MNs underpin the attribution of

[10] So, on this suggestion, it ought to be possible to get ‘garden pathing’ effects; thus we would expect that where an action seen in one setting (e.g. picking up cup before tea) primes one action (e.g. moving cup to lips) but the experimenter does something else (e.g.

intention to others because witnessing a conspecific perform a given action causes patterns of neuronal activity corresponding to our own performance of that action together with the performance of a functionally related action appropriate to the initial action in the given context. This neuronal activity is taken 'off-line' and constitutes a simulation of the mind of the conspecific. The question then is: does this revised MN hypothesis provide a plausible model of mindreading?

To start with, let's consider one piece of experimental evidence which might be thought to tell in favour of Iacoboni et al's proposal. This comes from Andrew Meltzoff's experiments with copying behaviour in children (Meltzoff, 1995). In these experiments children were exposed to novel toys and, in the crucial contexts, witnessed an experimenter apparently trying but failing to perform a given action with the toy. For instance, presented with a toy containing a hook and a loop, the experimenter mimed trying to put the loop over the hook, but on each occasion he failed, dropping the loop ineffectually on one side of the toy. Children then given the toy performed not the behaviour they had actually witnessed (e.g. picking up and dropping the loop) but the action the experimenter seemed to have been aiming at, in this case hanging the loop successfully on the hook. Meltzoff's suggestion is that children are able to 'see through the action to the intention'.

Given Iacoboni *et al.*'s hypothesis then we might think of these cases as instances where the children's mirror systems had fired in a way appropriate to seeing the picking up of the loop and the agent moving it toward the hook, and this had led to the firing of logically related neurons appropriate to putting the loop on the hook. In this way, when children themselves came to re-enact the behaviour they had, from a neurological perspective, already completed the sequence of actions they went on to demonstrate (picking up the loop and hanging it on the hook). This would explain the way in which children 'see through action to intention', as Meltzoff originally suggested. However, attractive as this explanation is, it is not at all clear that the experiment lends itself to our revised MN hypothesis. The problem is that the experiment was specifically designed using toys the child had never been exposed to before, yet given this it is not clear how children (or their mirror systems) could have an understanding of what constitutes 'typical' behaviour with such objects. Why would hanging

---

turns the cup upside down) there should be some evidence of forced reinterpretation or recognition of an anomaly by the viewer.

the loop on the hook in this context count as ‘what’s most likely to come next’, independent of the fact that this was what the experimenter intended to do? This worry I think highlights a fundamental concern with the revised MN hypothesis, to which I turn now.<sup>11</sup>

I think the main problem with the revised MN hypothesis is that it depends on the idea of ‘actions which are functionally related to observed actions’. Yet what action is appropriate, or functionally related, to another is a matter that can only be settled given a *prior* grasp of the mental states of the agent. Thus an account which tries to spell out grasp of agent intentions by appealing to a pre-existing connection between one kind of physically described gesture and another seems problematic, because any such patterns of behaviour are established only via an appeal to the intentions they are supposed to explain. This worry surfaces in a number of ways:

- (i) Not all intentions result in future action. I might intend to pick up the cup because I want to see it more closely, or I want to feel how heavy it is, or because I don’t want you to have it, yet there need be no action ‘sequentially associated’ with any of these intentions. Yet if there is no (typical) next action then, on the current proposal, there is no way for MNs to capture these intentions.
- (ii) Patterns of MN stimulation alone seem too fine-grained to encode intention, since different patterns of MN stimulation may map to a single intention. For instance, take two different motor actions – say a grasp and hold versus a grasp and put down – which will be mirrored by two different patterns of MN stimulation. Nevertheless it is clear that both these sequences of actions could subserve a single intention, such as the intention to tidy up. This seems to undermine the idea

---

[11] It is also interesting to note that one of the points offered in support of the MN hypothesis in Gallese and Goldman’s 1998 paper (and also appealed to in Gallese 2001) might be difficult to accommodate on the revised version of the MN hypothesis in Iacoboni *et al.* Gallese and Goldman appeal to MNs to explain the repetitive copying behaviour seen in certain patients with cognitive impairments. They suggest that such behaviour can be understood as the result of MN activation combined with damage to the mechanism responsible for taking this activation ‘off line’. However, if this is the right explanation in these cases, Iacoboni *et al* ought (*ceteris paribus*) to predict that the behaviour such patients end up displaying will reflect not just the gestures they actually observe but also some typical ensuing movement (due to the — now on-line — firing of logically related MNs). Yet, as far as I know, this is not the sort of compulsive copying witnessed in these cases. Of course, this could be explained by positing damage to both the off-line mechanisms and to the firing of logically related neurons. However work would be needed to show that such an explanation was not *ad hoc*.

that there is a functionally related action following grasping which encodes the intention to tidy up.<sup>12</sup>

- (iii) Patterns of MN stimulation alone seem too coarse-grained to encode intention, since a single pattern of MN stimulation might map to different intentions. For instance, MN stimulation appropriate to grasping a cup and then putting it to one side might subservise an intention to tidy up or an intention to place the cup nearer to you so that you can finish your tea. Thus, even in a situation in which, first, an agent observes the grasping of a cup (which causes a certain pattern of MN stimulation) and, second, features of the context determine a functionally related movement (inducing the firing of logically related neurons mirroring the action of moving the cup to one side), this gets us no closer to intentional attribution, since one and the same related action could still indicate a range of very different intentions.<sup>13</sup>

The general worry here is that Iacoboni et al seem to assume that it is reasonable to talk about the typical series of actions which would realise a given intention, but at the level at which MNs operate (namely physically described motor actions) it simply doesn't seem right, as the above points show, to think that there are any such sequences of actions which are constitutive of agent intentions.

Perhaps, however, the claim that we can determine a typical series of actions for a given intention might seem more compelling with respect to evolutionarily more basic intentions, like eating and drinking. So that even if the revised MN hypothesis can't give us an account of intentional attribution across the board, it could serve to show how attribution of certain basic intentions comes about.<sup>14</sup> Thus we might think of taking Gallese and Goldman's modesty about the

[12] Nichols and Stich (2003, p. 139) raise a similar worry for the overarching simulation theory claim that an observer may 'retrodict' the intentions of a target by, first, observing their behaviour and, secondly, making use of one's own intentional system to figure out what intentions could have led to such an act.

[13] Objection (ii) concerned a many:one mapping from patterns of MN stimulation to intentions. While this might help to cast doubt on the proposed connection between the two, it is not a knock-down objection. However, the current objection argues for a one:many mapping from patterns of MN stimulation to intentions and this, if correct, would seem to show the proposed identity between MN stimulation and intentions is unworkable.

[14] Indeed this is a claim to which Iacoboni *et al.* seem sympathetic: 'The stronger activation of the inferior frontal cortex in the "drinking" as compared to the "cleaning" Intention condition is consistent with our claim that a specific chain of neurons coding a probable sequence of motor acts underlies the coding of intention. There is no doubt that, of these two actions, drinking is not only more common and practiced but also belongs to a more basic repertoire, therefore, that the chain of neurons coding the intention of drinking is

MN hypothesis to heart once again and claim that any such system does not provide a realisation of a full-blown mind-reading system but rather an account of a limited set of evolutionarily basic goals and the actions they subsume (that is to say, MNs get us to see a given grasp as an act of ‘picking up the cup because X wants a drink’, though some other system underpins attributions like ‘picking up the cup because X wants to clear up’). Yet notice that even in other apparently evolutionarily basic cases, it is not obvious that talk of typical ensuing actions can help. Picking up a cup with the intention of drinking might typically be followed by moving the cup to one’s lips, but what’s the appropriate next action after pointing to a predator (presumably something which is pretty ancient), is it running away, or climbing a tree, or (more problematic still for the MN theory) staying still? Or again, picking something up because it interests you, or because you want to see what is underneath it, seems pretty basic in evolutionary terms but once more the idea of a typical next action seems extremely strained. Furthermore, it is simply not obvious that we can place any weight on the notion of ‘typical’ (or ‘functionally related’) here without appealing back to the other mental states of the target. If one picks up a cup with the intention of drinking from it, moving it to one’s lips is a typical next move *only if* one doesn’t also believe that the fluid is currently too hot to drink, or that if the cup contains red fluid then it contains poison and should not be drunk, etc. What one is likely to do next depends not just on features of the context and the initial act in that context but on the complex network of one’s beliefs and desires. Yet this fact threatens to render the proposed explanation circular: the revised MN hypothesis seeks to attribute mental states to agents via an appeal to typical behaviour, but one can isolate typical behaviour only in terms of one’s attribution of mental states.<sup>15</sup>

---

more easily recruited and more widely represented in the inferior frontal cortex than the chain of neurons coding the intention of cleaning’ (p. 5).

- [15] A reader for this journal noted that Iacoboni et al could, at this point, respond that in many mindreading tasks subjects represent the target’s intention without explicitly representing their beliefs and desires. Thus it would seem that MN activity alone might suffice in these cases. However I think this somewhat misplaces the above challenge. The claim is not merely that MN activity is insufficient to capture the representation of beliefs and desires (so that some other mechanism must be in place to explain this aspect of mindreading) but that without some assumptions about the target’s beliefs and desires (be these explicit representations or some sort of tacit inferential base) witnessing an action in context is always insufficient for attributing an intention to another. So, if MN activity doesn’t furnish us with access to the beliefs and desires of others this entails that MN activity alone is insufficient to furnish an observer with access to the intentions behind the acts of others.

### 3. Conclusion

To summarise then: it seems that the discovery of MNs might tell us something about mindreading. Recording the conditions in which MNs are triggered might, it seems, help to limit the extension of the term ‘conspicuous’ for an individual: it might go some way towards determining which entities in the world a person was prepared to treat as ‘minded’. But any stronger claim, either along the lines that MN activation is a necessary feature of intentional attribution *per se*, or that the discovery of MNs supports the simulation approach to mindreading, remains dubious. The MN hypothesis remains dubious because either we need an argument to show that a simulation-based precursor to mindreading entails a simulation-based account of mindreading-proper (the worry for the original version of the MN hypothesis) or we need a way to characterise sequences of functionally related actions which does not in itself appeal to the intentions of the agent (the worry for the revised MN hypothesis).

The objections to the revised MN hypothesis turn, I think, on familiar anti-behaviourist lines. According to the behaviourist an agent counts as being in a particular mental state just in case they are disposed to engage in certain kinds of behaviour. So being in pain becomes a matter of being disposed to cry out and rub the affected area, etc. Similarly, then, for the revised MN hypothesis having a specific intention (and attributing it to others) is a matter of one’s motor system being disposed to fire in a specific way — wanting a drink is a matter of one’s MN system firing in a way appropriate to picking up a cup and moving it to one’s lips. However, as standard objections to behaviourism have stressed, there simply does not seem to be any (non-circular) way to move between descriptions of behaviour and claims about mental states.<sup>16</sup> So, if there is any general lesson to be gleaned from the MN hypothesis as considered in this paper, it seems

---

[16] One objection here, as noted by a reader for this journal, is that the challenges raised in this paper for the MN hypothesis are themselves premised on too behaviourist an outlook, for they take the stimuli to which MNs respond to be ‘bare bones’ behaviour — that is to say, mere physical behaviour stripped of all meaning. If we instead assume a more McDowellian picture (see McDowell, 1994) where nature is ‘re-enchanted’ and meaning is found as part of an external reality waiting to be discovered by the subject then we might, as the reader commented, construe MNs as ‘interpreters giving us a direct channel for discovering meaning in what we perceive out there in the world’. In this case MNs would clearly be crucial in our grasp of meaning. The reader also suggested that my use of the technical term ‘intentional attribution’ in the paper was misleading as it disguised this role for MNs. To respond to the latter point first: intentional attribution is a kind of meaning detection (so debates about meaning will indeed range wider than debates about intention detection) but I would suggest that the arguments of this paper carry over from concerns about intention-detection to any other intentional notion: the general point is that

to be that cognitive science should resist the lure of behaviourism, even when it comes in disguised form.<sup>17</sup>

---

physically-described aspects of the world underdetermine intensional descriptions and thus that, if this is what MNs respond to, then they are insufficient to explain intensional-level accounts (see also Hurford 2004 for a criticism of the claim that MNs can explain grasp of specifically linguistic meaning). This brings us to the former objection: that MNs are not responding to mere physical behaviour but to enriched, meaning-saturated aspects of the world. It is not possible to respond fully to this challenge in the confines of a paper but I would like to make two brief points: first, embracing this kind of metaphysical picture is a serious commitment (and one which brings significant challenges in its wake). Thus, at the very least, advocates of the MN hypothesis should make it clear that this is the kind of world view they endorse. Secondly, even if this picture is the one in play it still seems unclear to me to what extent MNs really support the simulation theory of mindreading: if meaning is part of an objective reality which MNs put us in touch with, then why think we need to simulate the agent who performs the act in order to grasp the meaning she endowed it with? At most it would seem that the simulation claim is now exhausted by the idea that we have a single mechanism (MN activity) which is solely responsible both for the production and the detection of meaning. However even this claim seems mistaken to me, for just as I've been arguing that MN activity is insufficient to explain meaning (or intention) detection, so it seems that MN activity is insufficient to explain meaning (or intention) production. As far as I'm aware MN activity in action-production is insensitive to fine-grained intentions: that is to say, we witness the same pattern of MN stimulation if an agent engages in a precision grip of a cup with the intention of drinking or if they engage in a precision grip of a cup with the intention of examining the cup's contents. But if this is right then MNs can't be the whole story in meaning-generation anymore than they can be the whole story in meaning-detection.

- [17] The revised MN hypothesis seems particularly problematic to me as it seems to constitute a genuine appeal to behaviourism. However we might also note that there seems to be a rather more subtle re-emergence of behaviourism in some other recent work in this area. Thus the appeals to MNs to answer questions of language acquisition in Arbib et al 2005 seem to rely on what we might term 'epistemic behaviourism'. According to this approach, although behaviour may not provide an answer to the constitutive question 'what is it to be in mental state F', it does provide the answer to the epistemic question 'how do I know that you are in mental state F?'. For instance, Arbib et al write: 'Our approach integrates perception, the building of action, and the meaning of words, despite the fact that many studies of language acquisition assume that gestures entail ambiguity of reference... These authors rely on Quine's classic essay (1960) in which he discussed the ambiguity of reference entailed in, say, speaking about and pointing to, a rabbit. But caregivers tend to focus attention with precision. They do not simply say an unfamiliar word (such as Quine's *gavagai*) while pointing. Instead, caregivers may rub a rabbit's fur while saying, 'fur'; trace the topography of its ears while saying, 'ear', stroke the entire rabbit or rotate the whole animal when saying 'rabbit', etc. ... Successful teaching entails marking the correspondence between what is said and what is happening' (p. 247). However, I think many philosophers will suspect that this underestimates the strength of the Quinean challenge: a tabula rasa child faced with a caregiver tracing the topography of a rabbit's ear while saying 'ear', is no more warranted in taking the meaning of the term to be *ear*, than she is in taking it to be *rabbit-ear*, or *temporal slice of rabbit-ear*, or *rabbit-eared shape*, etc. While behaviour is no doubt very important on route to intentional explanation, I think serious questions still remain to be asked about the validity of any such epistemic behaviourism. However such questions cannot be explored here.

### *Acknowledgement*

This paper was presented at the annual conference of the European Society for Philosophy and Psychology in Belfast in 2006, thanks are due to the audience there for comments. Work on this paper was made possible by funding from the Arts and Humanities Research Council.

### **References**

- Arbib, M., Oztop, E., and Zuckow-Goldring, P. (2005), 'Language and the mirror system: A perception/action based approach to communicative development', *Cognition, Brain, Behaviour*, **3**, pp. 239–72.
- Gallagher, S. (2006), 'Perceiving others in action', *Fondements cognitifs de l'interaction avec autrui*.
- Gallese, V. (2000). 'Agency and motor representations: new perspectives on intersubjectivity', ISC working papers 2000–6. <http://www.isc.cnrs.fr/wp/wp006.htm>.
- Gallese, V. (2001), 'The "Shared Manifold" hypothesis: from mirror neurons to empathy', *Journal of Consciousness Studies* **8** (5–7), pp. pp. 33–50.
- Gallese, V. and Goldman, A. (1998), 'Mirror neurons and the simulation theory of mind reading', *Trends in Cognitive Science*, **2**, pp. 493–501.
- Gergely, G., Nadasdy, Z., Csibra, G. and Biro, S. (1995), 'Taking the intentional stance at 12 months of age', *Cognition*, **56**, pp. 165–93.
- Heider, F. and Simmel, M. (1944), 'An experimental study of apparent behaviour', *American Journal of Psychology*, **57**, pp. 243–59.
- Hurford, J. (2004), 'Language beyond our grasp: What mirror neurons can, and cannot, do for language evolution', in *Evolution of Communication Systems: A Comparative Approach*, ed. D. Kimbrough Oller and U. Griebel (Cambridge, MA: MIT Press), pp.297–313.
- Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. and Rizzolatti, G. (2005), 'Grasping the intentions of others with one's own mirror neuron system', *PLoS Biology* **3**, e79, pp. 1–7.
- McDowell, J. (1994), *Mind and World* (Cambridge, MA: Harvard University Press).
- Meltzoff, A. (1995), 'Understanding the intentions of others: re-enactment of intended acts by 18-month-old children', *Developmental Psychology*, **31**, pp. 838–50
- Nichols, S. and Stich, S. (2003), *Mindreading* (Oxford: OUP).
- Rizzolatti G. and Craighero L. (2004), 'The mirror-neuron system', *Annual Review of Neuroscience*, **27**, pp. 169–92.
- Rizzolatti G., Fogassi L., and Gallese V. (2001), 'Neurophysiological mechanisms underlying the understanding and imitation of action', *Nature Reviews Neuroscience*, **2**, pp. 661–70.
- Williams, J., Whiten, A., Suddendorf, T. and Perrett, D. (2001), 'Imitation, mirror neurons and autism', *Neuroscience and Biobehavioural Reviews*, **25**, pp. 287–95.

Paper received June 2006; revised March 2007.