Philosophical Issues
A Supplement to NOÛS

Philosophical Issues, 26, Knowledge and Mind, 2016
doi: 10.1111/phis.12079

SIMULATION AND UNDERSTANDING OTHER MINDS

Sherrilyn Roush
King's College London

There is much disagreement about how extensive a role theoretical mind-reading, behavior-reading, and simulation each have and need to have in our knowing and understanding other minds, and how each method is implemented in the brain, but less discussion of the epistemological question what it is about the products of these methods that makes them count as knowledge or understanding. This question has become especially salient recently as some have the intuition that mirror neurons can bring understanding of another's action despite involving no higher-order processing, whereas most epistemologists writing about understanding think that it requires reflective access to one's grounds, which is closer to the intuitions of other commenters on mirror neurons. I offer a definition of what it is that makes something understanding that is compelling independently of the context of cognition of other minds, and use it to show two things: 1) that theoretical mind-reading and simulation bring understanding in virtue of the same epistemic feature, and 2) why the kind of motor representation without propositional attitudes that is done by mirror neurons is sufficient for action understanding. I further suggest that more attention should be paid to the potential disadvantages of a simulative method of knowing. Though it can be more efficient in some cases, it can also bring vulnerability, wear and tear on one's personal equipment, and unintended mimicry.

Introduction

The view that we know the beliefs, desires, intentions, and feelings of other people by, in some sense, simulating or replicating them has been around for a long time, but in cognitive science in the 1980s it began to be developed more fully as a rival to the view that we know each other's mental states by interpreting them using concepts and theories. There are many versions of this *simulation* view of our mindreading capacities, but broadly speaking on such a view we know each other's minds primarily by running through a similar process in our own mental equipment to what is going on in theirs, and seeing what mental states and behavior we ourselves would have. On a *theoretical* mind-reading view, we know each other's minds primarily by having beliefs about others' behavior and mental states to which we apply a theory of mind and reason to conclusions about other mental states and behavior.[1] It is clear that we are capable of theoretical mind-reading, but the plausibility of our at least sometimes also simulating was secured with the discovery of mirror neurons at the end of the 20th century. This phenomenon gives concrete evidence that in some cases we actually do acquire a kind of replica of the other person's mental states in our own mental states, and that we do this not only in imagination but at the level of our brains, in body-coded patterns of firings of the same neurons that fire when we do or feel the thing we witness the other doing, a phenomenon called *embodied simulation* (Gallese and Goldman 1998, Gallese, Keysers, and Rizzolati 2004, Gallese and Sinigaglia 2011, Goldman 2012, Casile 2013).

This firing of motor neurons is the simulative response that stands in the most extreme contrast to a theoretical response, since motor neurons definitely do not carry out reflection and their firings are not believings or other kinds of propositional attitudes using concepts. This leads to the question just how much understanding can be achieved through them alone. If we get something epistemic out of matching that is this devoid of thought then simulation is powerful indeed, but how could it be so?

Mirror neurons display three properties that I take to be required for simulation in this context (Cf. Goldman 2006). One is that a simulation executes a process, another that the process is dynamically isomorphic to the process undergone by the person being evaluated, hereafter the actor, and third that the process is a re-deployment of the mechanism that serves the mind-reader when she herself behaves along the dimensions relevant to the behavior of the actor. I will argue that the first two features, which are shared by theoretical mindreading, are what makes either method able to deliver understanding. The qualitative nature of the dynamically isomorphic states the understander possesses—whether beliefs or primitive firings—and whether the process is a redeployment or not are what I take to distinguish simulation from theory. These features are not per se relevant to whether or how much understanding is delivered, but they are pragmatically relevant to how efficiently and accurately understanding is achieved in a given case, and with how much risk to the mind-reader. As I will explain below, the simulator may think less and respond faster with the same accuracy in some kinds of cases, but he also has "skin in the game" and this can carry costs that the theorizer avoids.

Social Cognition as Knowledge

Many think of simulation as a process, one that takes "pretend" behavior and mental states as input and by grinding these inputs through a person's own belief-desire mill, produces "pretend" mental states and behavior as output, output that is then projected onto the mind one is trying to understand (Goldman 2006). As a process, simulation lends itself well to the application of Alvin Goldman's epistemological theory of justified belief as a belief that resulted from a reliable process, that is, a process that produces true beliefs most of the time. Simulation can be a way of gaining such justified beliefs about other minds on the assumption that human beings' belief-desire mills are similar to each other, despite wide variation in the beliefs and desires themselves. If our mills are similar enough, and I put relevantly similar beliefs, desires, and intentions as input into my mill as are input in yours, then I will usually come up with accurate beliefs about your beliefs, desires, intentions and behavior. If this process does get it right most of the time then that means, on Goldman's view, that each one of its verdicts is a justified belief.

Reasoning with a theory, and inferring from past to future behavior in more primitive ways by association, are also processes, and Goldman's view of justified belief applies in the same way to them. Insofar as the theoretical assumptions we make about other minds are true, and we accurately observe other minds' behavior, infer their mental states from it, and appropriately apply the theoretical assumptions to it, the beliefs that we get about their future mental states and behavior will also be accurate. That in this way we get it right most of the time means that every occasion of it yields a justified belief. Goldman's concept of justified belief applies in a similar way to less intellectually developed or consciously available processes of predicting behavior on the basis of other behavior. It does not matter how unsophisticated the process might be or how inaccessible its workings might be to the subject herself, as long as it gives her true beliefs most of the time. Justified belief is the main ingredient of knowledge on Goldman's view[2] so despite the differences in the three processes that could be methods of knowing other minds, the methods all would produce knowledge of other minds in virtue of the same property: the reliability of the process through which those beliefs were formed.

Many theories of knowledge will apply neatly to simulations and other forms of mind-reading even though they do not take the process through which a belief was formed to be the crucial bearer of epistemically relevant properties. A more traditional view of justified belief, and thereby knowledge, has an internalist requirement that the subject have access to the reasons for his belief, and be able to give an account of those reasons, and this is a condition that theorizing about other minds can easily achieve. This is because that reasoning is usually imagined to take place via meta-beliefs about the other mind and minds in general, and these will usually be accessible to the subject, at least in principle.[3]

But this traditional view will not count all simulations as giving justified belief. If a simulation is of the high-level variety where one uses imagination to stage the play that results in a conclusion about the other mind, then one will often be able to give some account of why one believes that conclusion, and by that fact have a justified belief. However embodied simulation does not use imagination[4]—the replicas of the other mind's states are coded more primitively in the firings of mirror neurons—and a subject whose conclusion about the other person has been affected by these neurons may not be able to give informative reasons why he has the conclusion he has. If we think that embodied simulation can give a subject beliefs with the same level of justification as beliefs reasoned from a good theory can have, then Goldman's view of justified belief, which is *externalist* since it does not require the subject to be able to call up the reasons for his belief, has an advantage over the traditional view of justified belief and knowledge.

There are other externalist theories of knowledge that can be applied easily to our mentalizing beliefs about other people. My own favorite view takes the properties relevant to the epistemic achievement of knowledge to be two tracking requirements (Roush 2005). The first is the *variation* or *sensitivity* condition, which says that in order for the subject to *know* proposition p her dispositions must be such that the probability she does not believe p given that p is false is high (or, equivalently, the probability that she believes it given that it is false is low):

$$P(-b(p)/-p) > .95 \qquad \textit{variation (sensitivity) condition}$$

The second is the *adherence* condition, which says that the probability she believes p given that p is true is high:

$$P(b(p)/p) > .95 \qquad \textit{adherence}$$

These conditional probabilities are modal properties. In order for the subject to satisfy them in a particular case of belief she must have dispositions that would make her have the same belief or non-belief responses to the truth and falsity of p in a variety of non-actual situations.[5] On this picture knowledge of p is a matter of being disposed to have one's belief states follow p through changes in circumstances and p's truth value.

The application of the tracking conditions to mentalizing about other minds is straightforward. If a subject has dispositions of using a theory and observations about others' behavior and states in such a way as to avoid false ascriptions of mental state m to others and to tend to make true attributions of mental state m, then in any particular case of attributing m she tracks that mental state, and if in addition she is actually right in that case then she also knows.[6] If her more primitive behavior-predicting mechanisms track behavior in a similar way, and so yield beliefs about the other that track the other, then she knows for similar reasons. The criteria are the same for simulations. If via simulation a subject truly believes that another has mental state m and the subject's simulative powers and tendencies are such that were the mental state not there in the other mind she wouldn't believe it was, and were that to be the mental state in a range of other situations she would also believe it was[7], then that belief about the other is knowledge. Tracking can also be defined for states more primitive than belief, by substituting some other state in for b(p) in the equations. A person will track where the actor's hand is if her mirror neurons fire selectively, that is, fire in a certain pattern when the hand is in a given state, and do not fire when it is not in that state. Whether the subject uses a theory of mind, simulation, high-level or low-level, or primitive behavior-predicting heuristics, the results will be knowledge or knowledge-like, if they are, in virtue of the same property: fulfillment of the tracking conditions.

However we come to a belief, in order for it to count as knowledge it must have a robustness by its accuracy being more than an accident. Whereas that is guaranteed in Goldman's view by the requirement that your process get it right most of the time, and in an internalist justified-belief view by the fact that you can give good reasons for your belief, in tracking it is insured by your disposition to get a belief state about p that is appropriately responsive to p's truth value in all probable non-actual circumstances.[8]

On this tracking view it does not matter whether simulation or theorizing is the process one uses because what method or process you use to come to your beliefs does not matter at all per se to whether a belief counts as knowledge. It matters only insofar as the method and your tendency to use that method affect your tendency to come to an appropriate belief state. Similar to Goldman's view the tracking view is also externalist: it does not matter whether under questioning you could come up with any reason at all for your belief. What matters here is whether your belief actually stands in a tracking relation to the truth. Thus according to the tracking view embodied simulation is at no epistemic disadvantage if one who uses it is unable to give an account of how he came to his conclusion about the other person.

The tracking view is consistent with the plain fact that the process or method used matters pragmatically insofar as given the creatures we are some methods will be available and effective while others will not. For example, if there is no fairy godmother in our world then we might need to use perception, or, due to the limits of our automatic inferences in complex cases we might need to use explicit reasoning, which we would also be able to report on questioning. But these are contingent facts about the means to knowledge, not relevant, on the tracking view, to the question what it is to achieve the goal of knowledge.

These reflections make the simple point that on a variety of views of knowledge behavior-reading, simulation, and theoretical mind-reading are capable of delivering knowledge of other minds, and though each theory of knowledge sees a different feature as crucial to whether a belief is knowledge, for a given theory its key feature applies naturally to all three methods of social cognition; the reasons the methods give knowledge when they do are the same within each view. The one wrinkle in this neat story is that internalist views of justified belief or knowledge may have difficulty finding a way to count the deliverances of embodied simulation as knowledge. However, that will depend on the details of the particular epistemological theory and the detailed facts of embodied simulation.

Knowledge versus Understanding

Though the terms "knowledge" and "understanding" are sometimes used synonymously, the word "understanding" as applied to persons has warm, fuzzy, and respectful connotations that are prima facie not well-captured by the idea of accurate identification and prediction of behavior and mental states, even if we add that one must achieve this accuracy robustly. The general concept of understanding as distinct from knowledge has received a great deal of attention in recent epistemology (e.g., Kvanvig 2003, Elgin 2006, 2009), and several features stand out as consensus views among those who theorize about what it is to understand a subject matter. One is that understanding requires appreciation of the relations among the facts in that domain, another that one who understands must be able to use her beliefs in further inferences about or activities concerning the subject matter, still another that understanding is more valuable than and not a subspecies of knowledge. These would not be implausible as requirements on what it is to understand another person's action or desire, but notice that none of the views of knowledge above explicitly imply that one will have any of these properties with regard to a person whose mental state one knows or a proposition that one knows about them. Understanding is a different epistemic property from knowledge, and can be expected to be so in application to persons as much as to subject matters.

The only explicit definition of understanding that I know of in terms of necessary and sufficient conditions is my own view based on relevance matching (Roush forthcoming) which I will use to explore the similarities and differences among the methods of mentalizing about other minds. This is a view not of what it is to understand a subject matter, which occupies most epistemologists in this area, but what it is to understand why p, a proposition, is true, as opposed to merely knowing *that* it is true. The contrast is easily seen by thinking of cases of knowledge got by using mere indicators. I can know what the temperature is outside by using a good thermometer, but that would not give me an understanding of why the temperature is what it is. For that I would need to appeal to things like what season it is, and that a cold front has just come in to my area. A better understanding would come from wider and deeper knowledge of the detailed dynamics of snowstorms, world weather and climate at the current time, and general meteorology.

The relevance matching view of understanding takes its cue from the idea that understanding why p is true requires an appreciation of the relation of p's being true or false to other matters being true or false. Like the tracking view of knowledge it makes the idea rigorous in terms of a particular kind of dispositional covariation of properties of one's belief in p with properties of the world, expressed in conditional probability, and in which the direction of fit has the subject following the world. Unlike in tracking, the relevant matters in the world that one's belief must by definition be well-disposed toward include not just p but many other propositions. All understanding that finite human beings actually achieve will be partial on this view, but this is not counterintuitive.

Any proposition, q, has a degree of relevance to p that is, intuitively, the difference that q's being true or false makes to p's being true or false, and that degree of relevance might of course be none. q's being positively relevant to p can be expressed in probability as follows:

$$P(p/q) > P(p/-q)$$

This says that, other things equal, p is more likely when q is true than when q is false—q's being true rather than false makes a positive difference to whether p is true—and the further the left hand side exceeds the right hand side the more the positive relevance. We can measure how much difference it makes by considering the ratio of the two terms, where positive relevance means $P(p/q)|P(p/-q) > 1$, and I will take how far this value is greater than 1 to measure the degree of positive relevance.[9] If q is not relevant to p the greater-than signs are replaced by equals signs. If q is negatively relevant to p, then not-q will be positively relevant.

A distinct question is whether q's being true or false makes a difference to your *believing* p. If q's truth makes a positive difference to whether you believe p, that is, to b(p), we can express that situation as follows:

$$P(b(p)/q) > P(b(p)/-q)$$

You are more likely to believe p given that q is true than you are given that q is false. q's truth value is something you rely on in opting to believe or not believe p. This can also be written as the ratio of those two terms being greater than 1:

$$P(b(p)/q)|P(b(p)/-q) > 1$$

and measured by how far it exceeds 1. Again if q is irrelevant to your believing p then these greater-than signs will be replaced by equals signs, and when q is negatively relevant not-q will be positively relevant.

If you are to appreciate the relations p has to other matters then the relevance that the truth values of propositions other than p have to your believing p had better match, to some degree, the relevance that they have to p's being true. I will call it a *relevance match for* p *on* q when

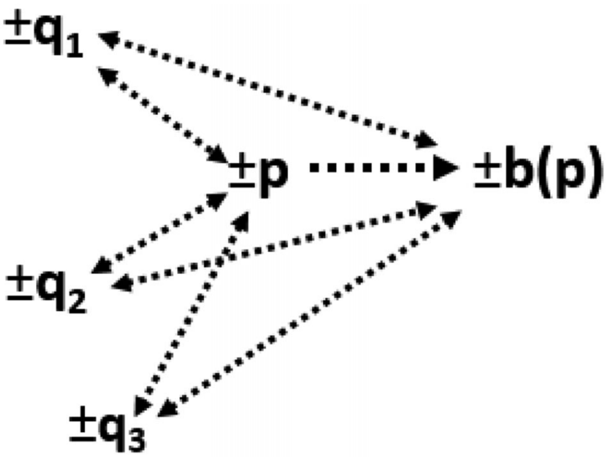
$$P(b(p)/q)\,|P(b(p)/-q) \approx P(p/q)|\,P(p/-q)$$

the difference that q makes to whether you believe p (the term on the left hand side) is approximately the same as the difference that q makes to whether p is the case (the right hand side). We have a *relevance mismatch for* p *on* q whenever

$$P(b(p)/q)|P(b(p)/-q) \not\approx P(p/q)|P(p/-q)$$

that is, when q's truth value makes significantly more of a difference or less of a difference to whether you believe p—the left-hand side—than it does to whether p is true—the right-hand side. Understanding why p is true then requires that you relevance match for p on an appropriate domain of q's. For the case of temperature, reading the thermometer would give you little understanding of why the temperature is low if you would believe it is low regardless of whether a cold front is or is not coming in and of whether you do or do not live near the equator, since the temperature depends on those things.

Pictorially, the relevance matching conditions require the following kind of relation of your dispositions to believe p and p's relations to other matters:

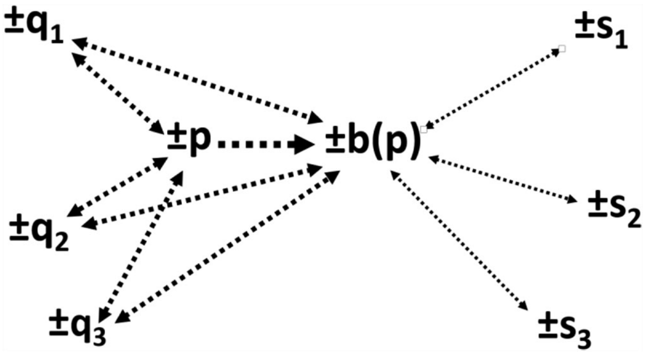


The arrow from ±p to ±b(p) indicates that your belief in p tracks p, where the ± indicates the two possibilities for the truth values of the proposition; tracking has you following the world and not the other way around, so the arrow is single-sided. The arrows between the q_is and your belief in p indicate the relevance matching by being isomorphic to the arrows between the q_is and p.

While this picture is good as far as it goes, it does not capture a notion of understanding that goes beyond tracking, for perfect tracking—the relation in the diagram between p and b(p)—logically implies perfect relevance matching, and vice versa. This is because a perfect indicator does your relevance matching for you. A perfect thermometer perfectly follows the temperature, and so follows the existence or not of a cold front in precisely the way that that temperature does. In perfectly following the thermometer your belief about the temperature follows the cold front in the only way you need to do for relevance matching.

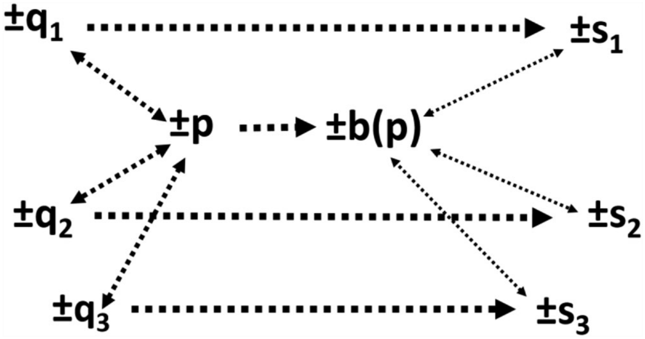
Imperfect relevance matching and imperfect tracking diverge in ways that are important pragmatically to how best to go about pursuing either of them (Roush forthcoming). However they do not differ in the right way to make imperfect relevance matching into imperfect understanding, since any degree of mere matching of your believing p to the factors in p's web of relevance is still compatible with the matching being secured by something other than yourself. You may do imperfect but good relevance matching to the factors relevant to a storm via an imperfect but good thermometer. Or you might consult what a good computer simulation programmed by someone else concludes about the likely course of a storm, and to be good it must make use of many factors relevant to the storm. Your belief about the course of the storm will relevance match to those factors because the simulation does, but if you have no clue or sense of how the simulation gets its conclusions then this gives you no understanding of why the storm will follow that path.

This inadequacy of mere relevance matching corresponds to the observation of several authors, noted above, that understanding is something that its possessor can use in further inferences or activities. Relevance matching is a direct way of expressing sensitivity to the factors in p's web of relevance and is part of understanding on my view, but the subject can achieve it without being able to take the understanding home with her and leave the computer at work, without herself having a capacity to know how to go on or exploit the relevance matching in her further endeavors. The additional condition besides relevance matching that is needed for a subject to understand why p is true, as I have argued elsewhere (Roush forthcoming), is that the subject herself possess states whose dispositional relations to her belief in p match the dispositional relations that the q_i's have to p, a situation that can be depicted as follows:



The subject who possesses the network of four states in the right half of the diagram not only has a belief-state with regard to p that is dispositionally related to the q_i's in the same ways that p is related to those q_i's but also has states s_i that her believing p or not is dispositionally related to in the same ways as p is related to the q_i's. For example, if p's truth is positively relevant to q_3's falsehood to some degree, then the subject's belief in p is positively relevant to state s_3 being in the off mode to a similar degree. And so on for all of the possible permutations. For ease of exposition I have represented the possibilities as binaries—true or false, plus or minus, on or off—but naturally they could come in degrees.

The subject in the next figure relevance matches to p on q_1, q_2, and q_3 (omitted for simplicity) as in the previous diagram, but this figure displays that having s_i's with the property just explained gives the subject a dynamical model of the web of relevance around p, because the subject's s_i states *track* the q_i's. This is because it follows logically from her states, $\pm s_i$, relating to her belief state $\pm b(p)$ in the same way as the $\pm q_i$ relate to $\pm p$, and her belief b(p) tracking p, that her states s_i have a minimum level of tracking of the q_i respectively.[10]



In cases fitting these latest two diagrams, the subject has a set of states that is isomorphic to p and its web of relevance, which I call a *dynamical* *mirror model* because the replica is not a static image but fully dispositional. The s_i's do not just happen to coincide at a given time with b(p) in a way that mirrors the q_i's relations with p. The covariation of their values, indicated by + or −, with the subject's believing p or not co-varies with the covariation of the q_i's' values with p's truth or falsity in all relevant possible worlds.

Understanding why p is true is not itself a propositional attitude toward p but a co-variation of co-variations of states of the world and states of the understander. To understand why p is true is to own a dynamical working model of p's web of relevance, not in one's pocket but somewhere in one's person.[11] Owning that model implies that in any particular actual state of the world you very likely *actually* have the appropriate mirror states, whether that means "off" or "on" corresponding to true and false for p and each of the q_i. As above with tracking, what it might take to bring it about that you have this set of interacting dispositions, or to activate them in a particular way, is not relevant to whether the set of dispositions counts as understanding, a point that will be important in application to the case of mirror neurons below.

I do not require that the states s_i in one's model are themselves beliefs or even mental states more generally, because, on theoretical and intuitive grounds, I do not think understanding requires these. Theoretically, it is evident from the diagrams, I think, that what does the work of giving a subject appreciation of the relation of p to other matters is not the qualitative properties of the s_i, but the *relations* of the s_i's to other states of the subject, especially b(p), and the relation of these to the q_i's and p. It is appreciation of the relations that other matters have to whether p is true, the relevance matching, not the qualitative properties of the states one does relevance matching with, that makes it count as understanding. It is clear from the definitions that understanding is strictly stronger than tracking, and hence more than knowledge. And the fact that understanding resides in the relation of one's dispositional states to a network of other matters means that even when one's understanding is carried by states more primitive than belief, that's counting as understanding is not due to understanding's having been watered down to predictive success about p.

Intuitively, we can see why understanding does often involve but does not require beliefs about the q by thinking about examples. In a well-worn example from discussion of scientific explanation, we improve our understanding of why p, Mr. J has paresis, a distinctive kind of paralysis, when we discover that only people with tertiary, untreated syphilis have this condition. This is borne out by my picture because the discovery means that our belief or lack of belief that someone has paresis now covaries appropriately not only with beliefs about his paralysis symptoms but with belief or lack of belief in another factor, the presence of latent untreated syphilis, so this counts as an improvement in our understanding of Mr. J's having the paralysis. This is a case where the appreciation of relevance relations is carried by beliefs: that a patient has untreated syphilis, and that this is a necessary condition for paresis, are things one acquires beliefs about. Explicit reasoning about Mr. J would display one's understanding. Reasoning from a belief that latent untreated syphilis is a necessary condition for paresis to the belief that the patient has latent untreated syphilis would exhibit that one has beliefs whose dispositions to be held or not correspond to the logical and probabilistic relations that the beliefs' propositional contents have in the world. In other words, the s_i's in this case are beliefs in the q_i's; s_i is $b(q_i)$.

However, the appreciation of relevance relations around p may be carried by states that a person does not have reflective access to and that may not have the compositional complexity of belief. We have this in cases of expert intuition, for example when, as the story goes, Linus Pauling, on learning of the sickled appearance of the blood cells in sickle-cell anemia, immediately declared that this must be due to a genetic abnormality in the hemoglobin molecule. It took years of dogged work (by Harvey Itano) to discover what the chemical abnormality is that causes the sickling, but it was considered worth the investigation because everyone knew that Pauling's vast experience of molecular chemistry gave him understanding—appreciation of the extent and types of relevance of chemical properties to phenomena at a larger scale—that supported the plausibility of his specific conjectures. He appreciated relevance relations beyond what he would probably have been able to say explicitly at the time.

Understanding without the ability to report beliefs or give explicit explanations is common in science (Lipton 2009). Visual imagination and a good orrery can give one understanding of the retrograde motion of Mercury that one might not be able to give an equivalent verbal account of. One can acquire sophisticated understanding of a machine by becoming expert at using it, and this manipulational ability stores causal information that one may not be able to articulate in sentences. The inarticulateness, which on my model is due to the states s_i being something more primitive than belief, does not prevent understanding from being usable. One who understands Mercury's motion via the orrery would be able in the same way to see what change in its position or motion would iron out the retrograde appearance. One who understands the machine via manipulation will know how to use it on a new case. Pauling's understanding gave him hunches about which specific properties to check. Understanding is often used to make inferences from some beliefs to others, but it need not be carried by beliefs in order to be usable.

Social Cognition and Understanding

The two main models of the way that we cognize other minds, simulation and theoretical mindreading, can both be seen as capable of delivering understanding on this view. Both simulation and reasoning are processes, and, speaking abstractly, they are each an activation of a complex of dispositions that lies in wait in the understander. In the case of reasoning these are dispositions to change one belief in response to changes in another belief. For a very basic example, a good reasoner will be disposed to believe p whenever he believes p-and-q, and will be disposed to withdraw belief in p-and-q when he comes to believe not-q.[12] For deductive relations among propositions the dispositional relations among his beliefs can mirror those relations in the world without him having done empirical investigation. When his beliefs and the dispositional relations among them are graded, degree of belief in p being disposed merely to increase degree of belief in q for example, those relations among his degrees of belief will mirror the relations among the propositions, such as p and q, when he has it right about how the matters of p and q relate in the world.

Suppose one knows that actor A is wealthy, with much more in resources than he can use himself. In conversation with A one forms the belief that A is sympathetic to the plight of B, and one encourages A to help B financially. One knows that B is proud, so there is some likelihood he will reject the help, but there is a decent chance he will accept it although that would also bring him ambivalence and resentment. But you feel sure A will not tell B that you encouraged him because A will want to be seen as coming to the idea himself. Thus you think that your encouragement of A will increase the chances that he offers help to B, and not be revealed to B, and that in this way you can raise the chances of B getting help without B's resenting you or being embarrassed for it in front of you.

In this case reasoning with a theory of mind gives a plausible model of your cognition of A's and B's actions and mental states. You have beliefs about how pride affects acceptance of help and inclination to reveal the causes of one's actions, as well as the mental states of ambivalence and resentment and embarrassment, and about the characteristics of A and B in particular in these dimensions. These beliefs are premises in your reasoning, and they are also s_i in the model of understanding. You can understand to some extent why A offers help, if he does, or why B accepts it, if he does, because you have beliefs about these q_i and their relevance to, their potential to affect, A and B's behavior, and your beliefs about the q_i are disposed to change relative to each other and to your belief in p, say the proposition that B accepts the offer with ambivalence, in the same way as the q_is are disposed to change relative to each other and to p. Moreover, you would be able to call up at least a sketchy version of these beliefs and their relationships on questioning. Thus, the dynamical mirroring model of understanding can easily count theoretical mindreading as showing understanding.[13]

A simulation also activates a dynamical mirror model, a web of dispositional relationships among the understander's mental states, but in this case instead of reasoning, inferring from belief to belief about pride and offers of help, one redeploys one's own affective and motivational machinery. One settles in to an offline version of the feeling of pride, for example, pretends that one needs help and that someone offers it, and then notes the offline emotional reactions one has. If one's own emotional equipment is relevantly similar to that of the actor, and one has donned the appropriate offline beliefs and emotions to mirror his, then the outcome offline emotional reaction will to a good extent match the reaction of the actor. Some of the s_i are (offline) beliefs in this case, $b(q_i)$, such as that one has been offered help, but others are emotional states, m_i. We certainly possess concepts of pride and embarrassment, but that alone does not imply that we always use them as concepts, in beliefs, to come to our conclusions, even when they are relevant.

Action Understanding and Mirror Neurons

Some of the motor neurons that fire in a monkey when he grasps for an object with his hand also fire when he sees someone else's hand grasping for an object in a similar way, though his mirroring neurons are not inducing the same action on his part. Similar phenomena of reuse or redeployment have been found in humans, and over a wider range of mental states including not only action intentions but sensations and emotions (Gallese et al. 2004, Keysers and Gazzola 2009, Rizzolatti and Craighero 2004, Casile 2013). One dominant interpretation of the phenomenon is as a distinctively embodied type of simulation, where what makes the firing of motor mirror neurons embodied is that the information is formatted in bodily rather than propositional form (Goldman and Vignemont 2009), and the possible patterns are constrained by dynamical possibilities of physical movements of the body.

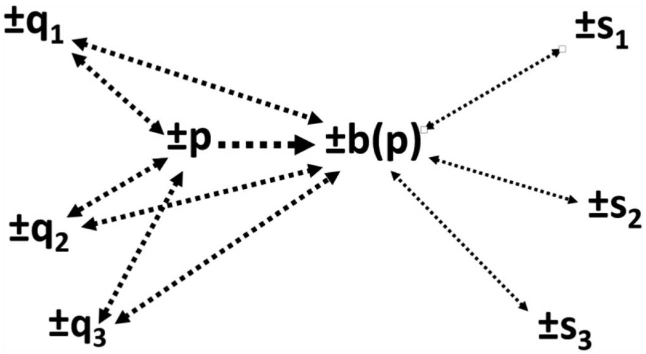
The mirror neurons for, e.g., grasping objects have remarkable selectivity to the goals of the witnessed action. In many cases the monkey's neurons that fire when he witnesses a genuine action do not fire for mimicry, where the observed hand grasps just as if to pick up an apple but the monkey knows no apple is there. When the grasping does have a goal, some mirror neurons fire differentially for different goals. These can also fire selectively at different stages of the action, at the beginning or later, in a way that shows responsiveness to the level of abstraction of the goal (Casile 2013). Some mirror neurons selectively respond to the goal independently of the proximal means used to achieve it, whether by hand, foot or mouth, or by a variety of different tools (Sinigaglia 2013).

As more evidence accumulates about how this process works, much controversy surrounds what precise role this kind of simulative firing plays in social cognition. In the understanding of motor action in particular, the chief question has been what role the mirroring plays in understanding the goal of an observed action. Two points of disagreement have been prominent in the discussion. One is whether the mirror neuron firings are sufficient for identification of the goal of an action. Some think the goal is identified by the mirror neurons via a *direct matching* (Rizzolatti et al. 2001, Gallese 2006, Gallese and Sinigaglia 2011), while others think the evidence points to a higher-order process that first identifies the goal and tells the mirror neurons to model the observed action as having that goal rather than another by firing selectively in the pattern corresponding to that goal (Csibra 2013, Jacob 2013). A second prominent point of disagreement is around whether mirror neurons firing in the pattern that corresponds to a particular goal counts as *ascription* of that goal to the actor.

Both of these issues are taken to be crucial to whether mirror neurons can deliver understanding all by themselves. I will take issue with this and argue that which level of processing identifies the action as having one goal rather than another, and whether matching firing by itself counts as ascription of a goal, do not matter to whether the selective firing of mirror neurons counts as understanding why a hand moves as it does. I will argue for these things by showing that according to the independently sensible view given above of what it is to understand why p is the case, the mere selective firing of mirror neurons counts all by itself as understanding why a hand moves as it does. There is also a kind of understanding of why the hand moves as it does that is not captured by identifying which goal the actor has or ascribing that goal rather than another to the actor. This is appreciation of the goal *that it is a goal* rather than just a likely future event that makes some intermediate events more likely than others, and this kind of understanding the matching mirroring delivers directly. Overall, with my definition we can capture what Rizzolatti, Gallese, Sinigaglia, and others are seeing in the mirror neuron phenomenon, without running afoul of some of the objections that have been made to their particular points.

Csibra (2013) and Jacob (2013) both argue that direct matching is not sufficient to identify which goal an actor has in moving his hand in the observed way. Their point can be illustrated through an interpretation of the experiments showing mirror neurons' selective behavior in response to mimicry, not firing when the monkey knows that there is no apple the observed hand can get to. The only observed feature that could tell the monkey that the motion is mimicry is the contextual cue of presence or absence of an apple, but the mirror neurons cannot themselves be processing that cue since all that they can do is fire, or not fire, in a pattern that mirrors the firing of the motor neurons making the hand move. Some other process, this view says, must interpret the scene using the indicator of the presence or absence of the apple, to tell the mirror neurons whether to fire in the grasping way or not. As Csibra says, "... such interpretation [of goals] precedes rather than follows from action mirroring" and "goal understanding is not the output but the input of the mirroring process" (Csibra 2013, 450, 446, respectively).

However this may be, it does not matter to whether the mirroring by itself counts as understanding why the hand is moving as it does. This is because the mirroring by itself counts as a dynamical mirror model of the action in the sense above, regardless of what further things cause the neurons' activation or the selectivity of their response. Take the p in the diagram below to be the proposition that the hand moves thus and so. Some of the q_i relevant to whether the hand moves thus and so will be propositions concerning the actor's relevant motor neurons firing. The subject who observes the hand moving thus and so takes it to move so, that is, b(p), and that belief is related to his own corresponding motor neuron firings, s_i, in the same way as the actor's hand moving, p, is related to the firings of his motor neurons, q_i. The subject thus possesses understanding of why p is the case, and this is in virtue of actually mirroring and possessing the selective tendencies required to fire the relevant neurons when the action is observed, and not when it is not. Rizzolatti et al. put it simply in introducing their direct matching hypothesis: "... we understand an action because the motor representation of that action is activated in our brain".



The understanding delivered by the mirror neurons alone is incomplete of course. There will usually be other relevant matters that one could come to have states covarying with, e.g., further goals, such as eating the apple after picking it up, that this particular mirroring set of neurons is not matching to, but to which the subject could relevance match via some sort of state if further clues about the action were forthcoming. But all understanding is incomplete and can be filled out further by acquiring states that are disposed to covary with more states of the world probabilistically relevant to whether p occurs.

In the case of the moving hand, if there is a process separate from mirror neurons that does the job of identifying which goal the actor has, then if that node also corresponds to some state in the actor that is related to his moving his hand, such as the intention to do this rather than that, then the observer's having that identifying process gives him greater understanding of why the hand moves that way than he had without it. It is just that having that further node is not required in order to have understanding at all of why the hand moves as it does. This is so even though on the assumption that there is a separate process identifying the goal, that process is causally responsible for activating the mirror neurons' dispositions in one way rather than another. As drawn out above, what it might take to bring it about that you have a given set of interacting dispositions, or to activate them in a particular way, is not relevant to whether that set of dispositions and activations counts as understanding.

Nor, on my view, is ascription of the goal required for understanding why the hand moves as it does. With Jacob (2013) and others, I take ascription in the full-bodied sense that I think is natural, which is that "to ascribe an intention to another is to judge or believe that they intend so and so" (Jacob 2013, 1139). So, like these authors, I would not classify the selective firing of mirror neurons as ascription of a goal or intention. But both those in the discussion who think mirror neurons' firing does count as ascription (e.g. Gallese) and those who think it does not (e.g., Jacob), are assuming that ascription of a goal or intention is needed for understanding of an action, and this is where I disagree.

This is for two reasons. First, to suppose that goal ascription—understood as some further thing beyond mirroring itself—is necessary for understanding why a hand moves as it does is to make the same mistake as we saw above with assuming that the cause of the activation of the mirroring must be included in order to count the mirroring as understanding. Second, to suppose that ascription in the full sense involving belief is required for understanding is equivalent in my terms to supposing that in order to understand why p is the case the states s_i that the observer has that correspond to and covary with the q_i that are relevant to p must be *beliefs* about the q_i—e.g. a belief that the actor intends to have the apple—whereas I have argued above that those states need not even be mental.

We see this with someone whose understanding of a machine, that gives him an ability to use it, is carried by muscle memory from his experience of using it. The understanding is carried by the covariation of covariations, whatever the qualitative nature of the states that carry it. In the case of the hand, there is evidence that the neurons' mirroring allows the observer to predict or anticipate the future steps of the movement, and the outcome, e.g., of acquiring the apple. If those anticipations are reliable then the observers' states are covarying with future states of the hand, and even if the states that subserve those anticipations are not beliefs, but information in bodily format, the covariation relationships they carry count as part of the understanding. There is no need to count the mirroring as ascription of a goal in order for it to qualify as understanding.

There is a further kind of understanding of the action that mirror neurons can achieve all by themselves and that it may be difficult for discursive beliefs to capture as richly or efficiently. This is understanding not that the goal or intention is this rather than that, but that it is a goal or intention at all, as opposed to merely a likely future state of affairs which will make some intermediate states of affairs more likely than others. The difference can be illustrated with a clock. We can identify the future state towards which a clock is set to run—e.g., it is set to display the same time as a particular satellite does 10 minutes from now. But that is not an intention of the clock. We do not understand it as an intention and would not simulate the movement of the clock's second hand with our mirror neurons (unless, perhaps, we were pretending the clock was a person, or that we were the clock).

Mirror neurons allow us to appreciate that the movement of another person's hand is an action, not just an unfolding sequence of states of the world. They do this in virtue of the fact that they are redeployments of the mechanism that we use when we act ourselves. It would be difficult to give a discursive definition of what it is to act or to intend, as opposed to events just happening, but with mirror neurons we do not have to use such a definition. Whatever action and intention are in themselves we can recognize them by reference to our own acts; acting is whatever it is I am doing when my own motor neurons are firing and causing movement of my hand. Sinigaglia (2013, 60) makes a similar point that ". . . [t]he mirror mechanism enables the onlooker to understand the goal of the other's action from the inside as an outcome to which her own action can be directed and not just from the outside as something that can happen, as a mere event among others". It allows one to appreciate of a goal or intention that it is a goal or intention.

Efficiency, Vulnerability, and Morality in Social Cognition

> The continual emotion that is felt in the theatre excites us, enervates us, enfeebles us, and makes us less able to resist our passions. And the sterile interest taken in virtue serves only to satisfy our vanity without obliging us to practice it.
>
> — Jean-Jacques Rousseau, *Politics and the Arts*: Letter to M. D'Alembert on the Theatre

In the justified excitement about mirror neurons and simulation of other minds in general, there is too little discussion of the disadvantages of these ways of knowing others' minds.[14] Obviously one dimension of trade-off is between efficiency and accuracy. It seems that simulation in general and use of mirror neurons in particular can be more efficient than conceptualizing and theorizing in simple cases, insofar as it bypasses at least some reasoning processes, thus lightening the computational load on the brain and making us able to judge and respond faster. But of course some matters are complicated and for those a slower and more laborious reasoning with propositions will have enough better accuracy to compensate for the computational cost and loss of speed. So far this is merely the well known difference between thinking fast and slow, but understanding by redeployment of one's own system of action-intention, sensation, or emotion introduces a new dimension of cost to the observer, the cost of feeling or virtually doing what the other feels or does.

Redeploying one's own pain neurons when observing someone else in pain has the immediate cost of feeling bad, but the costs can also ramify. One's natural impulse in response to one's own pain is to take steps to alleviate it, but if it is caused by witnessing someone else's pain one could achieve this not only by alleviating the other's pain, but also by moving far enough away so as not to observe the other anymore; one could help or one could flee, which may cause feelings of guilt later. Processing the other's pain as a fact to have beliefs about rather than re-experiencing it oneself would make it possible to respond on the basis of ethical principles and avoid sympathetic pain. If a nurse had an impulse to flee when re-experiencing someone else's pain that would be a reason not to simulate on the job (or not to be a nurse), but equally if simulation prompted impulses to help, delivery of the help would nevertheless be made less efficient by her wincing at the patient's pain.

Feeling another's pain can also make one vulnerable to exploitation. Suppose that you have a tendency to mirror another's pain when you observe it and a tendency to help rather than flee in response to it, and that your decision whether to help is strongly influenced by the intensity of the pain or need for help. Then someone who is good at faking the expression of pain or neediness can exploit your tendencies and be more likely to gain what they want. If instead of simulating you had evaluated the situation entirely in the "pale, detached" medium of belief—representational attitudes about propositions—you would observe the visible cues that the person is feeling pain, but you would consider various other relevant matters as well, such as whether there are other possible motives for the person to display pain signs, and whether you have the spare resources to help. Even if the good faker succeeded in getting you to believe they were in pain, these other considerations would be weighed alongside, and that attributed pain would also not give you the same degree of motivation to help that it would if it were experienced as your own pain.

Redeployment of one's own apparatus for feeling and doing when observing others feeling and doing may have moral costs as well. Rousseau famously condemned theatre, arguing that an audience's emotional identification with the characters in theatre was morally disastrous. This was, roughly, because empathy has a natural role in our innate moral sense motivating us to moral action, and the experience of it and its resolution on stage gives an empathizing viewer the false sense that moral duty has been discharged (Banerjee 1977). The intensity of the emotions identified with in some drama might also dull one's sensibilities by wearing them out.

Jane Heal puts the point nicely when she explains what a simulation is by imagining using one's heart "offline" to predict what will happen to someone else's heartrate under various physical exertions: "All this sounds pretty risky. Perhaps it would be better to carry around a spare heart to do the experiments on" (Heal 1994, 134).

It is easier for an observer to distinguish virtual acting from his taking real action than it is to distinguish virtual feeling from real feeling. Yet there is an additional worry about actions, that identification with or re-experiencing of another's intentional actions can make similar action on one's own part more probable, and those actions may not always be good (Cf. Khalil 2011). This could be expected if the experience of dancers is any guide. "Marking" is imaginatively or with only token physical movements putting oneself through the motions of a routine—not just imagining someone else do it. It is a ubiquitous practice because without exhausting the muscles it succeeds in training them to be more likely to carry out the routine correctly. When mirror motor neurons are firing in observation they are either inhibited from or not turned on for leading to action, by the firing or not of other neurons, but the regulatory system must be more sophisticated than that if virtual dance gets through to affect one's future actual dance. Learning more about how the switching system works will be important for evaluating the cost of simulative ways of knowing.

Conclusion

I have argued on the basis of an independently plausible account of what it is to understand why p is the case that theoretical mindreading and simulation achieve understanding in virtue of the same feature, which is the ownership of a set of states disposed to vary with the subject's believing p or not in the same way that states of the world relevant to p vary with p's truth value. I have used this concept of understanding to argue that the fact that mirror neurons fire selectively in response to observed actions in the same way that they fire when the subject so acts herself is sufficient for the subject to have an understanding of why the actor's hand moves as it does. This understanding is achieved whether there is or is not a further process that causes the network of dispositions to be activated selectively in a particular case, and whether or not the subject has a further belief state that is ascription of a goal, though having those additional nodes in his network would increase or improve the subject's understanding in a clear way. I have further argued that the matching that mirror neurons do delivers understanding of an intention that it is an intention in a simple and direct way. Finally, I have suggested that the potential costs of social cognition by the method of simulation, and the relation of those variables to how the brain works, are worthy of further study.

Notes

1. Note that to avoid clumsy terminology in what follows I am using the term "mindreading" generally, for understanding another's mind regardless of method, and the qualifier "theoretical" when referring to that particular method.
2. Knowledge is defined by adding to Goldman's view of justified belief a requirement that the belief actually be true, and a clause that rules out troublesome Gettier cases.
3. An exception to this is described in footnote 13, but there too the reasoning would be consciously accessible to the subject.
4. This is so even if higher-level interpretation is required in order for those mirroring simulations to be delivered selectively depending on goals. (See Csibra 2013, Jacob 2013.)
5. The conditional probabilities listed for the tracking conditions are shorthand for a universal quantification over a set of probability functions that corresponds to a set of possible situations; that is what makes the tracking conditions modal and dispositional. How to define which situations among all possible situations are the ones the subject is responsible for is a long-standing question which I answer via further probabilistic conditions (Roush 2005, Ch. 3).
6. The tracking conditions are not strictly closed under known implication, leading to the consequence that the level of tracking can fall off with each inference, and eventually have the level of tracking in one's conclusion belief fall below the threshold. I address this in Roush (2012), where I show that under natural conditions the tracking level falls off gradually with acceptable lower bounds, a property I call "closure enough".
7. For ease of exposition I formulate the conditions counterfactually rather than probabilistically in this sentence though the distinction is quite important (Roush 2005, Ch. 3).
8. Externalist theories of knowledge are weaker than internalist in that they do not require conscious access to reasons, but stronger along another dimension. Typical internalist theories do not require for knowledge any reliable relation of the subject's belief to how the world is. The only relation the belief must have to the world is being true.
9. There are other ways of measuring degree of probabilistic relevance. This measure corresponds to the likelihood ratio, which I prefer on other grounds (Roush 2005, Ch. 5), and which makes for a clear link between this view of understanding and the tracking view of knowledge.
10. This tracking is achieved to a degree bounded from below by the level of matching the s_i-b(p) relations have to the relations between the q_i's and p, and the level of tracking that b(p) does of p, a property I call "transitivity enough". See Roush (2005, Chapter 5).
11. The force of my phrase "in one's person" is to include the body and not just the mind, and to require an intimate sense of ownership of the states that co-vary with the world, but it is not intended to exclude a priori the possibility that the mind extends beyond the skull to include cognitive aids in the environment (Clark and Chalmers 1998). I say that a person does not achieve understanding of why the temperature will be high merely by consulting a computer simulation's output

prediction that the temperature will be high, if he has no knowledge or sense of the dimensions of the model the computer is using. The computer itself may understand, since it does possess the dynamical model of the weather. And the computer's computation can be said to be part of the person's cognitive process for *identifying* the temperature, and so part of his extended mind. However to be hooked up to the variation in the computer's states that correspond to variable dimensions in the world—i.e. to be hooked up to the computer's understanding—in such a way as to make use of it would ipso facto give one a dynamical mirror model inside oneself, meaning that there would be no reason to say one's understanding was located in the mind's extension rather than inside one's skull. Understanding may be a bit like taking a bath—something you have to do yourself.

12. I myself think of beliefs as dispositions, but any view of belief is consistent with the model I am describing because the question what beliefs are can be separated from questions about the dispositions to have a belief or to have one belief when one has another.
13. The fact that theoretical mindreading qualifies as a dynamical mirror model may raise the worry that on my view implicit theory use collapses to simulation, since the dynamical mirror model idea has a simulationist feel to it. This is ironic since traditionally the worry has been how to define simulation in such a way as to avoid its collapse to implicit theory (e.g., Heal 1994). However, theorizing remains distinct on my view because it uses beliefs *about* the actor in its determination of a conclusion about the actor, rather than the subject's re-deploying her own psychological mechanisms. The only case when this distinction would potentially not apply is when the subject is trying to determine whether the actor believes p given that the actor believes a set of premises q, r, and s. The most efficient way to figure this out would not be to manipulate beliefs about the beliefs of the actor, but to suppose that the actor is rational like you and to use your own reason to infer from q, r, and s to p or not-p. In this case you are re-deploying the equipment that you use to get beliefs like p and not-p, and thus count as simulating, but I see nothing counter-intuitive about calling it that, and it does not change the distinction present in other cases.
14. An exception is Khalil, who is concerned about how mirroring can lead to contagion, whether the action mirrored is good or bad.

References

Banerjee, Amal (1977), "Rousseau's Concept of Theatre", *British Journal of Aesthetics* 17 (2): 171–177.

Casile, Antonio (2013), "Mirror Neurons (and Beyond) in the Macaque Brain: An Overview of 20 years of Research", *Neuroscience Letters* 540: 3–14.

Clark, Andy and David J. Chalmers (1998), "The Extended Mind", *Analysis* 58 (1): 7–19.

Csibra, Gergely (2007), "Action Mirroring and Action Understanding", in P. Haggard, Y. Rosetti, and M. Kawato (eds.), *Sensorimotor Foundations of Higher Cognition: Attention and Performance XXII*. Oxford: Oxford University Press, 435–480.

Elgin, Catherine Z. (2006), "From Knowledge to Understanding", in S. Hetherington (ed.), *Epistemology Futures*. Oxford: Clarendon, 199–215.

——— (2009), "Is Understanding Factive?", in D. Pritchard, A. Miller, A. Hadock (eds.), *Epistemic Value*. Oxford: Oxford University Press, 322–330.

Gallese, Vittorio (2013), "Mirror Neurons, Embodied Simulation, and a Second-Person Approach to Mind-Reading", *Cortex* 49: 2954–2956.

——— (2011), "Response to de Bruin and Gallagher: Embodied Simulation as Reuse is a Productive Explanation of a Basic Form of Mind-Reading", *Trends in Cognitive Science* 16: 99–100.

Gallese, Vittorio and Alvin Goldman (1998), "Mirror Neurons and the Simulation Theory of Mind-Reading", *Trends in Cognitive Sciences* 2: 493–501.

Gallese, Vittorio, Christian Keysers, and Giacomo Rizzolatti (2004), "A Unified View of the Basis of Social Cognition", *Trends in Cognitive Science* 8 (9): 396–403.

Gallese, Vittorio and Corrado Sinigaglia (2011), "What is So Special about Embodied Simulation?", *Trends in Cognitive Science* 15 (11): 512–519.

Goldman, Alvin I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*. Oxford: Oxford University Press.

——— (2012), "A Moderate Approach to Embodied Cognitive Science", *Review of Philosophy and Psychology* 3 (1): 71–88.

Goldman, A. and Frederique de Vignemont (2009), "Is Social Cognition Embodied?", *Trends in Cognitive Science* 13 (4): 154–159.

Heal, Jane (1994), "Simulation vs. Theory-Theory: What is at Issue?", *Proceedings of the British Academy* 83: 129–144.

Keysers, Christian and Valeria Gazzola (2009), "Unifying Social Cognition", in J. A. Pineda (ed.), *Mirror Neuron Systems*. New York: Humana Press, 1–35.

Khalil, Elias L. (2011), "The Mirror Neuron Paradox: How Far is Understanding from Mimickry?", *Journal of Economic Behavior & Organization* 77: 86–96.

Kvanvig, Jonathan (2003), *The Value of Knowledge and the Pursuit of Understanding*. Cambridge: Cambridge University Press.

Jacob, Pierre (2013), "How from Action-Mirroring to Intention-Ascription?", *Consciousness and Cognition* 22: 1132–1141.

Lipton, Peter (2009), "Understanding without Explanation," in H. de Regt, S. Leonelli, and K. Eigner (eds.), *Scientific Understanding: Philosophical Perspectives*. Pittsburgh: University of Pittsburgh Press.

Rizzolatti, Giacomo and Laila Craighero (2004), "The Mirror-Neuron System", *Annual Review of Neuroscience* 27 (1): 169–192.

Rizzolatti, Giacomo, Leonardo Fogassi, and Vittorio Gallese (2001), "Neurophysiological Mechanisms Underlying the Understanding and Imitation of Action", *Nature Reviews Neuroscience* 2: 661–670.

Roush, Sherrilyn (2005), *Tracking Truth: Knowledge, Evidence, and Science*. Oxford: Oxford University Press.

——— (2012), "Sensitivity and Closure", in *The Sensitivity Principle in Epistemology*. Oxford: Oxford University Press.

——— (forthcoming), "The Difference between Knowledge and Understanding", in P. Klein, C. de Almeida, and R. Borges (eds.), *Explaining Knowledge: New Essays on the Gettier Problem*. Oxford.

Sinigaglia, Corrado (2013), "What Type of Action Understanding is Subserved by Mirror Neurons?", *Neuroscience Letters* 540: 59–61.