The mirror-neuron system: a Bayesian perspective

James M. Kilner, Karl J. Friston and Chris D. Frith

Wellcome Trust Centre for Neuroimaging, UCL, London, UK

Correspondence to James M. Kilner, Wellcome Trust Centre for Neuroimaging, Institute of Neurology, UCL, 12 Queen Square, London WC1N 3BG, UK
Tel: +44 20 7833 7472; fax: +44 20 7813 1472; e-mail: j.kilner@fil.ion.ucl.ac.uk

Received 24 January 2007; accepted 11 February 2007

Is it possible to understand the intentions of other people by simply observing their movements? Many neuroscientists believe that this ability depends on the brain's mirror-neuron system, which provides a direct link between action and observation. Precisely how intentions can be inferred through movement-observation, however, has provoked much debate. One problem in inferring the cause of an observed action, is that the problem is ill-posed because identical movements can be made when performing different actions with different goals. Here we suggest that this problem is solved by the mirror-neuron system using predictive coding on the basis of a statistical approach known as empirical Bayesian inference. This means that the most likely cause of an observed movement can be inferred by minimizing the prediction error at all cortical levels that are engaged during movement observation. This account identifies a precise role for the mirror-neuron system in our ability to infer intentions from observed movement and outlines possible computational mechanisms. *NeuroReport* 18:619–623 © 2007 Lippincott Williams & Wilkins.

Keyword: Bayesian, inference, mirror-neurons, predictive coding

Introduction

Humans can infer the intentions of others through observation of their actions. Here we define action as a sequence of acts or movements with a specific goal. Little is known about the neural mechanisms underlying this ability to 'mind read', but a strong candidate is the mirror-neuron system (MNS) [1–3]. Mirror-neurons discharge not only during action execution but also during action observation. Their participation in both action execution and observation suggests that these neurons are a possible substrate for automatic action understanding. Mirror-neurons were first discovered in the premotor area, F5, of the macaque monkey (see [3] for a review) and have been identified subsequently in an area of the inferior parietal lobule, area PF [4]. Neurons in the superior temporal sulcus (STS), also respond selectively to biological movements, both in monkeys [5] and in humans [2], but they are not mirror-neurons as they do not discharge during action execution. Nevertheless, they are often considered part of the MNS [6].

Mirror-neurons and the MNS have been the focus of much interest since their discovery because they have been proposed as a neural substrate that enables us to understand the intentions of others through observation of their actions [1]. Actions can be understood at many different levels. Following Hamilton and Grafton [7], we will consider actions that can be described at four levels: (i) the intention level that defines the long-term goal of an action – for example, to pour a glass of wine; (ii) the goal level that describes short-term goals necessary to realize the intention; for example, reaching and grasping a wine bottle; (iii) the motor signals that describe the pattern of muscle activity through which the action is executed; and (iv) the kinematic level that describes the configuration of the hand and the movement of the arm in space and time. To understand the intentions or goals of an observed action, the observer must be able to describe the observed movement at either the goal level or the intention level having only access to a visual representation of the kinematic level. It is, however, unclear how the visual information from an observed action maps onto the observer's own motor system and how the goal of that action is inferred [8–11]. Gallese [12] recently noted that '... we do not have a clear neuroscientific model of how humans can understand the intentions promoting the actions of others they observe'. Therefore, the question remains how do mirror-neurons mediate understanding of actions performed by others? In answer to this, Rizzolati and Craighero [3] wrote 'The proposed mechanism is rather simple. Each time an individual sees an action done by another individual, neurons that represent that action are activated in the observer's premotor cortex. This automatically induced, motor representation of the observed action corresponds to that which is spontaneously generated during active action and whose outcome is known to the acting individual. Thus, the mirror-neuron system transforms visual information into knowledge'.

Implicit in this, and many other descriptions of the MNS, is the idea that information carried by visual signals is transformed as it is passed by forward connections between areas of the MNS from low-level representations of the movement kinematics to high-level representations of intentions subtending the action. Specifically, the observation of an action drives the firing of neurons in the STS, which drives activity in the parietal mirror neuron area PF, which in turn, drives activity in inferior frontal (premotor) mirror neuron area F5 (Fig. 1a).

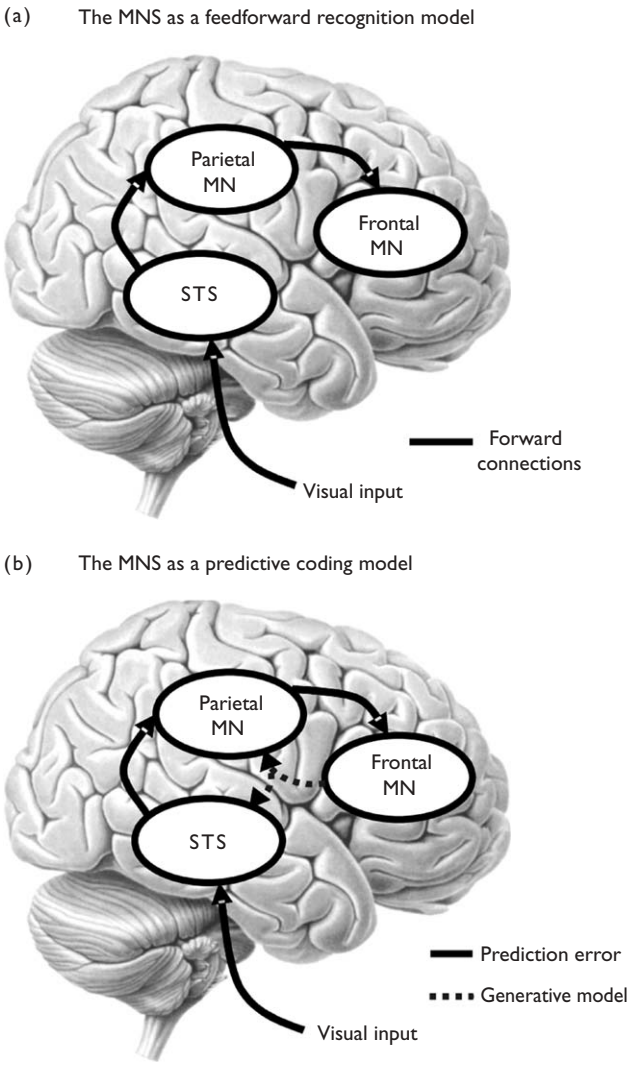

Fig. 1 Schemas of the mirror-neuron system, showing the STS, parietal mirror-neurons (MN) and frontal mirror-neurons. STS, superior temporal sulcus.

A problem arises with this account. Formally, this is a recognition model that implicitly inverts a generative model of how the observations to be recognized are generated. When executing an action, the representation of our goal in F5 generates the appropriate muscle commands, which cause the action and the associated visual kinematics. A direct relationship exists between the motor commands and the resultant visual kinematics, so that the brain can readily use goals to generate a model of the predicted visual kinematics (i.e. goal→predicted visual kinematics: the forward model). During action observation, however, can the brain covert visual signals into knowledge about goals by simply inverting this process, that is visual kinematics→predicted goal? The problem with such a simple feedforward recognition scheme is that it will only work when the processes generating the sensory inputs from the causes, the generative models, can be inverted. This can only occur when one sensory input is associated uniquely with one cause. In general, this is not the case as the same sensory input can have many causes. This is known as an ill-posed inverse problem. In the specific case of action-observation, the same visual kinematics can be caused by different goals and intentions. For example, while walking along the street, if someone suddenly waves their arm, are they hailing a taxi or swatting a wasp? Furthermore, the empirical data do not support this simple feedforward recognition account. Mirror-neurons in area F5 that discharge when a monkey is observing a reach and grasp action also discharge when the sight of this movement is partly hidden behind a screen [13]. Critically, this result shows that mirror-neurons in area F5 are not driven simply by the visual representation of an observed movement.

Here, we propose that the role of the MNS in inferring the goals of observed actions can be understood within a predictive coding framework (see [14,15]). The predictive coding account resolves the inverse problem and furnishes some specific hypotheses about the functional architecture of the MNS. The essence of this approach is that, given an expectation of the goal of the person we are observing, we can predict their motor commands. Given their motor commands, we can predict the kinematics on the basis of our own action system. The comparison of the predicted kinematics with the observed kinematics generates an error. This prediction error is used to update our representation of the person's motor commands. Similarly, the inferred goals are updated by minimizing the error between the predicted and the inferred motor commands (Fig. 2). By minimizing the error at all the levels of the MNS the most likely cause of the action will be inferred at all levels (intention, goal, motor and kinematic). This approach provides a mechanistic account of how responses in the visual and motor systems are organized and explains how the cause of an action can be inferred from its observation. Many accounts of the way that the brain encodes the most likely cause of its sensory input rest on minimizing the error of predictions [14,15].

Predictive coding (see Fig. 2) is based on minimizing error through recurrent or reciprocal interactions between cortical areas that are organized hierarchically. In the predictive coding framework, each level of a cortical hierarchy employs a model that generates a prediction of the representations in the level below. This prediction is conveyed to the lower level of the hierarchy via backward connections where it is compared with the representation in this subordinate level to produce an error. This prediction error is then returned to the higher level, via forward connections, to adjust the neuronal representation of sensory causes, which in turn changes the prediction. This self-organizing, reciprocal exchange of signals continues until the error is minimized and the most likely cause of the observed action is inferred (see Figs 1 and 2). It can be shown that this scheme is formally equivalent to empirical Bayesian inference, in which prior expectations emerge naturally from the hierarchal models employed [14,15]. In brief, empirical Bayesian inference harnesses the hierarchical structure of a generative model, treating the estimates at one level as expectations for the subordinate level. This provides a natural framework within which each level provides constraints on the level below to treat cortical hierarchies in the brain. This approach models the brain as a hierarchy of systems where supraordinate causes induce and moderate changes in subordinate causes. These expectations offer contextual guidance towards the most likely cause of the sensory input.

Predictive coding is highly appropriate for understanding the function of the MNS; predictive coding provides an

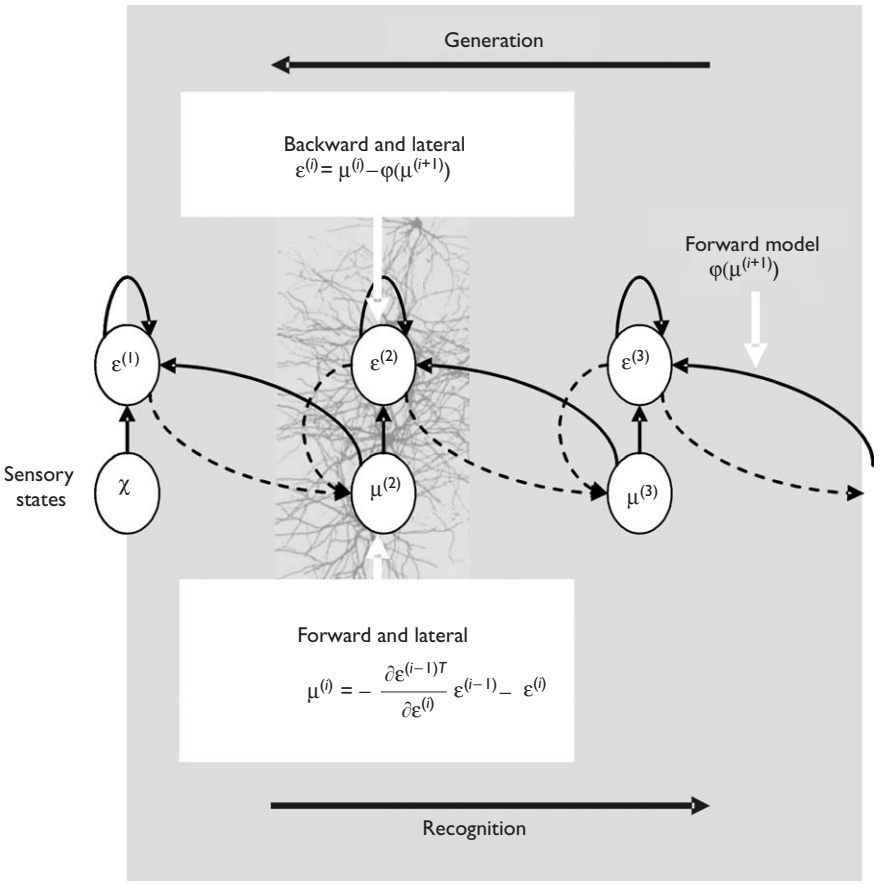

Fig. 2 Hierarchical architecture for predictive coding with empirical Bayes. The empirical Bayesian perspective on perceptual inference suggests that the role of backward connections is to provide contextual guidance to lower levels through a prediction, φ of the lower level's inputs. Given this conceptual model, a stimulus-related response can be decomposed into two components corresponding to the transients evoked in two functional subpopulations of units. The first encodes the conditional expectation of perceptual causes, μ. The second encodes prediction error, ε. Responses are evoked in both, with the error units of one level driving appropriate changes in conditional expectations through forward connections. These expectations then suppress error units using predictions that are mediated by backward connections. These predictions are based on the brain's generative model of how sensory states are caused [15].

established computational framework [15] for inferring the intentions, goals and motor commands, from an observed movement. It does not, however, explain why mirror-neurons fire during both action-observation and execution. To explain this we can appeal to a compelling body of work on the use of generative or forward models in motor control and learning: It is now generally accepted that when we execute a movement, we predict the sensory consequences of that movement through generative or forward models (see [16]). These predictions can then be used to circumvent motor control problems induced by delayed feedback and sensory noise. In short, forward models that generate predicted kinematics from motor commands might be an integral part of motor execution. The suggestion here is that the same models are used to infer motor commands from observed kinematics produced by others during perceptual inference (see [17] for a similar computation in the domain of language perception). In execution, motor commands are optimized by minimizing the difference between predicted and desired kinematics, under the assumption that the desired kinematics (i.e. goals) are known. Conversely, in action-perception, these goals have to be inferred. In both optimizations, however, a forward model of motor control is required. In the predictive coding account of the MNS, the same generative model used to predict the sensory effects of our own actions can also be used (with appropriate transformations) to predict the sensory effects of the actions of others (see Ref. [15] for a description of the relationship between forward and inverse models and predictive coding).

There have been several previous accounts that have proposed the use of forward and inverse models in action-observation [6,16,18]: 'skilled motor behaviour relies on the brain learning both to control the body and predict the consequences of this control. Prediction turns motor commands into expected sensory consequences, whereas control turns desired consequences into motor commands. To capture this symmetry, the neural processes underlying prediction and control are termed the forward and inverse internal models, respectively' [19]. First, forward and inverse models have been proposed as a mechanism for imitation; the inverse model (mapping kinematics to motor signals) is identical to the inversion model of the MNS (shown in Fig. 1a). The logic is that the inverse model can be used as recognition model and therefore can infer the cause of an observed action. Once the cause of the observed kinematics is inferred the action can then be imitated. Second, the hierarchical modular selection and

identification for control (HMOSAIC) account of motor control has recently been extended to provide a framework for understanding social interactions [16]. In the HMOSAIC account several predictor-controller pairs are organized hierarchically. The predictor is a forward model that predicts the consequences of a motor act. The controller is an inverse model that computes the motor act required to achieve the required consequences. Thus, there are several links between the HMOSAIC account and the predictive coding account described here.

Although these generalizations of forward-inverse models in motor control to imitation and social interactions are exciting; they are formally distinct from, and more complicated than, the predictive coding account of the MNS. In contrast to the HMOSAIC account, in predictive coding there is no separate inverse model or controller; a forward model is inverted by suppressing the prediction error generated by the forward model. In the predictive coding account, this inversion depends on the self-organizing, reciprocal exchange of signals between hierarchical levels (Fig. 2). In distinction to the HMOSAIC account, this simplicity translates into a set of computations that could be implemented by the brain [14,15,18].

Predictions and evidence for the predictive coding account: neurophysiology

Predictive coding requires that each level of the MNS predicts the activity in the level below. An increasing evidence that the MNS is predictive exists. First, mirror-neurons in area F5 that discharge when a monkey is observing a reach and grasp action also discharge during observation of the same movement even when the sight of this movement is partly hidden [13]. This result shows that these neurons are not simply driven by an observed movement (as would be the case with the recognition model described in Fig. 1a). Rather, this pattern of discharge suggests that these neurons are predicting the most likely kinematics of the action irrespective of whether the whole sequence is visible or not. Similarly, neurons in the inferior parietal lobule area PF show different patterns of discharge when observing the same movement performed during two different contexts [4]. In this study, monkeys observed movements where the experimenter reached and grasped an object. The experimenter then either placed that object in their mouth or in a container next to their mouth. During the period when monkeys were observing the initial grasp, a movement that was common to both action-observation conditions, some of the mirror-neurons in area PF showed a differential discharge pattern that depended on the subsequent movement. Again this cannot be explained by a feedforward recognition model such as that outlined in Fig. 1a, but is entirely consistent with a predictive model.

In humans, there is more direct evidence of prediction during action observation. When healthy humans make a movement there is a slow negative-going potential, the readiness potential, which can be recorded using EEG in the 1 s before the movement execution. A similar potential is also present during action observation before an observed predictable movement [20]. The authors argue that this suggests an active role for the MNS in making a prediction of another person's action, endowing the brain with the ability to predict his/her intentions ahead of their realization. Finally, when participants watch a predictable action, their eye gaze leads the observed movement. In other words, the coordination between their gaze and the actor's hand is predictive, rather than reactive [21]. It should be noted that predictive coding pertains to the predicted effects of causes, not the forecast or prediction of future effects from past effects. In generalizations of predictive coding for dynamic systems, the motion or trajectory of an effect, however, is predicted from the trajectory of causes. This is particularly relevant for forward models in motor control that have to operate in real time.

Predictions and evidence for the predictive coding account of the MNS: anatomy

The predictive coding account of the MNS requires that the areas engaged by action-observation are arranged hierarchically and the anatomical connections between these areas are reciprocal. Furthermore, predictive coding posits functional asymmetries in forward and backward connections; forward connections furnish feedback that is linear (i.e. driving) on the prediction error (see Fig. 2), whereas backward connection implementing nonlinear generative models should show modulatory influences. These predictions can be tested empirically and there is already some supporting evidence. The three cortical areas, which constitute the MNS, the STS, area PF of the inferior parietal lobule and area F5 of the premotor cortex, are reciprocally connected. In the macaque monkey, area F5 in the premotor cortex is reciprocally connected to area PF [22] creating a premotor-parietal MNS, and STS is reciprocally connected to area PF of the inferior parietal cortex [23,24] providing a sensory input to the MNS (see [6] for a review). Furthermore, these reciprocal connections show regional specificity. Although STS has extensive connections with the inferior parietal lobule, area PF is connected to an area of the STS that is specifically activated by observation of complex body movements. An analogous pattern of connectivity between premotor areas and inferior parietal lobule has also been demonstrated in humans [25]. Given that the areas of the MNS are reciprocally connected, what is the hierarchical organization of the MNS? The notion that the area F5 is the highest level of any hierarchy is implicit in many accounts of the MNS. This is the hierarchical arrangement shown in Fig. 1. No direct evidence, however, exists to support this view and the results of recent studies suggest that the inferior parietal lobule area may be supraordinate to premotor areas in the MNS hierarchy [7]. Under a predictive coding account, this organization will determine which connections are mediating the generative model and which the prediction error.

Summary

Social interaction depends upon our ability to infer beliefs and intentions in others. It has been suggested that the MNS could underlie this ability to 'read' someone else's intentions. Here we have proposed that the MNS is best considered within a predictive coding framework. One of the attractions of predictive coding is that it can explain how the MNS could infer someone else's intentions through observation of their movements. Within this scheme the most likely cause of an observed action is inferred by minimizing the prediction error at all levels of the cortical hierarchy that is engaged during action-observation. This

account specifies a precise role for the MNS in our ability to infer intentions and formalizes the underlying computations. It also connects generative models that are inverted during perceptual inference with forward models that finesse motor control. Furthermore, it makes the following specific predictions about the behaviour and organization of the MNS that can be tested empirically.

Anatomically: the areas engaged by movement observation are arranged hierarchically and the anatomical connections between these areas are reciprocal. These reciprocal connections should show functional asymmetries with context-sensitive or nonlinear influences being exerted by backward connections.

Functionally: the MNS is predictive with (empirical) Bayesian properties and should show context-sensitive responses during action-observation.

In terms of functional anatomy: prediction error encoding higher-level attributes (e.g. goals and intentions) will be expressed as responses in higher cortical levels of the MNS.

In terms of effective connectivity: backward connections play an essential and demonstrable role in forming responses, to the extent that disabling them will preclude inference about observed actions.

Acknowledgements

The Wellcome Trust funded this work. The authors would like to thank Professor G. Gabella for helpful comments in revising an earlier version of this manuscript.

References

1. Gallese V, Goldman A. Mirror-neurons and the simulation theory of mind reading. *Trends Cogn Sci* 1998; **2**:493–501.
2. Frith CD, Frith U. Interacting minds: a biological basis. *Science* 1999; **286**:1692–1695.
3. Rizzolatti G, Craighero L. The mirror-neuron system. *Annu Rev Neurosci* 2004; **27**:169–192.
4. Fogassi L, Ferrari PF, Gesierich B, Rozzi S, Chersi F, Rizzolatti G. Parietal lobe: from action organization to intention understanding. *Science* 2005; **308**:662–667.
5. Oram MW, Perrett DI. Responses of anterior superior temporal polysensory (STPa) neurons to biological motion stimuli. *J Cogn Neurosci* 1994; **6**:99–116.
6. Keysers C, Perrett DI. Demystifying social cognition: a Hebbian perspective. *Trends Cogn Sci* 2004; **8**:501–507.
7. Hamilton AF, Grafton ST. The motor hierarchy: From kinematics to goals and intentions. In: Rosetti Y, Kawato M, Haggard P, editors. *Attention and performance xxii*. (In press).
8. Gallese V. Intentional attunement: a neurophysiological perspective on social cognition and its disruption in autism. *Brain Res* 2006; **1079**:15–24.
9. Iacoboni M. Neural mechanisms of imitation. *Curr Opin Neurobiol* 2005; **15**:632–637.
10. Jacob P, Jeannerod M. The motor theory of social cognition: a critique. *Trends Cogn Sci* 2005; **9**:21–25.
11. Saxe R. Against simulation: the argument from error. *Trends Cogn Sci* 2005; **9**:174–179.
12. Gallese V. Embodied simulation: from mirror neuron systems to interpersonal relations. In: Bock G, Goode J, editors. *Empathy and fairness*. Novartis Foundation. 2006.
13. Umilta MA, Kohler E, Gallesse V, Fogassi L, Fadiga L, Keysers C, Rizzolatti G. I know what you are doing. A neurophysiological study. *Neuron* 2001; **31**:155–165.
14. Friston KJ. Learning and inference in the brain. *Neural Netw* 2003; **16**:1325–1352.
15. Friston K. A theory of cortical responses. *Philos Trans R Soc Lond B Biol Sci* 2005; **360**:815–836.
16. Wolpert DM, Doya K, Kawato M. A unifying computational framework for motor control and social interaction. *Philos Trans R Soc Lond B Biol Sci* 2003; **29**:593–602.
17. Chater N, Manning CD. Probabilistic models of language processing and acquisition. *Trends Cogn Sci* 2006; **10**:335–344.
18. Miall RC. Connecting mirror neurons and forward models. *NeuroReport* 2003; **14**:2135–2137.
19. Flanagan JR, Vetter P, Johansson RS, Wolpert DM. Prediction precedes control in motor learning. *Curr Biol* 2003; **13**:146–150.
20. Kilner JM, Vargas C, Duval S, Blakemore S-J, Sirigu A. Motor activation prior to observation of a predicted movement. *Nat Neurosci* 2004; **7**: 1299–1301.
21. Flanagan JR, Johansson RS. Action plans used in action observation. *Nature* 2003; **424**:769–771.
22. Luppino G, Murata A, Govoni P, Matelli M. Largely segregated parietofrontal connections linking rostral intraparietal cortex (areas AIP and VIP) and the ventral premotor cortex (areas F5 and F4). *Exp Brain Res* 1999; **128**:181–187.
23. Harries MH, Perrett DI. Visual processing of faces in temporal cortex: physiological evidence for a modular organization and possible anatomical correlates. *J Cogn Neurosci* 1991; **3**:9–24.
24. Seltzer B, Pandya DN. Parietal, temporal, and occipital projections to cortex of the superior temporal sulcus in the rhesus-monkey: a retrograde tracer study. *J Comp Neurol* 1994; **343**:445–463.
25. Rushworth MFS, Behrens TEJ, Johansen-Berg H. Connection patterns distinguish three regions of human parietal cortex. *Cerebral Cortex* 2006; **16**:1418–1430.