

The Architecture of Representation

[To appear in *Philosophical Psychology*. Total word count, including abstract, notes and references: 10147]

Rick Grush
Philosophy-Neuroscience-Psychology Program
Washington University in St. Louis
rick@twinearth.wustl.edu
<http://www.artsci.wustl.edu/~rgrush/>

Abstract: In this article I outline, apply, and defend a theory of natural representation. The main consequences of this theory are: i) representational status is a matter of how physical entities are used, and specifically is **not** a matter of causation, nomic relations with the intentional object, or information; ii) there are genuine (brain-)internal representations; iii) such representations are really *representations*, and not just farcical pseudo-representations, such as attractors, principal components, state-space partitions, or what-have-you; and iv) the theory allows us to sharply distinguish those complex behaviors which are genuinely cognitive from those which are merely complex and adaptive.

0. Introduction:

There is no notion more crucial to the study of thought and cognition than *representation*. It is the fact that cognitive systems traffic in representations that sets them apart from merely complex and interesting, but non-cognitive, systems like elms, oceans, and microwave ovens. As crucial as this notion is, and as much theoretical attention and industry as has been devoted to it, it remains a frustrating enigma. Frustrating because it seems to be very simple -- a representation is something that stands for something else -- and yet it has proven quite resistant to any systematic, plausible and revealing analyses.

In this article I outline, apply, and defend a theory of natural representation which is, in the first instance, a theory of how representations are constructed and used by natural systems such as nervous systems. The main consequences of this theory are: i) representational status is a matter of how physical entities are used, and specifically is **not** a matter of causation, nomic relations with the intentional object, or information; ii) there are genuine (brain-)internal representations, contra theorists who maintain that only *external* symbols can be representations; iii) such representations are really *representations*, and not just farcical pseudo-representations, such as attractors, principal components, state-space partitions, or what-have-you;¹ and iv) the theory allows us to sharply distinguish those complex behaviors which are genuinely cognitive from those which are merely complex and adaptive, contra dynamical systems theoretic and related views which treat cognitive phenomena as just complex adaptive behavior on the same continuum with 'simple' sensorimotor integration.

Section 1 will briefly introduce some terminological distinctions and develop an example of representational activity which motivates these distinctions. Section 2 attempts to provide fairly precise characterizations of these concepts in terms of control theory and some associated mathematical apparatus. We will then be in a position to describe an architecture, which I call the Emulation Theory of Representation (ETR), that is a necessary condition on a system's representational status. Section 3 provides an example from robotics research which both instantiates the architecture I describe, and is a clear example of representational

¹ "...while dynamical models are not based on transformations of representational structures, they allow plenty of room for representation. A wide variety of aspects of dynamical models can be regarded as having a representational status: these include states, attractors, trajectories, bifurcations and parameter settings." (van Gelder and Port, 1995) This is just one of an increasingly large number of cases where the notion of representation is being bleached to the point where theorists feel comfortable calling any bit of theoretical exotica a representation if it is some state or process which allows the system to behave appropriately, even if there is no hint of an account of what the content of such a representation might be. Of course, such things may be genuine representations, but this will be because they have some specifiable content, and not just because some mathematically minded theorist appeals to it as part of a non-psychological behavioral explanation.

activity. Section 4 turns to the brain and provides evidence to the effect that the brain does instantiate this architecture for maintaining and using internal representations for use in motor control and imagery. Section 5 extends ETR in order to account for certain crucial features of representation. Section 6 answers some common objections.

Section 1: Representation and Presentation

Representations are entities which stand for something else -- or better, they are entities which are used to stand for something else. This second characterization brings out something not explicit in the first, that of a user. If this second definition, and my gloss on it, are correct, then a representation is a part of a three-way relationship which also includes a user and a target. So far so good. Some may quibble over the need for a user, but that is not where the real problem lies. The real problem has been, and continues to be, the choice of states for which theorists attempt to give a representational analysis. Specifically, sensory states have been used as a model for representational states, the idea presumably being that sensory states represent the world to the subject. The thought seems to be that an analysis of the representational status of sensory states, once adequately done, will then be able to be generalized to other sorts of internal representational states. This choice has the consequence that the theoretical role of a user is optional (e.g. the retinotopic projections in various parts of the CNS represent the visual scene even if these projections are not being attended to or used by anything else in the brain, one might argue).

But this is exactly wrong. A useful theory of representation must not treat sensory input as representational. Such information will be better treated as *presentational*. To see the distinction consider the following analogy. I am playing a game of chess with a friend,

however I am not in the immediate vicinity of my friend or the chess board. Rather, I am at some other location where I learn about my opponent's moves, and issue my own moves, via the telephone. Now I can't keep track of everything in my head, so I keep with me a board which I use to keep track of what the 'official' game board looks like. But, and this will mark the crucial distinction, I also keep a *second* board with me. This second board I use to try out moves, perhaps long sequences of moves, and to assess the possible consequences of those moves and counter-moves. These two uses for chess boards are quite distinct. The first board's use is to accurately mirror the state of the official board, accurate information about this board being crucial to my chances of success in the game. The second board is emphatically NOT used to accurately mirror the state of the official board. Its usefulness for assessing possible positions *depends* on its *not* having to carry information about the actual state of the official board.² It might be thought that one can get by with one board -- one can try out moves and then put the pieces back before making the official move. This is entirely correct. But note that what one is doing in this case is putting the same board to the two different uses I described. One uses the board now to present the real game position, and now takes it 'off line' to try out moves.³

According to the theory I will advance, only the second board is a representation, the first can perhaps be described as a presentation. Most of the recent philosophical literature on internal representation and content has been tying itself in knots because it has not distinguished internal representations from internal presentations, and has been trying futilely to give a theory of content for internal presentations, and then subjecting these

² I have been informed that Robert Cummins (1996) makes a similar distinction. I thank Pete Mandik for pointing this out to me.

³ As we shall see, there is reason to believe that certain cortical areas have exactly this character -- during perception they are driven largely by peripheral sensory organs, but they can be taken off-line to support imagery.

theories to constraints and intuitions which properly belong to representations.⁴ The pitfalls of this conflation should be nowhere more evident than where analyses of inaccurate perception (mistaking a horse on a dark night for a cow, for example) are taken to be relevant to the question why something can be a representation even when it is not an accurate mirror.

As I have remarked, what distinguishes presentations from representations is the use they are put to. A presentation is used to provide information about some other, probably external in some sense, state of affairs. It can be used in this way because the presentation is typically causally or informationally linked to the target in some way.⁵ The representation's use is quite different: it is used as a counterfactual presentation. It is, in very rough terms, a model of the target which is used off-line to try out possible actions, so that their likely consequences can be assessed without having to actually try those actions or suffer those consequences.⁶ Second, and this is implicit in the first point, the ability to use an entity as an off-line stand-in depends crucially on its not being causally linked to, and its not necessarily carrying information about, the entity it represents.

This issue is actually quite touchy. Given my claim that a representation need carry no information about the target, two sorts of objections have been raised. First, to stick with the chess board analogy, the structure of the representational board must mirror the structure of the real board if it is to be of any use. As one person put it, "The second board,

⁴ A few famous examples include Dretske (1981), Fodor (1987), and Millikan (1984).

⁵ It is not my purpose in this article to examine presentations in any detail. I will continue to gloss their function as providing accurate information about the environment, even though I think that this is not quite correct. See Akins (1996) for a critique, with which I am sympathetic, of the view that the role of sensory systems is to provide accurate or veridical information.

⁶ This statement is reminiscent of Kenneth Craik's (1943) theory of representation and cognition. Indeed, major portions of my project, as expressed in Grush (1995, in preparation) can be seen as spelling out Craik's insights in more precision and detail.

of course, carries many sorts of information about the first board that are necessary for it to be a representation of the first board, including the board's structure and the kinds of pieces that can be involved."⁷ This is correct. If I might be allowed to distinguish two sorts of information, say *state* information and *structure* information, I want to say that representations need carry none of the first, but I admit they must have at least some of the second. The distinction would be roughly this: structure information tells one about the laws and generalizations which govern the operation of a system as well as the systems gross and relatively permanent features, while state information tells one about specific contingent features of a system, and hence which laws and generalizations are in effect. I will say a bit more about this in section 5, footnote 20. But for now I will note that when I discuss information, I mean state information. I feel justified in doing this because this is the way that the term is used by proponents of informational, covariational, and teleological theories of representation and content, and it is with these theorists that I am quarreling.

Second, it might be thought, even if we restrict the discussion to state information, the representation will carry information about the target. The fact that my black bishop is currently on the second board entails that it has not already been captured on the official board. Thus the second board does in fact carry information about the official board, in that knowledge of the state of the second board reduces the uncertainty about the state of the official board. But this isn't because of any constraints on representation, but is rather because in this case I only have need to represent certain classes of situation, namely those which I might encounter on the real board. Thus, given my purposes, and the fact that I'm not completely daft, the state of the second board will reduce the uncertainty about the state of the official board. This information is more properly a result of my purposes and not a result of requirements for representation, however. I can, after all, put my black bishop

⁷ This point was made by an anonymous referee for this journal.

back on the second board after it has been captured to determine where I went wrong, or if there were other moves I might have made to avoid the capture.

I will be arguing in due course that the proper way to understand the representational brain is to understand the interplay of three distinct sorts of entity; controllers (these correspond to the players which actually make and try out moves in the chess game analogy), presentations (these will typically be sensory or perceptual states), and internal representations (or emulators). With these distinctions in hand, many of the problems which have frustrated attempts to understand natural representation dissolve. But before we can apply these distinctions with any clarity, they must be refined.

Section 2. Emulation

The place to look for illumination on this topic is control theory. It will be useful to start with open-loop, or feed-forward, control. Such a system is shown schematically in Figure 1. The system includes a target system (often called the 'plant' in the control literature, thus I will use 'plant' and 'target system' interchangeably) that can be described as a set of system variables, some of which will be control variables, and some of which will be output variables, and perhaps others of which are neither. For instance, one might describe a car with a number of variables (amount of fuel, engine rpm, acceleration, mass, etc.), some of which are control variables (pressure on the gas pedal, torque on the steering wheel), and some of which are output variables (the position of the speedometer needle). The goal is to get some of the target's system variables to certain goal values -- one wants the car's speed to be between x and y mph, and its orientation to be exactly n degrees.



Figure 1. Open loop control.

The controller is a system whose purpose is to get the crucial variables of the target system to within their goal parameters -- the driver would usually be the controller of a car, for instance. Controllers typically need two pieces of information to do their job, the current state of the target, and a specification of the goals. (In order to get the car to 55 mph, one needs to know not only the goal speed, but the car's current speed as well. What action is appropriate will depend on both these factors.)

So in the simplest case, we can imagine the following sort of process. A controller is given a goal state and the current state of the target. From this information it determines an appropriate set of actions, perhaps an action sequence, which will get the target within the goal parameters. It issues these commands to the target. The target system, which starts in its initial state, then undergoes state changes as a function of the commands sent to it, and if everything is working correctly, the target ends up in the appropriate goal state. Once we notice that the entire system, controller plus target, implements an identity mapping (from goal states to goal states), we can characterize the controller as the inverse of the target (see Figure 2). That is (modulo the initial state specification), the controller implements a mapping from goal states ('x' in the figure) to command sequences (the 'y_i's in the figure), and the target implements a mapping from command sequences to goal states. The mapping

performed by the target is called the forward mapping, while that performed by the controller is the inverse mapping.

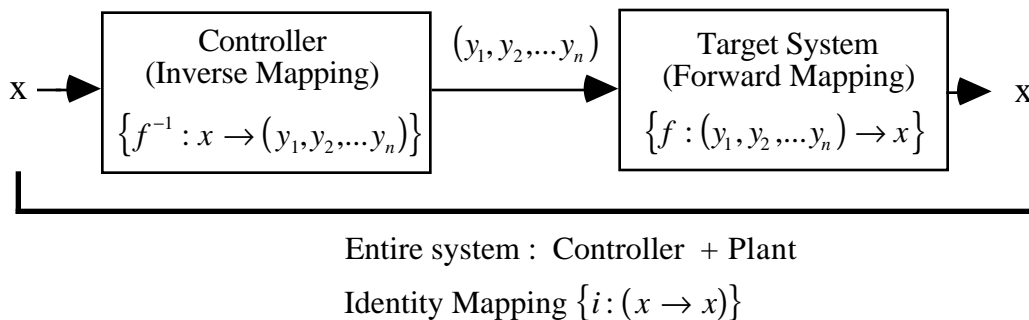


Figure 2. Forward and Inverse Mappings

Open-loop control has the virtue of being conceptually simple, and allowing us to introduce useful distinctions between types of mappings. Its usefulness as a control architecture is more questionable. *Closed-loop* control (also known as feedback control) is often more effective, flexible and efficient than open-loop control. In a closed-loop control system, there are sensors that are sensitive to various parameters of the plant, and these sensors feed this information back to the controller, which can then effectively change or continue its on-going command sequence in real time as needed (see Figure 3).

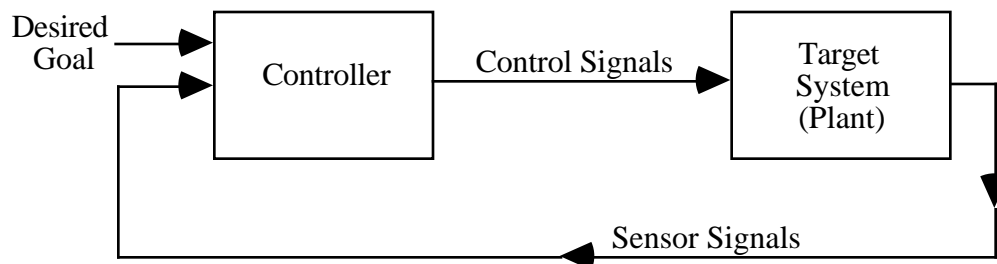


Figure 3. Closed loop control.

So for instance, a thermostat is a simple sort of closed-loop control system. The thermostat itself is the controller, and the plant in this case consists of everything else: the heater, cooler, thermometer, and the room. The thermostat sends commands to the plant (specifically, the heater and the cooler) in order to influence the temperature of the room, and it gets constant information (feedback) about the current state of the room, particularly its ambient temperature, from a thermometer. This frees the thermostat from having to determine the entire command sequence at once, as an open-loop controller would. Such an open-loop scheme would be brittle anyway, because the system would not be able to adjust the command sequence on the basis of unexpected perturbances. Feedback control allows the controller to produce and update the command sequence in real time as a function of its information about the progress of the plant. Because of this, in most cases, closed-loop control allows for much simpler controller design than open-loop systems do. For instance, the closed-loop thermostat can simply compare the current temperature with the desired temperature at each time step, and on the basis of that comparison, do one of three things; turn (keep) heater on, turn (keep) air conditioner on, turn (keep) both off.

Closed-loop control is often revered as the state of the art in control technology. Indeed, when the only competitor countenanced is open-loop control, the this conclusion is understandable. But closed-loop control has a number of problems as a control architecture, and it has an additional problem if one attempts to use it as a model of cognitive activity. One general problem with closed-loop systems is that they can be quite sensitive to feedback delays. To illustrate: suppose a thermostat is situated such that the information it receives concerning the temperature of the room it controls is, say, four hours old (perhaps the thermostat is controlling, via radio signals, the temperature of a

room on a space station some four light-hours away). If the thermostat is trying to get the room to 70 degrees, and is getting information to the effect that the room is 60 degrees, it will turn on the heat. In fact, because the feedback is delayed four hours, the thermostat will keep the heater cranked for four hours longer than it needs to -- by which time the room can be expected to be nice and toasty.⁸ At that point, the thermostat gets information to the effect that the room has reached 70 degrees. It turns off the heater, but then it continues to be told that the heat is rising. So it turns on the air conditioner. For the next four hours, the thermostat continues to be told that the temperature in the room is rising, so it keep the air conditioner on. Of course, the room will be below 70 degrees for 4 hours before the thermostat turns the air conditioner off. You get the idea. In general, depending on various parameters of the system, such as how much delay there is in the feedback, and how responsive the plant is to control signals, the plant may go into oscillations or instabilities as a result of such delays.

In addition to problems with closed-loop control itself, there is an additional problem arising from the attempt to use closed-loop control systems as a model for cognition.⁹ This is the *problem of obligatory action* -- a controller is only a controller when in closed-loop contact with the target system. When decoupled, like a detached thermostat or a Watt governor laying on the shop floor, controllers do *nothing*, and in particular they do nothing *cognitive*. Humans, on the other hand, seem able to do all sorts of cognitive things even when not actively engaged with a target system or interactive environment, contra some of the stronger claims issued from the embedded cognition camp.¹⁰ I can close my eyes and

⁸ I am here ignoring the fact that in this example the total delay in the control loop would actually be eight hours, because the command signal will take four hours to get to the room.

⁹ Timothy van Gelder (1995) urges us to think of the Watt governor, a closed-loop control system, as a model for cognition.

¹⁰ "...we must learn to think of an agent as containing only a latent potential to engage in appropriate patterns of interaction. It is only when coupled with a suitable environment that this potential is actually

plan the quickest route home after work when I learn about a traffic jam on my normal route. Additional examples of this sort of thing are, I imagine, easy to construct.

I do not intend to get into a discussion of the merits or demerits of looking at matters this way, but rather to render the entire debate obsolete by explaining exactly how systems represent targets even when not coupled to them; that is, how cognitive systems avoid the problem of obligatory action by being able to act on representations in lieu of acting on external environments. The key to this is to recognize that a *cognitive system* is not just a controller, but a controller together with a forward model (or as I shall call it, an *emulator*).

To see what a forward model is, consider the following solution to the feedback delay problem faced by the thermostat. Suppose that before the space station was sent away, that is, before there was a feedback delay problem, researchers decided to train a neural network to mimic the forward mapping of the plant. In this case, the plant is the room, its heating and cooling systems, and its associated sensors (the thermometer). Inputs to the plant are commands to the heater and cooler, and outputs are thermometer readings. The training proceeds by letting the thermostat control the temperature of the room, and at every time interval (preferably very short), the neural network is given as input the current temperature and current command, and the training signal is the temperature which is produced at the next time step. After training, the neural network will exactly implement the forward mapping. It is thus called a forward model, or *emulator*.

Once one has an emulator, it is possible to implement pseudo-closed-loop control (see Figure 4). Here, the controller's output is split into two copies. One is sent to the plant, as

realized through the agent's behavior in that environment." Beer (1995). Amazingly, Beer (and many others) see the problem of obligatory action not as a problem, but as a premise on which to base a call for a reconception of cognitive activity.

usual, and the other is sent to the emulator. Because the emulator implements the same input-output function as the plant, the controller can use feedback supplied by the emulator just as well as it could use the real feedback from the plant itself. Once this happens, it will no longer matter if the real feedback signal is delayed -- it's not being used anyway. Provided that the emulator is accurate, and that *its* output is not delayed, all will proceed without a hitch.

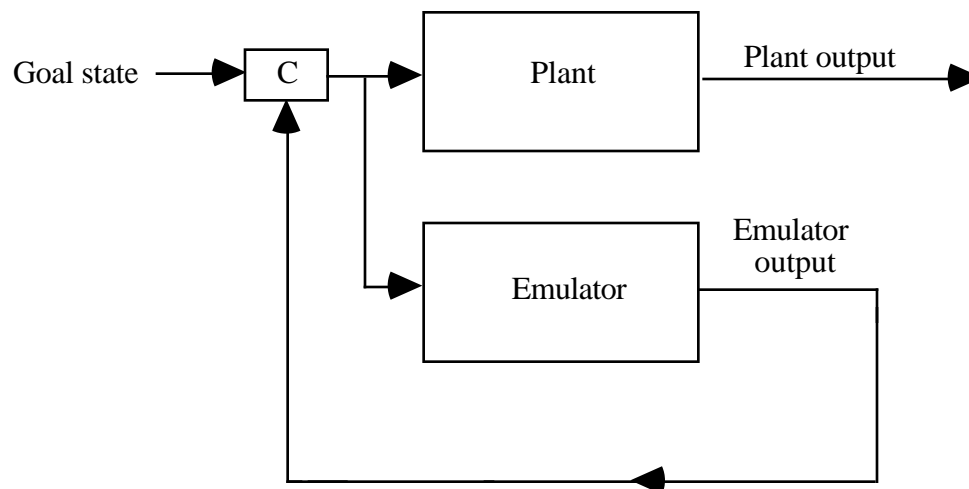


Figure 4. Pseudo-closed loop control.

Of course doing this introduces other problems. If the emulator is not perfect, for instance, and there is no means of keeping the state of the emulator close to the state the plant is in, it will eventually wander, and then the control signals are likely to be inappropriate. But for the present conceptual point we can safely ignore these issues. The important point is the notion of an emulator, an entity which implements the forward mapping, which is plug-compatible with the plant, and can thus be run in parallel with the plant (as in pseudo-closed loop control) or can be run *instead of* the plant (examples of this will be provided shortly).

This brief exposition of a few control architectures should allow us to gain clarity on the distinction between presentations and representations. A presentation, as I shall use the term, can be thought of as information about the current state of whatever actual system one is dealing with. This will often be sensory information. A representation, by contrast, is an emulator of some system, which can be run off-line to provide mock sensory information about what the real target system would do under various conditions. It is, to belabor the point, something which is used to stand for something else. If this is correct, then we can say that a *necessary* condition for a system being a representing, cognitive system is that it has the capacity to internally emulate some external system with which it can also interact overtly.

Section 3: An Example from Robotics

I have distinguished representations from presentations, and attempted to spell this difference out in terms of control loops -- a presentation is perceptual information from a target, while a representation is an internally maintained forward model of the target, which can supply mock information for various purposes, including counterfactual reasoning. But an example can be worth a thousand definitions.

Murphy (Mel, 1990) is a robot whose task is to maneuver its arm around obstacles in order to get its hand to an object which it will then be in a position to grasp. Murphy's arm has three joints, a shoulder, elbow, and wrist, all of which have their range of motion limited to a single plane. Above this plane a video camera whose signal is used to drive a 64 x 64 grid of units which act as a sort of visual display -- a 64 x 64 grid of pixels. This grid of units is what Murphy uses to guide its arm movements around the workspace without contacting obstacles.

Each of the units of the grid is actually a connectionist unit whose activity is normally driven by the video camera. But each of these units also receives inputs from the robot's motor output.¹¹ That is, the motor output is split into two copies, one of which normally gets sent to the arm, and a second which gets sent to the grid units. Each unit then gets information from two sources, visual information from the video camera, and information concerning the motor commands which Murphy issues.

¹¹ In Murphy's case, the motor output is a joint angle configuration, and not an effector command. The forward mapping Murphy learns is thus the forward kinematics, and not the forward dynamics. This is not of consequence to the present point, however.

During the first phase of Murphy's operation, it simply moves its arm through a sample of arm configurations. During this time, the grid units monitor both the motor commands sent to the arm, as well as the resultant video input to the grid which drives the units. Because these units see both the input and the output of the forward mapping which the arm/camera system implements -- it gets a copy of the motor command and is fed with a visual image which that command leads to -- they are thus able to learn the forward mapping of the plant. In this case, the forward mapping is a function from joint angle commands to visual grid displays. The *plant* implements this mapping by actually sending the joint angle command to the arm, which moves accordingly, and the video camera sends an image of the resultant configuration to the grid. But once the units have learned the forward mapping, they can implement this function without going through the real arm/camera system. Thus, with the real arm and the video camera off-line, the grid is able to process a copy of the motor command to create a mock visual display which is similar to the display which would have been produced by the video camera pointed at the arm as it carried out those same motor commands.

And this is exactly what Murphy does after learning. It gets an input of a visual scene including its arm's initial location, the object location, and obstacle locations. It then takes the real system (its arm and the video camera) off-line, and drives the visual grid, which is an emulator, directly with an efferent copy of the motor command. Murphy is then able to create an image of its arm moving around the work space. When it sees its arm (now an image of its arm) run into an image of an obstacle, it backs up and tries again, just as it would when actually operating its arm. It continues this way until it discovers a route from the initial arm configuration to the goal location. It then puts the arm and camera system back on line, and implements this solution.

According to the terminology I introduced earlier, the visual grid, when on-line and being driven by the actual arm/camera system, is a *presentation* of the workspace. When the arm/camera system is taken off line and the grid is being used as a forward model driven by an efferent copy, the visual grid is a *representation* of the workspace.

But we can go further than this. What the forward model does is allow Murphy to engage in *counterfactual reasoning*. It can determine that if it were to issue joint-angle command sequence R, its arm would impact an obstacle, but if it were to implement joint-angle command sequence S, its arm would move to the goal without impact. At the end of this reasoning process, it implements the correct sequence. And even though the system is a relatively simple one whose operation could be described in terms of the flow of activation in connectionist units, it is also easy to see how we can provide representational explanation of its operation. Why did Murphy **not** move its arm to location X? Because it found(/believed) that if it did, its arm would hit an obstacle. What is Murphy thinking about now? It is contemplating the use of command sequence R. No doubt my use of the terms 'believe' and 'contemplate' are a bit strained with such a simple system. But I hope that doesn't divert attention from the fact that some sort of representational description is almost forced by Murphy's use of the emulator to stand for the real system.

So this, in brief, is the Emulation Theory of Representation (ETR). A representation is something which is used to stand for something else. And this amounts to a controller which normally operates some system being able to decouple from that system, and couple instead to a plug-compatible emulator. When it does this, the emulator is standing for the real system. It represents it. Not because it is causally linked to the real system. Not because it carries information about the state of the real system. But because it is used to stand for it.

Section 4: The Human CNS

We have before us a starting point for a theory of internal representation and an example of a robot which implements the architecture described in the theory. Now I will turn to the human nervous system and provide evidence that it uses emulators to represent entities external to it. I will discuss two sorts of case; emulators of the musculoskeletal system (MSS) and emulators of the motor-visual loop.¹²

We saw in section 2 how feedback delays can cause problems for closed-loop control systems. There is considerable evidence to the effect that, in part because of relatively slow axonal conduction velocities,¹³ proprioceptive feedback from the limbs is too slow to be used to guide fast voluntary movements which are executed in less than about 200 - 450 ms.¹⁴ This might lead one to suspect that such movements are executed open loop. But this seems not to be correct either, as there is evidence that there are adjustments made to the

¹² Much more detail on these types of emulators and others, and more evidence for them, can be found in Grush (1995; in preparation).

¹³ The muscle spindle and Golgi tendon organ signals, which constitute the major proprioceptive mechanoreceptors, are Ia afferents, which are large myelinated fibers having the fastest conduction velocities of all afferent axons. Such fibers are fast *compared to other types of axons*. But the delay is not exclusively a product of the axonal conduction velocity, but of synaptic relays, and central processing of the signal, as well as delay introduced from the efferent side of the process. Thus the feedback delay is the total delay in the loop. I will continue to gloss this total delay, somewhat inaccurately, as due to 'slow axonal conduction velocities'.

¹⁴ Cf. Dennier van der Gon, J.J. (1988), Ito (1984). There is considerable debate as to the exact delay involved in the proprioceptive loop, most proposals seem to indicate it as being between 200 and 450 msec. The fastest latency for human arm motions I have ever seen defended in print is 125 msec, and even this is longer than the 70 msec which is when adjustments seem to be made to the motor sequence (see below). The usual paradigm for testing feedback delays is the tendon vibration technique (see Redon et al. (1991)), where a vibrator is placed on a tendon, causing the muscle spindles to misjudge the joint angle. Even with proprioceptive information being distorted, there is no change in the details of a given movement, as compared to a non-vibration control, in less than about 200 - 450 msec, which implies that proprioceptive information cannot influence the motor program in less than that amount of time.

motor program, as quickly as 70ms after movement initiation, which have the effect of correcting initial inconsistencies in the first part of the trajectory.¹⁵

This apparent paradox -- that it looks as though there are corrections made to the motor program on the basis of peripheral feedback before peripheral feedback can be used effectively -- is dissolved when one realizes that the feedback used to make adjustments to the motor program could be supplied by an internal emulator. One way to do this would be via the pseudo-closed-loop architecture described in section 2. In other words, when the movement must be executed too quickly for peripheral feedback to be of use, an internally generated 'mock' proprioceptive signal, generated with a musculoskeletal emulator, can be used to adjust the motor program instead.

A number of researchers¹⁶ have developed models based on this insight. According to Kawato (1990), there is a neural circuit involving the cerebellum (especially the dentate nucleus) and the red nucleus which acts as a model of the MSS. The models of Wolpert et al. (1995), and Gerdes and Happee (1994) are more sophisticated, positing not a simple pseudo-closed loop architecture, but rather variants of Kalman filters which use forward models. In the model of Wolpert et al., to a first approximation, this filter combines the deliverances of the real target and the forward model in a time-varying manner, so that the motor centers rely exclusively on the output of the forward model during the initial phases, and as time progresses rely more and more on the 'real' proprioceptive information.

Whatever the details, there is converging experimental and theoretical evidence from human

¹⁵ Cf. van der Meulen et al. (1990).

¹⁶ E.g. Kawato (1990), Wolpert, Gharamani and Jordan (1995), Kawato, Furukawa and Suzuki (1987), Gerdes and Happee (1994).

motor performance which indicates that the human CNS in fact uses internal forward models of the body.

It should be noticed that this solution to a motor control problem has an unexpected benefit: the tools nervous systems evolved to solve this problem can also be used to generate motor imagery if the MSS is simply taken off-line. Motor imagery is no more than an internally generated proprioceptive image (like visual imagery, only not visual), and this is exactly what the forward model generates in order to solve the feedback delay problem. Consistent with this hypothesis is the finding that many of the same motor areas which are used to drive overt motor behaviors are also active during motor imagery.^{17 18}

It is possible to address visual imagery with the same architecture. The account of motor imagery I sketched exploited the fact that there are regularities in the mapping from initial proprioceptive states and motor commands to future proprioceptive states, and thus that an emulator driven off-line by efferent copies can produce proprioceptive imagery. The same is true for visual imagery. Given a visual input (retinal projection, primary visual cortical projection, whatever) and a motor command (such as a saccade, or a step forward), the next visual input is at least in part predictable (e.g. a translation in the direction opposite the saccade, or an enlargement of the projection of the object one is walking towards). Mel

¹⁷ Decety, Sjöholm et al. (1990); Roland, Larsen et al. (1980); Fox et al. (1987); Ingvar and Philipsson (1977).

¹⁸ To pick out one of a hundred examples of converging evidence: Vilayanur Ramachandran (personal communication) has found that most phantom limb patients fall into one of two groups; those who can voluntarily control their phantom limbs and those who cannot. It turns out that in almost all cases, those who cannot move their phantoms experienced a period of pre-amputation paralysis, while those who can move their limbs did not. On the present theory, this is to be expected. If the musculoskeletal emulator learns and updates the forward model by monitoring the operation of the musculoskeletal system, then when there is a period of paralysis, the mapping learned is that no matter what the motor command is, the proprioceptive result is 'no movement'. When the amputation occurs without a pre-operative paralysis period, there is no information to contradict the operation of the emulator, and hence no reason for the forward model to change (keep in mind the crucial difference between i) proprioceptive information to the effect that there was no movement, and ii) no proprioceptive information about any movements).

(1986) has developed a model of mental rotation, zoom and pan using this insight. The model is a robot that can move around, towards, and away from objects while looking at them, has the ability to learn this forward mapping, and can then take itself off-line, and 'mentally' rotate, zoom and pan images of objects.

In the human case, there is increasing evidence that many of the same areas that subserve vision are also implicated in visual imagery, that is, evidence that these areas are driven not only by sensory organs, but can also be driven by internal efferent copies, much like Murphy. To take but one example of dozens, Martha Farah (Farah et al., 1992) conducted a series of imagery experiments on a patient before and after unilateral occipital lobectomy. The subject was asked to imagine a number of common objects (ruler, car, bicycle, etc.), and to imagine them getting closer and closer until they just began to overflow the edges of the visual image boundary. The fascinating result was that after the unilateral occipital lobectomy, the distance at which imagined objects began to overflow the boundary increased dramatically for objects which were primarily horizontally oriented, but did not significantly alter for vertically oriented objects. One plausible explanation for this finding is that visual imagery is the result of the use of a forward model which drives (at least some of) the same visual areas driven by overt vision. After the removal of one occipital lobe, the forward model has a screen, so to speak, only half as wide to drive. Imagined horizontal objects will accordingly be farther away when they begin to push the edge of the mind's eye.

Section 5: Articulation

Even though the theory of representation I have outlined here may be a promising starting point for understanding natural cognition, there are a number of serious inadequacies as it stands. While addressing all of them is beyond the scope of this article, in this section I will address one which is crucial.

It will have been noticed that when I have talked about ETR to this point, I have claimed that the emulator represents the target system (e.g. the musculoskeletal emulator (MSE) represents the MSS). But even though this is representation, it is a weak variety. While this might perhaps work for imagery (which is what I have primarily been discussing), it is not clear that it will work for representation in general. This is because we have one entire system which represents, in some holistic way, some different system. What might be preferable in some cases is to have entities which represent, individually, components of the target system. So for instance, given that specific physical parameters of joints and muscles are relevant components of the MSS for purposes of its movements, one would like to be able to point to entities in the CNS which represent, say, specific physical parameters of the forearm, or the index finger. A similar problem is manifest with Murphy -- it has a representation of its workspace (the visual grid when run off-line), but there is no distinguishable entity which has the dedicated task of representing Murphy's hand. As Murphy imagines a movement, first some units will be active, then those will spin down as others become active (this is what movement across the grid amounts to), and this is about as close as one is able to get to a representation of the hand.

But suppose that the emulator not only implements the forward mapping, but does so because it is composed of parts, or *articulants*, each of which represents some aspect of the target system. A target system will typically be describable as a dynamical system with N parameters, e_1, e_2, \dots, e_n . Some of these, let us say e_1, e_2, \dots, e_i will be input parameters, that is, parameters whose equations of evolution include one or more parameters of the controller. Others, say $e_{i+1}, e_{i+2}, \dots, e_k$, will be output parameters, that is, parameters of the target which directly influence the evolution of one or more of the parameters of the controller. The rest of the target system's $N-k$ parameters we can call internal parameters -- these will neither be directly influenced by, nor will directly influence, any parameters of the controller. Now let us suppose that we have an emulator which implements the forward mapping of the target because it also is an N parameter dynamical system, with $e^*_1, e^*_2, \dots, e^*_i$ as inputs, $e^*_{i+1}, e^*_{i+2}, \dots, e^*_k$ as outputs, and the rest of the $N-k$ e^* s as internal parameters, just like the real target (the first two conditions guarantee that the emulator and target are 'plug compatible'). Suppose further that the dynamic of the emulator is formally equivalent to the dynamic of the target.¹⁹ In such a case, not only does the emulator represent the target system, but it will also be true that the emulator articulant e^*_h represents target system parameter e_h . In such a case we can say that the emulator is articulated into components which represent components of the target system.²⁰

¹⁹ The same equations of evolution are obtained by replacing the e s with e^* s, the only difference being what the parameters are physically parameters of, but that need not effect the dynamic. E.g. $de_n/dt = -(k/m)e_m$ might be the dynamic for a mass-spring system (where e_m is displacement and e_n is velocity), while $de^*_n/dt = -(k/m)e^*_m$ might be the dynamic of a neural oscillator, where e^*_m and e^*_n are firing frequencies of appropriately coupled neurons. For a more detailed example of this, as well as evidence to the effect that the human motor emulator is articulated in this way, see Grush 1995, chapters 3 and 4.

²⁰ Moreover, it is arguably the case that the articulated emulator is best thought of as a *theory of the target domain*. Doing so allows me to make contact with some of Paul Churchland's views, especially as expressed in Churchland (1989). I address the comparison of my views and those of Churchland to some degree in Grush (1995), chapter 4. Finally, I can now make clear the distinction, mentioned in Section 2, between state information and structure information. The articulated emulator carries structure information about the target because it implements the same dynamical system as the target. The emulator's structure in such cases then carries information about the target's structure. But the emulator carries no state information about the target, because, in general, one can set the emulator's state (the specific values of the e^* s) to any

This has several advantages. First, it makes semantics much easier. To the question What represents this aspect of the target system? there will be a clear answer. Second, and more importantly, this gets us out of the regress problem.²¹ A cognitive system (a system which not only interacts with its environment, but consists of a controller and an articulated environmental emulator to which it can couple to represent its environment) can have an internal representation of some feature of its environment F . What makes the emulator articulant F^* represent F is the fact that the controller interacts with the emulator, and the emulator's articulant F^* , in a way analogous to the way it interacts with the environment, and the environmental feature F . Furthermore, the controller does not need to *interpret* anything *as* anything else, except in the innocuous and non-question begging sense that it interacts with something *as* it interacts with something else. It simply enters into dynamical interaction with one system rather than the other. Representation is cashed out in terms of use, and use is cashed out in terms of selective dynamical coupling. We are thus not forced into a representational regress, antirepresentationalism, mere instrumentalism about representational descriptions; nor does the problem of obligatory action arise.

Furthermore, we are in a position to see one clear criterion that distinguishes genuinely cognitive systems from merely complex and adaptive, but non-cognitive, systems. It is a necessary condition for a system to be a cognitive system that it consist not only of a controller (which may, by itself, be able to interact adaptively with the system's

value whatsoever to see what the target system would do if its e s had those values, and there need be no correlation whatsoever between the values of the e^* s and the e s.

²¹ By 'regress problem' I mean a common objection to theories of representation that maintain that a representation requires a user. The objection is that it seems that the user must interpret the representation to be a representation of something, which is tantamount to saying that the user must be able to represent in order to treat other entities as representations. The user then becomes a homunculus, whose capacities for interpretation are unexplained.

environment), but also that the system be able to internally emulate aspects of its environment, via use of a forward model, in such a way as to allow the controller to selectively decouple with the real environment and couple instead with the internal emulator.

Section 6: Objections and Replies

1. *Why do we need to describe such systems as representing anything? Rather, can't we just explain the behavior of Murphy, for example, as the operation of a complex dynamical system?*

This objection seems to presuppose that only in cases where one is *forced* to describe a system in representational terms is one licensed to legitimately do so at all. But I can't see why anybody should be moved by this. Indeed, any materialist will agree that there are, in principle, physical explanations for all behaviors -- but this is not seen as *prima facie* incompatible with legitimate representational explanation. The best criteria for the legitimacy of a representational explanation are whether or not such explanations make sense of the behavior of the system, and whether or not it is useful to interpret aspects of the system's internal antics as cognition about something else. In Murphy's case, to take the example at hand, both of these criteria are met. A representational account of what Murphy is doing when it goes off-line is natural and, in some sense, just plain right.

2. It has not been established that the motor control systems of the human CNS use motor emulators, as your theory suggests they do.

This objection misunderstands my rhetorical strategy. The existence of musculoskeletal emulators is not a *premise* on which the larger argument is based, but merely an example of an application of the architecture. It may in fact turn out that motor control does not use such emulators. It would not follow from this that the emulation theory of representation is wrong, but only that fast voluntary motor control does not employ representations.

Nonetheless, there is very compelling evidence that the human CNS does use such models (a small sample of which is cited in section 4). And the speculation that nervous systems do use such models for motor control purposes may shed light on the evolution of cognition (see section 7).

3. Even if ETR works for imagery, it is not at all clear how it is meant to work for other aspects of cognition.

It is not known to what extent cognition is in fact based on imagery, but it could be a much greater extent than is often supposed. To point out but a few examples: much current work in cognitive linguistics²² seeks to explain many key aspects of language in terms, in part, of imagery and image schemas. There are likewise rich connections between imagery and memory, suggesting that memory may not store information in an amodal propositional form, but rather as modality-specific images (see Paivio 1995 for review). Furthermore, given that emulators are just neurally implemented models, the door is open to accommodating many of the insights of Johnson-Laird's (1983) Mental Models framework for reasoning and inference.

²² Cf. Langacker (1987, 1990, 1991), Lakoff (1987).

How much of cognition can be accounted for in terms of imagery or image schemas is an open empirical question. Until such questions have been answered, the fact that emulators need not traffic in propositions or first order predicate calculus is no strike against them.

4. *It does not appear to be the case that the representations posited by ETR will be such as to exhibit well-known properties of representational content such as failures of substitution.*

Actually, one of the more exciting precipitates of the current account is that it does in fact address this issue, but again, I can only sketch this roughly.²³ Emulators do not represent the target system *per se*, but represent it *as interacted with*. Any given physical object or system actually implements an infinite number of dynamical systems, and it is possible to interact with a physical system *as* any number of these systems. So for instance a computer can implement a dynamical system which describes a Turing machine, as well as implementing a dynamical system describable as a mass-spring (imagine here the computer attached to a wall with a spring, and oscillating back and forth). I could then interact with the computer either as a Turing machine, or as a mass-spring system (I might program it, I might try to predict when it will reverse direction). The articulated emulator, *qua* emulator, explicitly implements only one dynamical system,²⁴ and so it is only that aspect which is represented. My musculoskeletal emulator represents my arm, to a rough approximation, as *the body part with such-and-such physical/dynamical properties*, and not as *the last item seen by Smith before he went unconscious* or as *the twentieth heaviest*

²³ I discuss opacity phenomena in more, though still greatly inadequate, detail in Grush (1995), chapter 7. Initial appearances notwithstanding, the treatment there is compatible with, and in some sense presupposes, the remarks of this section.

²⁴ The *qua* clause is important, because of course the emulator itself also implements an infinite number of dynamical systems -- my brain could be attached to a wall with a spring, for instance. But this mass-spring dynamical system is not an emulator used by my brain because it does not have the appropriate coupling parameters with my motor centers.

object in the room. Emulation is necessarily emulation of some aspect of the target, where aspects are, in the first instance, individuated by use and means of interaction and prediction.

5. *'Standing for' is not a relation distinctive of representations, because signals in the sensory/presentational systems also stand for something else, namely the current state of the target.*

But this is exactly what I am trying to get away from. The distinction between representation and presentation is a distinction between *standing for* and *providing information about*. Perhaps it is easier to grasp this distinction if one distinguishes *standing IN for* and *providing information about*. If I am on a football team and get injured, another player can take my place -- can, so to speak, stand in for me. There is all the difference in the world between this, and simply providing information about me, e.g. via a video camera, to the undermanned team. So much difference that I am frankly surprised at how often this distinction is not grasped. In order for X to stand for Y, X must do something that Y would do, and be doing it instead of Y. Healthy replacement players can do this, information about injured players cannot. Similarly, *sensory states do not stand in for the objects in any sense*. A certain pattern of activity in my primary visual cortex, say pattern T, when I look at a tree is not at all standing in for a tree, because pattern T is not doing anything which is normally done by a tree.

7. Conclusion.

In *Neurophilosophy*, Patricia Churchland writes:

To follow evolution's footsteps in discovering how basic principles of motor control are refined and upgraded to yield more complex systems is a productive strategy. Additionally, it may be a shift in focus that allows us a breakthrough in the attempt to understand the higher functions... If we can see how the complexity in the behavior that we call cognition evolved from solutions to basic problems in sensorimotor control, this can provide the framework for determining the nature and dynamics of cognition.
(p. 451)

This is exactly what I hope to have done. If the theory I have very tentatively sketched here is correct, then representation and cognition are dependent on emulation, and the strategy of emulation plausibly may have arisen as a solution to a clear problem in motor control.

Given fixed and relatively slow axonal conduction velocities, the twin evolutionary pressures of greater size and greater speed work at cross purposes. Pseudo-closed-loop control (or some other similar strategy employing a forward model, such as Wolpert et al.'s (1995) Kalman filter) provides a straight-forward solution to this problem, and requires only relatively humble ingredients: an efferent copy of the motor command, and some way to do associative learning. And once this simple solution is in place, nervous systems have a powerful new tool. A tool which makes possible imagery, representation, and cognition itself.

Perhaps we should be grateful that axons are as slow as they are.

Acknowledgments:

I would like to thank the following people for helpful discussion and feedback: Robert Hecht-Nielson, Patricia Churchland, Paul Churchland, Andy Clark, Adrian Cussins, Vilayanur Ramachandran, Bob Gordon, George Lakoff, Jordan Hughes, Pete Mandik, Joe Schear, Chase Wrenn, Chris Eliasmith, the members of Patricia Churchland's Experimental Philosophy Lab, the participants of the 1996 UC Berkeley Summer Research Seminar, and an anonymous referee for *Philosophical Psychology*. I am grateful to the McDonnell Foundation and the Philosophy-Neuroscience-Psychology Program at Washington University in St. Louis for supporting this research.

References:

- Akins, Kathleen (1996). Of sensory systems and the 'aboutness' of mental states. *Journal of Philosophy* 93(7):337-72.
- Beer, Randall (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence* 72:173-215
- Churchland, Patricia (1986) *Neurophilosophy*. MIT/Bradford.
- Churchland, Paul (1989) *A neurocomputational perspective*. MIT/Bradford.
- Craik, Kenneth (1943) *The Nature of Explanation*. Cambridge University Press.
- Cummins, Robert (1996) *Representations, targets and attitudes*. MIT/Bradford.
- Decety, J., Sjolholm, H., Ryding, E., Stenberg, G., and Ingvar, D. (1990). The Cerebellum participates in Cognitive Activity: Tomographic measurements of regional cerebral blood flow. *Brain Research* 535: 313-317
- Denier van der Gon, J.J. (1988). Motor control: Aspects of its organization, control signals and properties. in Wallinga et al. eds. *Proceedings of the 7th Congress of the International Electrophysiological Society*. Amsterdam: Elsevier Science Publishers
- Dretske, Fred (1981) *Knowledge and the flow of information*. MIT/Bradford.
- Farah, Martha, Soso, Micheal J., and Dasheiff, Richard M. (1992) Visual angle of the mind's eye before and after unilateral occipital lobectomy. *Journal of Experimental Psychology: Human Perception and Performance* 18(1):241-246.

Fodor, Jerry (1987) *Psychosemantics*. MIT/Bradford.

Fox, P.T., Pardo, J.V., Petersen, J.V. and Raichle, M.E. (1987). Supplementary motor and premotor responses to actual and imagined hand movements with positron emission tomography. *Neuroscience Abstracts* 398(10):1433.

Gerdes, V.G.J., and Happee, R. (1994) The use of an internal representation in fast goal-directed movements: a modeling approach. *Biological Cybernetics* 70:513 - 524

Grush, Rick (1995) *Emulation and Cognition*. PhD Dissertation, University of California, San Diego. At URL <http://www.artsci.wustl.edu/~rgrush/>

Grush, Rick (in preparation) *The worlds within: the neural construction of mind, language, and reality*.

Ingvar, D. and Philipsson, L. (1977). Distribution of the cerebral blood flow in the dominant hemisphere during motor ideation and motor performance. *Annals of Neurology* 2:230-237

Ito, Masao (1984). *The cerebellum and neural control*. New York, Raven Press.

Johnson-Laird, Philip (1983) *Mental Models*. Harvard University Press.

Kawato, Mitsuo (1990). Computational schemes and neural network models for formation and control of multijoint arm trajectories. in Miller, W.T., Sutton, R.S. and Werbos, P.J., eds. *Neural Networks for Control*. MIT/Bradford.

Kawato, Mitsuo, Furukawa, K. and Suzuki, R. (1987). A hierarchical neural network model for control and learning of voluntary movement. *Biological Cybernetics* 57:447-454.

Lakoff, George (1987) *Women, Fire and Dangerous Things*. University of Chicago Press.

Langacker, Ronald (1987). *Foundations of Cognitive Grammar (Volume I)*. Stanford University Press.

Langacker, Ronald (1990). *Concept, Image and Symbol*. Mouton de Gruyter.

Langacker, Ronald (1991). *Foundations of Cognitive Grammar (Volume II)*. Stanford University Press.

Mel, Bartlett (1986) A connectionist learning model for 3-d mental rotation, zoom, and pan. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, 562-571. New York: Erlbaum Associates.

Mel, Bartlett (1990) Vision-based robot motion planning. In W. Thomas Miller III, Richard S. Sutton and Paul Werbos (eds) *Neural Networks for Control*. MIT/Bradford.

Millikan, Ruth (1984) *Language, thought, and other biological categories*. MIT/Bradford.

Paivio, Alan (1995). Imagery and Memory. In Gazzaniga (ed.) *The Cognitive Neurosciences*. MIT/Bradford.

Redon, Christine, Hay, Laurette, and Velay, Jean-Luc (1991) Proprioceptive control of goal-directed movements in man, studied by means of vibratory muscle tendon stimulation. *Journal of Motor Behavior* 23(2):101-108.

Roland, P.E., Larsen, B., Lassen, N.A., and Skinhoj, E. (1980). Supplementary Motor area and other cortical areas in organization of voluntary movements in man. *Journal of Neurophysiology* 43: 1:118-136.

van der Meulen, J.H.P., Gooskens, R.H.J.M., Dennier van der Gon, J.J., Gielen, C.C.A.M., and Wilhelm, K. (1990) Mechanisms underlying accuracy in fast goal-directed arm movements in man. *Journal of Motor Behavior* 22(1):67-84.

van Gelder, Timothy (1995) What might cognition be, if not computation? *The Journal of Philosophy* 91(7):345-381.

van Gelder, Timothy, and Port, Robert (1995) It's about time. Editors' introduction to Robert Port and Timothy van Gelder (eds.) *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT/Bradford.

Wolpert, Daniel, Ghahramani, Zoubin, and Jordan, Micheal (1995). An internal model for sensorimotor integration. *Science* 269:1880-1882.