CHAPTER 4

DEFLATIONARY NATURALISM ABOUT BELIEF: THE CASE OF SIMULATION THEORY

Naturalists in the philosophy of mind have typically focused their efforts on questions of the *ontology* of belief, by attempting to show how mental states might either reduce to or supervene on the physical. In the previous chapter, I have shown how a wide range of proposals for naturalizing belief in this manner are not successful.

But Huw Price (2004) has noted that there are two ways to approach naturalistic quandaries about concepts like "belief": as an *object* naturalist, or as a *subject* naturalist. The first takes the primary philosophic question to be whether a given controversial type of entity X exists in the natural world, typically answered by deciding if the term "X" refers; the second takes the primary question to be about ourselves as natural creatures, thus primarily about the status of our use of the *term* "X."

According to Price, the object naturalist approach has independent problems apart from the ones I noted in the previous chapter. In particular, if we find it is impossible to naturalize "belief," and decide it does not refer, we are threatened by the incoherence, discussed by (Boghossian 1990), of saying that "content" has no content. And if we take seriously the arguments of Stich (1996), then insofar as there is no way to naturalize theories of reference, there is no determinate way to say *either* that "belief" refers *or* that it does not.

But Price tells us that the second approach to naturalization, subject naturalism, does not face the problems of the first. It avoids each of the problems faced by object naturalism, insofar as it avoids questions of reference entirely. The main purpose of subject naturalistic philosophy, instead, is to "account for the use of various terms . . . in the lives of natural creatures in a natural environment" (81). So a subject naturalism about "belief" is interested in pragmatic usage of the term "belief," not its reference. So while the first approach to naturalism is an inflationary approach to mentalistic ontology, the second is deflationary.

How, then, would subject naturalism propose to account for our use of the term "belief"? Appropriately enough, one of the first overtly naturalistic philosophers to propose such a project was none other than Quine himself. Quine (1960, 219) sketches a model of belief-attribution that he would later call "empathy":

> [I]n indirect quotation we project ourselves into what, from his remarks and other indications, we imagine the speaker's state of mind to have been, and then we say what, in our language, is natural and relevant for us in the state thus feigned…. Casting our real selves thus in unreal roles, we do not generally know how much reality to hold constant….We project ourselves even into what from his behavior we imagine a mouse's state of mind to have been, and dramatize it as a belief, wish, or striving, verbalized as seems relevant and natural to us in the state thus feigned.

In his later work Quine (1992) came to rely on the concept of empathy not only to account for "irreducibly mental" ways of grouping "neural realities" (an endorsement of Davidson's anomalous monism) (71–2), but even as an account of language learning (1992, 42–3; 1995, 89–90). Drawing on Quine's insight, and the work of Davidson (1968), Stephen Stich (1983) endorsed a similar view in his earlier work.

However, critics of Quine, such as Ebbs (1994), have worried that "empathy" is a notoriously subjective-sounding concept for a naturalist like Quine to invoke. Perhaps in light of Price's concept of subject naturalism, however, there is less cause for concern. If we can use natural science to account for the mechanism of empathetic practice, we may be able to naturalize empathy and, thereby, the terminology of the mental. This possibility has occurred to several philosophers writing prior to Price's discussion of subject naturalism. Picardi (2000, 131–2), for example, notes that literature on the debate between "simulation" and "theory-theory" approaches to folk psychology might help naturalize Quinean empathy . Picardi thinks that neither side of this debate could satisfy Quine's "behaviorist strictures." But I will argue that one of these options—"simulation" theory—has more affinity to Quine's behaviorism than Picardi suspects. Specifically, I will show that the "radical" simulation theory of Robert M. Gordon (1995a) offers the best promise for a thoroughgoing *subject* naturalization of "belief."

I will begin this chapter by presenting the nature of the dispute between advocates of theory-theory and advocates of simulation theory. I will show why simulation theory, rather than theory-theory, offers the best chance at a subject naturalization of "belief," and why Gordon's version of simulation theory, in particular, is the best-suited version to accomplish this. I will then examine two leading objections to Gordon's simulation theory—the problems of pretense and adjustment—and attempt to defend Gordon against them. In the end, however, I will argue that the best case for Gordon's view runs into a new problem. After presenting this problem, I will discuss some important philosophic lessons learned from an encounter with simulation theory, including the consequences of its failure to serve as a naturalization proposal vis-à-vis naturalized epistemology.

Theory-theory vs. simulation theory

The topic of folk psychology—the question of how human beings understand the minds and thoughts of others—is of considerable relevance in several areas of philosophy. In epistemology, the question of whether we must understand others beliefs as generally coherent and rational in order to interpret individual beliefs is relevant to arguments against skepticism (such as those advanced by Davidson). In philosophy of mind, the dispute between realism and irrealism about the mental often turns on the question of the meaning of mentalistic concepts. If belief-desire psychology, for example, is held to be largely false, it is thought by some that the theoretical concepts of "belief" and "desire" might not refer, with the consequence of irrealism.

The dominant theory of folk psychology for the last century has come to be called the "theory-theory." According to the theory-theory, one person's attribution of a mental state to another essentially involves conceptual representations of that state, usually via the grasp of inferential relations between a theoretical entity and its *explananda*. The theory-theory has been widespread: it is the view of folk psychology presupposed by the classical position that we infer the existence of mental states by analogy to our own introspected states, but also by the more recent socio-linguistic

(Sellarsian) view that claims the reverse: that understanding our own mental states is itself a byproduct of a theory positing states that explain the behavior of others.

In recent years, an alternative to theory-theory has emerged, one which could have important implications for philosophical positions that turn on questions of folk psychology. "Simulation theory" is the name for a family of views formulated independently by Robert Gordon (1995a) and Jane Heal (1995), and subsequently developed by Alvin Goldman (1995; 2006) and others. Simulation theory claims that the more primitive attributions of mental states involve no new representations, but only the use of first-order (non-mentalistic) descriptive abilities, imaginative projection, and decision-making capabilites.[25]

To consider a simplified description of the process simulation theory envisions, suppose, for example, that we want to describe Maxi's state of mind after he opens the drawer and discovers that the chocolate is missing. Rather than drawing on a theory relating perceptual inputs to behavioral outputs, we simply imagine ourselves in Maxi's position. We see ourselves opening the drawer, then we feel disappointed, and then perhaps we decide to say to ourselves that the drawer is empty. Of course we make this decision "offline" and do not act on it. Drawing on our simulation, we can now ascribe the associated feelings and beliefs to Maxi, and perhaps predict his behavior.

In this chapter, I wish to focus on the "radical simulation" theory of Robert Gordon, because I take it to be the most consistent departure from theory-theory, and therefore a good test case for evaluating the difference between the simulation and theory paradigms. Other versions of simulation theory, such as Alvin Goldman's, for example, do not attempt to account for the *meaning* of mentalistic concepts from the ground up in the way that Gordon's theory does. Goldman presupposes,

[25] In fact a useful way of understanding the difference between theory-theory and simulation theory is just to hold that the former thinks that decision-making capacities are separate from folk psychological prediction mechanisms, whereas the latter thinks prediction capacities are just a part of our decision-making capacities (Davies and Stone 2001). See also Nichols and Stich (2003). At least according to Gordon (1995c) and Goldman (1995), this means that the question of simulation vs. theory can be settled empirically: one need only show that the same neurological mechanisms involved in decision-making (such as those involved in motor skills) are components of systems involved in the response to others (such as the "mirror neurons" activated in response to the perception of other's action) (Decety 2002).

for example, that the ability to engage in simulation requires a prior grasp of certain basic mentalistic concepts, which are needed to describe the instrospective results of one's own simulation and ascribe them to others, and also to assign the "inputs" needed to adjust a simulation to approximate the differing conditions of the target of simulation. Whether or not Goldman's presuppositions are acceptable on their own terms, they certainly make it difficult to use his version of simulation theory as a subject-naturalization of belief. Even though Goldman's own account of naturalized epistemology is in dire need of a naturalization of belief (see chapter 1), he is not interested in using simulation theory to naturalize mentalistic concepts. He simply wants to explain our application of these concepts to other people without the need to invoke overcomplicated theory-theoretic generalizations:

> In the very posing of my question—how does the interpreter arrive at attributions of propositional attitudes (and other mental states)—I assume that the attributor herself has contentful states, at least beliefs. Since I am not trying to prove (to the skeptic) that there is content, this is not circular or question-begging. I assume as background that the interpreter has beliefs, and I inquire into a distinctive subset of them, viz., beliefs concerning mental states. (2000, 75)

> As far as the interpretation strategy is concerned, it indeed appears that if the simulation theory is correct, the interpretation strategy is fruitless. One cannot extract criteria of mentalistic ascription from the practice of interpersonal interpretation if that practice rests on a prior and independent understanding of mentalistic notions. (2000, 94).

Gordon's approach, on the other hand, is "radical" because it (1995c; 1996) denies that either introspection or a prior grasp of mental concepts is needed to simulate or ascribe mental states. According to Gordon, simulation only requires the imaginative projection of oneself into the *situation* of the other. To state the contents of one's own beliefs, one need only ask and answer first-order questions, and append the answers to expressions of belief. One need not ask oneself "Where do I believe the chocolate is?" but simply "Where is the chocolate?" and give the answer, "The chocolate is in the drawer." One then engages in an "ascent routine" in which one appends an "I believe" construction to this result. Therefore all that is needed to attribute beliefs to *others* is to engage in this

ascent routine while *in the context of simulating the other* (appending "he believes" instead of "I believe" to the result).

Now it may well be that a hybrid of simulation and theory is needed to understand our mental concepts in all their richness, in which case an approach like Goldman's may be appropriate. But to get to the point of accepting that, it is necessary to see where a purely simulation-based approach breaks down, and to see that, it is best to examine Gordon's theory. Furthermore, since Gordon's "ascent routine" account of simulation does not require any appeal to introspection, it is also the version of simulation theory most likely to find favor with naturalists in the philosophy of mind.

In this chapter, therefore, I will examine Gordon's theory exclusively and examine some leading objections to it. I will begin by examining some of the experimental evidence that motivates simulation theory, but which also poses preliminary challenges to it. After showing how simulation theory approaches that evidence, I will proceed to examine a serious problem that arises for Gordon's theory, the problem of adjustment, and speculate about how he could address it. In the final section, however, I will argue that Gordon's best solution to the problem of adjustment faces a new problem. After presenting this problem and indicating why I don't think Gordon can solve it, I will present an alternative non-simulation explanation, one which is, nonetheless, not the same as the traditional theory-theoretic explanation.

Preliminary challenges from false belief task evidence

In the widely-replicated "false belief task" experiment, researchers have children watch Maxi (a puppet) place his chocolate in a drawer and leave the room (Wimmer & Perner 1983). After he leaves, his mother moves the chocolate to another location, and then Maxi returns. The children are asked where Maxi will look for the chocolate. Five-year olds typically say Maxi will look where he last saw the chocolate: in the drawer. But three to four-year olds answer that Maxi will look where

they know the chocolate to have been moved by the mother, indicating some difficulty in understanding or ability to express Maxi's false belief.

This pattern in the false belief task is very often taken as *prima facie* evidence in support of the theory-theory. Early failure at the task seems to suggest a *conceptual* deficit which is remedied by the child's development of a new "theory." In recent years, the relevance of the false belief task evidence has fallen into some doubt, particularly as experimentalists have attempted to simplify the task in a way that controls for difficulties children may have with its purely verbal aspects (Bloom and German 2000; Onishi and Baillargeon 2005). Nevertheless, the view that the task suggests some important conceptual development remains paradigmatic, as recent meta-analyses controlling for task-performance difficulties continue to show a pattern of development (Wellman et al. 2001), and as doubts linger about the significance of the early competence experiments (Perner and Ruffman 2005).

In any case, although recent simulation theorists such as Goldman (2006) draw on recent experimental developments to challenge the significance of the false belief task, at least at an early stage, Gordon (1995a) thought that the false belief task results could be equally or even better explained by simulation theory. He suggested that theory-theory could not explain the development through the acquisition of a new theory, because prior to that acquisition, a child would simply be *unable* to make predictions about human behavior, rather than what actually happens: making *unreliable* predictions (69–70). Now this objection is probably addressed by the theory-theorist's contention that children may have an early theory accounting for the possibility of unreliable predictions: instead of thinking of the mind as a representational device (as they do later), young children could think of the mind as a "copying" device, on which only real objects impress themselves.

But even if theory-theory could explain early false predictions adequately, Gordon (1995a, 70) could still respond that simulation theory offered a *better* explanation:

> Suppose…that the child of four develops the ability to make assertions, to state something as fact, *within the context of practical simulation*. That would give her the

> capacity to overcome an initial egocentric limitation to the actual facts (i.e., as *she* sees them). One would expect a change of just the sort we find in these experiments.

In other words, prior to passing the "false belief task," the child only has the ability to engage in the ascent routine without being able to project himself into the position of the other. From this perspective, the simulation theory explanation of egocentric error seems *simpler* than theory-theory's. Theory-theory has to posit a special children's theory of mind just to account for these errors, whereas simulation theory can simply draw on one's existing descriptive, and decision-making abilities, whose existence no one would dispute. Of course what accounts for the alleged deficit in the imaginative capacity is a matter of some controversy, but it is also not controversial that we eventually develop this capacity, and plausible that young children don't start out with everything.

The important question, then, is whether simulation theory can offer an adequate account of which particular imaginative capacity develops in such a way as to permit the eventual passing of the false belief task. And hopefully, the simulation theorist can offer this account in such a way that retains its simplicity advantage over the theory-theory. Positing unlikely and complicated imaginative capacities could counteract whatever simplicity advantage simulation theory has because of its reliance on existing descriptive and decision-making capacities.

Gordon's musings on this subject are only barely suggestive. He says that if, in the context of simulation, a child is led by evidence from the situation of the simulation to make assertions that conflict with assertions from one's "home" perspective, she will be making *de facto* "motivated attributions of false belief" (1995a, 61). There are questions we should ask immediately about this proposal. First, what kind of evidence is supposed to motivated assertions incompatible with one's home beliefs (hereafter, "home-incompatible assertions? The answer to the first question is simple enough. Presumably a child is motivated to make a new and different assertion because of some anomalous action on the part of the simulation target. The child may see Maxi headed to the "wrong" place: the place where the child knows there is no chocolate. Frustrated by the anomaly, the child looks for an explanation. Of course this would not yet explain children's ability to *predict* that Maxi

will go to the wrong place (since in the case of a prediction, the anomalous action has not yet occurred). But perhaps the ability to make such predictions is strictly theoretical, deriving from a simulation-based capacity to give these situation-based explanations.

Next, however, we should ask what imaginative capacities are available to enable this sudden development in the ability to make home-incompatible assertions—and in a way that would allow for the development of the genuine understanding of false beliefs. This is an important question because if the situational evidence that motivates the child to make a home-incompatible assertion is merely the target's anomalous behavior, there is at least one explanation a child could invoke that would explain the anomaly by way of an incompatible assertion, but which would not be the same as a false belief explanation. Supposing that Maxi is seen heading to the "wrong" place (where the chocolate *used* to be), the child could just as easily explain this by making an assertion about what Maxi is *pretending*: that the chocolate is in the drawer.

Now some psychologists do see pretense itself as involving some grasp of representational states in others, but it seems there are some important differences between pretense and false belief that would call this into question. Paul Bernier (2002) thinks that not just *any* mental process leading to home-incompatible assertions suffices for the ascription of beliefs that the simulator could genuinely comprehend as being false. To genuinely comprehend a mental state as a belief, Bernier notes, the simulator needs to know something about the function or aim of the state. Beliefs, in particular, are mental states that "aim at describing one and the same objective reality" (42). [26] A mental state counts as a false belief only insofar as it fails in this aim. But of course pretense does *not* aim at describing objective reality, and so a motivated attribution of a pretense which fails to correspond to reality is not yet the attribution of a representational state: pretense does not misrepresent reality because it is not even *trying* to represent reality.

[26] Jacob (2002, 100-3) seems to be making a similar criticism of Gordon.

In fact, one study by Perner et al. (1994) suggests that whereas most 3-year olds can discriminate between cases of knowledge and cases of pretense, they cannot discriminate between pretense and false belief. Young children surely display that they know their make-believe is not real at an early age (Perner 1991, 66–7, 76–78).[27] Children first acquire the ability to engage in knowing pretense as early as their second year, fully two years before passing the false belief task (Perner 2000, 385). So even if there is some rationale according to which understanding pretense involves understanding a representational state, it is not the *kind* of representational state the grasping of which is necessary to explain the ability to pass the false belief task. If we are trying to explain that ability, then, by some mechanism that leads to home-incompatible assertions, we had better specify that mechanism enough so that it explains more than pretense-attribution can explain.

Gordon, it seems, is aware of this problem, and recognizes that the mere ability to make home-incompatible assertions is not enough to count as the genuine comprehension of false-beliefs. He tells us that the route to a more "sophisticated" understanding of "belief" comes through a novel kind of simulation within a simulation:

> To see her own present beliefs as distinguishable from the facts she will have to simulate another for whom the facts are different—or, more broadly, adopt a perspective from which the facts are different, whether this perspective is occupied by a real person or not—and then, from the alien perspective, *simulate herself.* (1995c, 61–2)

This, however, does not yet seem to give us what we want, because the main question is how, in the first place, a child *can* come to "simulate another for whom the facts are different." The child cannot do it simply by engaging in the "ascent routine" on the basis of the child's own apprehension of the facts, and it also cannot be done through ascription of pretense to the other.

Now the second, broader suggestion, to "adopt a perspective from which the facts are different, whether this perspective is occupied by a real person or not," is slightly more suggestive, if only because it is broader. If, by "different facts," Gordon means "*incompatible* beliefs about the

[27] Children first acquire the ability to engage in knowing pretense as early as their second year, fully two years before passing the false belief task (Perner 2000, 385).

facts," then we are right back where we started. But perhaps Gordon means a different kind of difference, difference without incompatibility. Perhaps the child can simply imagine that he has more or less information about a given circumstance, and in some way the possibility of genuine false-belief attribution emerges out of this. The question of how a child might be able to do this, using existing imaginative capacities, is the problem we will examine in the next section: what Daniel Weiskopf (2005) has called "the problem of adjustment."

The problem of adjustment

To understand how existing imaginative capacities could be used to "adopt a perspective from which the facts are different" in the second sense I've described above, let's first present an example of Gordon's that is meant to warm us up to a kind of simulation explanation that could successfully generate the genuine comprehension of false beliefs.

Imagine that Sam (the simulator) and Tom (the target of simulation) are hiking a mountain trail together.[28] Suddenly Tom yells "Go back!" and Sam follows him, looking over his shoulder. Sam looks over his shoulder, looking for "menacing, frightening things" up the trail. Sam spots a grizzly bear. By looking at the same environment as Tom, and experiencing the same fear, Sam simulates Tom and understands why Tom runs: Tom runs *because of the bear*. Even though this is not yet anything like false belief explanation, it involves a kind of simulation that is more than attributing a pretense, as regards both its cause and effects: it is caused by the apprehension of an anomaly in Tom's behavior, and it has the effect of enabling further predictions about Tom's behavior (that he will continue to run away to safety once he is out of the bear's range), which predictions Sam can use to his own benefit (by following Tom).

Gordon calls this kind of simulation "total projection," because it involves the use of the total set of Sam's own knowledge, beliefs, and desires to understand Tom. No "adjustment" is needed,

[28] The names here are mine, not Gordon's.

since Sam is, for all practical purposes, in the exact same position as Tom. Granting for the moment, then, that "total projection" has explanatory power, we must consider whether simulation of situations more closely resembling false belief explanation have even more power. Gordon (1995b, 106) considers these situations to be ones in which "total projection" must be "patched". In these situations, the reliability of total projection fails, which failure itself constitutes a new anomaly necessitating a new explanatory tactic that will restore reliability. The *problem* of adjustment is just the problem of how total projection can be patched to provide the full extent of explanatory power usually attributed to false belief explanation.[29]

The failure of total projection is seen by theory-theorists as evidence for why interpreters need theory, not simulation, to understand the behavior of others. But Gordon (1995b) suggests ways in which total projection can be patched without resorting to theory. For example, if Sam sees Tom run from the grizzly bear, he may not at first understand the action *if the bear is heading towards Tom but away from Sam*. Since Sam is not in any danger himself, he cannot immediately empathize with Tom's action. The needed adjustment is easy enough: Sam imagines himself looking at the bear from Tom's perspective. Now he imagines the bear coming towards *him*, feels the fear, and can explain Tom's action as stemming from the oncoming bear (107). This almost functions like a false belief explanation, because "there is a bear coming towards me" is *false* from Sam's perspective. What he discovers is that it is not false from Tom's. But there is a question we can ask here and return to later: *how* is Sam prompted to consider what things look like from Tom's perspective?

To begin to understand why that is an important question to ask, consider a modified version of another one of Gordon's examples, this time one that involves Tom's being "epistemically handicapped." This time Sam sees the grizzly in Tom's path, twenty feet ahead, and expects him to run (having used the patching described above). But Tom does not run: he keeps walking towards the

[29] The problem of adjustment involves more than just adjusting to different informational situations. The simulator may also have different evaluations than the other being interpreted. See Weiskopf (2005) for more discussion of Gordon's attempted solution to this version of the problem, and the problems involved in his solution.

bear. As it happens Tom is myopic and sees only a blur ten feet in the distance.[30] Gordon would say that Sam can also explain this action by simulating Tom's blurry vision. Now Sam imagines that he doesn't see the bear in front of him, and doesn't want to run. This helps him grasp that Tom doesn't run because he doesn't see the bear. In another example, Sam sees Tom run from something he *shouldn't* run from: a large Newfoundland (the affable dog breed). Gordon proposes that Sam can explain the behavior by imagining the dog to be from some actually dangerous breed, or imagining the dog to look like a grizzly.

Gordon's strategy here is clearly inadequate. The first problem is that there is no obvious mechanism accounting for why Sam considers these particular hypotheses—or, more precisely, the particular pretend scenarios: to let his vision go blurry, or to confuse a Newfoundland with another dog or with a grizzly. There are many hypotheses that could also account for the behavior exhibited by Tom, and Gordon never tells us how any mechanism is supposed to pick one. Presumably if Sam is an adult, he has a great deal of theoretical knowledge that could be invoked, but at first glance it appears to be knowledge about Tom's mental states—and it is precisely this knowledge that we are not relying on in our attempt to explain how a *child* can come to attribute false beliefs in the first place. This is not a problem that attends the simpler types of simulation. Elsewhere Gordon reminds us of evidence documenting "imitative mechanisms" that lead young infants follow the gaze of others (joint attention), and imitate the facial expressions of others (emotional contagion). But he gives no evidence concerning mechanisms that could somehow lead us to lock on and simulate malfunctions or confusions in a putatively *internal* state of the other.

Perhaps this problem is an artifact of poorly chosen examples on Gordon's part. Consider a simpler example involving the hikers. Suppose Sam is far enough away to see a grizzly that Tom

[30] Gordon's original example involves the myopic hiker turning back anyway for an inexplicable reason. His point is that if the hiker is myopic, the ordinary explanation will not work, as he will not see the bear. The trouble with the example is that it is supposed to illustrate how simulating blurry vision could figure in an explanation, but it only succeeds in showing how a putative explanation can be dismissed. I have modified it so that the blurry vision actually explains the anamolous action.

cannot, because a boulder blocks Tom's line of sight. In this case, Sam could respond to the situation by placing himself in Tom's location, and then, by exploiting his joint attention imitative mechanism, look at the boulder rather than the bear. Feeling no obvious fear about the boulder, he would imagine himself walking forward rather than running away. This type of simulation seems to require no recourse to theory of mind, and yet provides something mimicking a false belief explanation: Tom doesn't believe there is a bear in his path, because he believes there is only a boulder in his path. This looks much simpler than Gordon's idea of the simulator as the young hypothesis-tester. Perhaps, by building on simple explanations like this, more complicated simulations are possible.

Unfortunately, the improved clarity of the example also brings a deeper version of the problem of adjustment to the forefront, one that applies even to the simplest examples of simulation that Gordon has offered us. We need to think more about the *motivation* of the simulator. Taking for granted that Sam is surprised by Tom's movement towards the bear, and is prompted to search for an explanation, what motivates him to imagine himself in Tom's *location* in the first place? This is the same question we asked even before we considered cases of "epistemic handicap," as when the bear was simply headed towards Tom rather than Sam. Now, appealing only to the joint attention mechanism, we can imagine why Sam would look at the rock. But from Sam's vantage point, he can see both the rock and the bear, and he isn't inclined to walk towards the bear as Tom is. So there are mechanisms for attending to what Sam is attending to, but attending to the same thing from different perspectives does not yield the same beliefs. There are mechanisms for imagining what it would look like from Tom's perspective, but none that account for the *motivation* for looking from Tom's particular perspective, as opposed to so many others. There are mechanisms for imitating the other, e.g., his facial expression, but none that account for the desire to imitate his location.

One is tempted to say, perhaps, that of course that location is relevant, because information about spatial location is relevant to how Tom perceives the world. But the question is: isn't this a piece of theoretical knowledge about Tom's mental states, rather than a product of simulation? Perhaps

advocates of the simulation theory who are willing to tolerate hybridized versions of it can admit as much, but it would seem that Gordon cannot—not if he wants to show how a primitive grasp of false beliefs is possible using nothing but simulation-based resources. In the next section, I will explore what I believe might account for the motivation to consider another's situation, and why recognizing it undermines simulation theory.

The problem of epistemological adjustment and the complexity of simulation theory

It would seem that the simplest explanation for wanting to imagine oneself in another's position is that Sam knows something about the connection between action and *knowledge*, specifically perceptual knowledge. He knows that people move towards good things they can see, and that they run away from the dangerous things they can see. So in order to determine what Tom can see from where he is, he simulates that position to explain Tom's action.

Gordon himself constantly insists that simulation involves the *implicit* attribution of knowledge to the other. In total projection, it involves projecting all of one's own knowledge to the other (1995b, 103). In patched projection, it may involve pinpointing "relevant ignorance" (1995b, 110–1). In saying this, however, his emphasis is on "implicit": knowledge attribution, which he thinks to be implicit in just the same way that belief attribution is supposed to be. Presumably simulators should be able to attribute it without the need for any explicit theory or concept of knowledge.

But Gordon's view neglects the fact that there is good reason to think that children form explicit beliefs about epistemic facts *before* they pass the false belief test. So there is good reason to think that any simulation with the power to render the same explanations as explicit false belief explanation *can* certainly draw on these explicit beliefs.

First we should say something about the formation of the concept "knowledge". It seems children make explicit mention of it at a very early age. Developmental psychologists have observed children using "know" as early as 15 months (Bretherton, et al. 1981). According to one study "know"

is the most frequently used mental term between 2 and 3 years of age (48% of all mental verbs in one child), and of these, "I don't know" is the most common construction among uses of "know" (62% of them, for the same child). Of course at the earliest stage children use the term 'know' merely as a conversational marker, in idiomatic phrases like "Know what?" or "You know?" Still, it is estimated that perhaps 12% of the earliest uses are in the context of a genuine assertion of correspondence with facts (Shatz et al. 1983) This is particularly clear when children don't simply say "I know," but contrast their usage with something they don't know, or talk about how they didn't used to know something, but now they do (Hogrefe et al. 1986). Clearly something like an "ascent routine" can facilitate the transition between idiomatic and genuine referential uses of "know." It's particularly interesting that in longitudinal studies, children regularly draw this knowledge-ignorance contrast in natural conversation months before they begin to draw belief-reality contrasts of the kind associated with passing the false belief test (Bartsch and Wellman 1995, 120).

Second, having an explicit concept of "know" enables children to develop explicit knowledge about the sources of knowledge. There is certainly a period of development in which children have *merely* implicit knowledge of the connection between perception and knowledge. In one study, children are asked which of two characters knows what is in a box: the one who has looked, or the one who has not. By the age of three, most children answer the question correctly (Pratt and Bryant, 1990). But in another study, while three year-olds can correctly identify the lookers as the knowers, the majority are not able to explain how it is that lookers know (they cannot explain where knowledge comes from). Only later, between the ages of four and five, do they acquire the ability to give explicit explanations (Wimmer et al. 1988; Gopnik and Graf 1988). Particularly interesting is a very recent study on the specific relationship between children's ability to offer epistemic explanations and their ability to pass the false belief test. The results suggest that while there is an initial period in which children fail to offer epistemic explanations and fail to pass the false belief test, there is a transitional period in which they can offer epistemic explanations without yet passing the false belief test (Burr

and Hofer 2002). So there seem to be three stages in the development of explicit knowledge: 1) drawing the knowledge-ignorance contrast, 2) identifying perceivers as knowers, and 3) drawing the knowledge-false belief contrast.[31]

It would be rash to conclude that simply because explicit epistemological theory precedes explicit "theory of mind," therefore children must be able to pass false belief tests *because* of development in their own epistemological theories. But this conclusion becomes more tempting if we can offer a hypothesis accounting for the conceptual process that would account for it. Recall that the fundamental distinction between understanding false-belief-motivated action and pretense-motivated action is that only the first involves the understanding of incompatible propositions aimed at the same reality. A child will see no contradiction between his own belief that there is no chocolate in the drawer, and Maxi's pretending that there *is* chocolate in a drawer.[32] But if an (older) child thinks that Maxi is in error, he thinks Maxi believes there is chocolate in the drawer, even though the child knows there isn't. This is because he understands Maxi's attitude as *aiming at reality*, and failing in its aim. In short, "believe" involves a *normative* element that "pretense" does not. I want to suggest that it is the child's *explicit* understanding of the causal connection between perception and knowledge that makes possible understanding the normativity of belief.

It is important that children begin with the contrast between knowledge and ignorance. They understand what they have (knowledge) and what they don't, and their naturally curiosity pushes them to get "more" of what they don't have. So knowledge is a goal of theirs: the question is, what is the

[31] There is of course a legitimate question about whether some of this explicit knowledge about the sources of knowledge might itself result from simulation. I have already noted how a child may form that concept through a kind of ascent routine. And there is possibly some connection between a child's understanding of other perceivers as knowers, and joint attention mechanisms, which are doubtless involved in seeing where others are looking. But it is not at all clear why simulation would be necessary for any of this. Children explicitly understand what knowledge is by reference to their own mental states. They surely also connect their own perceiving to their own knowing. I can imagine that simulation may play a role in discovering that other people know. They might notice that others have eyes, too, and consider themselves in the place of others—and conclude that others must know of particular things. But this would not account for the *concept* "know" or any *principled* knowledge about the sources of knowledge; it would only account for particular knowledge about who knows what. What's more, it's not at all clear why a simple argument by analogy, rather than simulation, would not account even for this particular knowledge.

[32] Indeed, we should wonder if pretense even requires any attitude towards a proposition at all.

means? This they begin to realize when they understand that knowledge originates in perception. Along with this comes an understanding of the *limitations* of our epistemic faculties: the understanding that one can only see one side of an object, a limited stretch of the environment, etc. By grasping these limitations, the child comes to see that knowledge is not just a goal, but a goal that must be achieved through effort–and that seekers of knowledge can fail in their quest. It is also plausible that learning about the limitations of perception, in particular, can help a child to understand how appearances can be misleading—a crucial point needed to understand how one can have a mental state that functions like knowledge even when it is not knowledge. Children, indeed, grasp the distinction between perceptual appearance and reality before they pass the false belief test (Gopnik and Astington, 1988, 34.) Once a child grasps that certain of his own mental states have functioned like knowledge in the past (e.g., had the same types of origins and applications) even when they later turned out to work more like ignorance (because they yielded unsuccessful predictions), he needs a concept to denote these states to explain unsuccessful actions resulting from them. This concept not only offers the aforementioned explanatory value, but also does so in a way that accounts for the difference between false belief and pretense. The child can now see false beliefs as aiming at reality in a way that pretense does not.

If this hypothesis is correct, then an interpreter does not need simulation to grasp the possibility of false beliefs. Of course supporters of the simulation theory do not argue that children *need* simulation to understand the concept of belief. They merely claim that simulation theory offers a simpler explanation of where that understanding comes from than theory-theory does. First of all, as I have shown, they claim that simulation explanations can draw on existing descriptive, decision-making and imaginative abilities, without the need to posit special childhood theories. Second, the *content* of the theories children would need to explain behavior be hopelessly complex in comparison to simulation theory. Gordon argues that if the folk have to rely on behavioral generalizations to explain and predict behavior, they will at best be able to generalize about "typical" behavior patterns,

but at the same time they would have to allow for countless *ceteris paribus* clauses. Gordon (1995a) certainly allows that behavioral generalizations are sometimes needed, but insists that the simplest way to understand how *ceteris paribus* clauses are filled in is by using them in the context of simulation, such that one can use one's own practical reasoning skills to predict the best course of action in exceptional circumstances (67).

But if we allow that interpreters also have epistemological generalizations about their targets—i.e., explicit knowledge about the connection between perception, knowledge and action—simulation is no longer needed to simplify the application of these existing behavioral generalizations. Consider how the use of simulation might be used to simplify the application of a behavioral generalization:

> S1. People run from danger (*behavioral generalization*).
> S2. Tom is approaching a dangerous bear (*anomalous observation*).
> S3. I only act to avoid danger that is present (*behavioral generalization*)
> S4. There is no dangerous bear present [said from Tom's perspective] (*simulation*)
> S5. Tom is walking towards the bear because he believes: there is no dangerous bear(*new explanation resulting from the ascent routine*).[33]

But now consider how the attributor's practical knowledge can be used just as easily to inform the application of a behavioral generalization to the situation, without the need to have countless *ceteris paribus* clauses:

> T1 People run from danger (*behavioral generalization*).
> T2. Tom is approaching a dangerous bear (*anomalous observation*).
> T3. People only act to avoid dangers they know about (*epistemological generalization*).
> T4. Tom does not know about the bear which is not in his line of sight (*epistemic observation*).
> T5. Tom is walking towards the bear because he believes he is safe (*new explanation*).[34]

---

[33] Now it is possible that the epistemological generalization in (3) is a bit more specific than we would expect a young child to possess. But it is a conclusion reached easily from "people act on what they know," combined with "people act to avoid dangers." The observation in (4) is of course an application of another epistemological generalization, "people cannot see around solid objects." There is no reason to think a younger Sam would need to consider all of these more general generalizations explicitly at the moment in order to understand Tom, but it is plausible that he would have them stocked in his background knowledge explicitly at some point in the past.

[34] Now it is possible that the epistemological generalization in (3) is a bit more specific than we would expect a young child to possess. But it is a conclusion reached easily from "people act on what they know," combined with "people act to avoid dangers." The observation in (4) is of course an application of another epistemological generalization, "people cannot see around solid objects." There is no reason to think a younger Sam would need

Each of these processes involves parallel steps, and works by applying one's own practical knowledge to a situation (thereby obviating the memorization of complex exceptions). The difference is that while all of the steps in the second process are likely to be known or knowable by young children, step S4 in the first is knowable only if one adds the complexity of a special mechanism that motivates the simulator to envision himself in Tom's position, a mechanism that is not obviously accounted for by any currently known cognitive mechanisms (even though there are known mechanisms involving imagination, joint attention, and facial imitation).

So if simulation is not needed to simplify the application of behavioral generalizations—and if simulation mechanisms require positing new mental mechanisms to explain the motivation to imagine oneself in another's position—simulation theory loses its chief explanatory advantage: it's simplicity. It is not simpler than the explanation involving explicit epistemological generalizations, and may in fact be more complex. The use of explicit epistemological generalizations requires no new cognitive mechanism, just the usual kind: background conceptual generalizations of a type which children are likely to have.

It should be noted that while the style of explanation I have present above is not a version of radical simulation theory, this does not mean that I think simulation never has a role to play in folk psychology. My point about how epistemological generalizations can be used to offer theoretical explanations can apply equally well to simulation: some simulation may be very important in explaining action, but it still requires these epistemological generalizations, and these involve concepts of mental states. The upshot is that any account that attempts to derive the possibility of genuine false-belief attribution and comprehension from purely non-conceptual simulation—like Gordon's "radical" simulation—is doomed. No matter what we say about the relevance of simulation, it must be supplemented by other types of "mindreading." Some kind of hybridized account is necessary. My

---

to consider all of these more general generalizations explicitly at the moment in order to understand Tom, but it is plausible that he would have them stocked in his background knowledge explicitly at some point in the past.

suspicious is that this hybrid account does not require a *prominent* role for simulation, but that is another matter.

What's more, I think that rejecting the possibility that simulation is *basic* to the grasp of mental concepts does not mean that the content of all mental concepts is itself "theoretical," at least in the functionalist sense that is so often associated with the theory-theory. If I am correct that the concept "knowledge" can be formed by contrast with examples of ignorance, for instance—and the concept of "belief" can then be derived somehow from "knowledge"—then there is at least one mentalistic concept, "knowledge," which is not functionalistic, even if others (like "belief") are.

Conclusion: Implications for deflationary naturalized epistemology

The possibility that children might grasp concepts of intentional mental states by first grasping *epistemic* states has not been widely entertained by psychologists or philosophers, probably because of the longstanding tradition in philosophy of thinking of knowledge as requiring analysis in terms of (justified true) belief, not vice-versa. But this tradition has recently been challenged by epistemologists like Timothy Williamson (2002), who argues that "know" should be understood as a primitive, whereas "believe" should be understood as putative knowledge or as a functional equivalent of knowledge.[35] If my account of folk psychology as originating in folk epistemology is correct, there is psychological support for Williamson's hypothesis. Simulation theorists will be disappointed, but theory theorists will also need to reconsider the widespread view that philosophy of mind is independent of epistemology.

Interestingly, in considering simulation theory as a naturalization of "belief," we have also now come full circle. Naturalized epistemologists begin with an interest in naturalizing knowledge. Because philosophic tradition says that knowledge is to be understood as a kind of belief, and because

[35] The present account of how the concept of "believe" develops from the concept "know" also helps enrich Williamson's claim that "know" has greater psychological explanatory value than "believe": if we need to understand what a subject knows and what he does not know in order to understand his beliefs, then knowledge attributions explain everything that belief attributions explain, and then some.

understanding belief has always posed an independent challenge to naturalists, naturalists then seek to naturalize belief. In chapter three, we explored various proposals for naturalizing belief in an object-naturalist style, but found them lacking. In the present chapter, we have now explored the most likely subject-naturalist proposal, simulation theory, and found it lacking, too—because there seems to be good reason to think that simulation itself cannot be performed or understood without the possession of epistemic concepts. In a way, this consideration calls the very project of naturalized epistemology into question. If the concept of "knowledge" is so basic as to ground even our understanding of "belief," perhaps it is not a concept that is in need of explanation, much less naturalistic explanation.

At this point, however, naturalists might reasonably ask whether knocking down simulation theory as a naturalization proposal is knocking down a straw man. Perhaps there are other subject-naturalist proposals out there worth examining, which may not depend on folk epistemology in the way that simulation theory seems to. This point is fair enough. In what remains of this section, I want to argue that even if there are other interesting descriptions of folk psychological practice that might be conducive to subject naturalization of belief, the facts we have presently uncovered about the apparent dependence of folk psychology on folk epistemology have some interesting implications for deflationary naturalism about *knowledge*.

In chapter 1, I examined the views of Michael Williams (1996), whose views I classified as a kind of pessimistic naturalized epistemologist. Although Williams does not characterize himself as a naturalized epistemology, he does take a deflationary approach to knowledge that would recommend itself to the subject naturalist, in the manner described by Price. Williams opposes what he calls "epistemological realism," the view that knowledge is a real thing common to all of the instances we call "knowledge." Instead he argues that the task of epistemology, if anything, is to examine our actual practices of knowledge attribution, for instance by examining how standards of justification shift from the context of one discipline to another. The case for this kind of deflationary naturalism about knowledge of course depends on the idea that knowledge is *not* a real thing (or that we have no reason

to think that it is), and Williams supports this by claiming that cases of knowledge have no "theoretical integrity." What I will urge at present is that, interestingly enough, the evidence we have unearthed to oppose the deflationary view of "belief" helps to undermine Williams's deflationary view about knowledge by showing how knowledge *can* have theoretical integrity, though in a different way than Williams imagined it might.

Williams supports his contention that "knowledge" has no theoretical integrity by giving the example of Francis Bacon's early account of heat, which begins by listing a number of examples of heating: heating by radiation, by friction, by exothermic reactions, and by hot spices on the tongue. This list, says Williams, is at best a nominal kind. From the fact that we call all of them "hot" does not mean they all possess the property of *heat*. Likewise, simply because we say we know a number of things does not mean that there must be a theory of all things called "knowledge" (1996, 106–7). Of course in the case of heat, we know that there does end up being *some* coherent natural kind, described by the kinetic theory of heat: it just doesn't necessarily correspond to our pre-theoretic intuitions about "heat." So Williams goes deeper still, to argue that simply because a concept is teachable does not imply that it refers to something real. He mentions that the distinction between analytic and synthetic is teachable (a point made famous by Grice and Strawson in their critique of Quine), but that this does not imply that there really is a distinction between sentences called "analytic" and ones called "synthetic." For after all, the distinction between "witch" and "non-witch" was once teachable, but there are no witches (107–8). Presumably, (I am filling in some gaps in the argument here) Williams's point is that while scientists eventually found an underlying theoretical unity to *much* of our concept of "heat," there was none to be found for "witch." I take it that the application to "knowledge" is that, since scientists have not found anything equivalent to the kinetic theory of heat for "knowledge," the concept is more like "witch" than it is like "heat." Without any positive grounds for believing in theoretical integrity of the concept, if all we have learned is the mere ability to use the concept to make

various distinctions (like that between knowledge and ignorance) we have no reason to believe that "knowledge" refers to anything real.

But Williams considers an objection to this argument. Simply because we do not have a theory of knowledge anything like our theory of heat does not mean that knowledge does not exist as a real thing. Consider things like tables and chairs. We have merely "loose, functional classifications" of these everyday objects, nothing like a rigorous physical theory about them. Yet we would never say tables and chairs do not exist (109). Williams responds that this objection assumes that "knowledge of the external world" is more like "chair" than it is like "witch." What causes "witch" to fail to refer, he says, is its status as an "essentially theoretical" term. Essentially theoretical terms, says Williams, are ones which "we see no point in continuing to make, or even no way of drawing, once the theory behind them has been rejected" (109). Clearly what Williams has in mind here is very similar to the descriptivist theory of reference discussed in chapter 3, as exemplified in Lewis' view of the reference of theoretical terms.

Williams then needs only to argue that "knowledge" is essentially theoretical, like "witch" but unlike "chair" (and, presumably, that the theory behind it has no support). He claims it is essentially theoretical, because there is "no commonsense, pre-theoretical practice that this way of classifying beliefs rationalizes: its sole function is to make possible a certain form of theoretical inquiry, the assessment of knowledge of the world as such" (110). The concept that Williams says is essentially theoretical, however, is "knowledge of the external world," which he takes to be understood in the Cartesian sense in contrast with "experiential knowledge." This point recalls much of his previous development, which rejects Cartesian-style arguments from the "priority of experience" to conclusions about the external world, and with them, the possibility of foundationalism. The possibility of foundationalism, Williams has argued, is the one hope the concept of "knowledge" has for exhibiting "theoretical integrity": if all of our knowledge can be shown to reduce somehow to the senses, this would be an important fact that all of it has in common. So if we follow Williams in rejecting

foundationalism, and if its validity is a commitment of the concept of "knowledge of the external world," then it seems we should indeed decide that the very concept of "knowledge of the external world" fails to refer to some real fact common to all things we call "knowledge of the external world."

It is important to note, however, that this position assumes that the only relevant concept of knowledge is the Cartesian concept, and that the only relevant kind of foundationalism is the sort based on the "priority of experience." Arguably, there are other versions of foundationalism available which may not face the same problems as traditional Cartesian foundationalism. In particular, I have in mind direct realist foundationalism, which holds that basic beliefs are not beliefs about experience, but beliefs about ordinary middle-sized objects. (In general, I think many of Williams's objections to foundationalism trade on a confusion between versions which claim that *beliefs are based on experience* (which is definitive of any kind of empirical foundationalism) and versions which claim that *beliefs are based on beliefs about experience* (which is specifically Cartesian foundationalism). Now in fact I offer some arguments for direct realist foundationalism in my final chapter, but I do not need them presently to make the following point. As long as there are distinguishable versions of foundationalism, the failures of Cartesianism do not imply the failures of every foundationalism, and there is hope for the theoretical integrity of "knowledge" yet.

I can take this argument one step further. Even if many philosophers have assumed that the concept "knowledge" has Cartesian implications, this does not mean that the concept itself is committed to these implications. Just because the concept is not as non-theoretical as "chair" does not mean it is as theoretical as "witch." After all, what about "heat"? Williams himself seems to think that "heat" is something of a middle case between essentially theoretical and non-theoretical. Even though it was originally taken to have implications that we eventually rejected (e.g., that heat was caused by caloric fluid), we did not conclude that there is no heat. Williams calls concepts like these "classifications that have been theoretically rationalized but which retain independent utility" (110). "Distinctions like this," he says, "are apt to survive the rejection of theories with which they have

become associated" (110) Williams does not explain what makes this middle case possible; perhaps it happens when a theory's core commitments are retained while others are rejected.

Whatever is to be made of this possibility, Williams at least thinks highly of it, and offers to show that it does not apply to "knowledge." He says that the concept has no "pre-theoretical utility" or a "theory-independent way of drawing even approximately the right boundaries around it." But here I think Williams is just wrong, and the evidence we have considered about the dependence of folk psychology on folk epistemology proves it. That evidence suggests that at an age when children are far too young to have read Descartes, or even to consider the idea that they are only aware of their internal experience, not the outer world, they are still "little foundationalists." They come to see perception as the source of knowledge, and cite it as explaining how they know various claims. Not only that, but we have also now seen that being able to attribute knowledge or the lack of it is a prerequisite of attributing beliefs, and insofar as attributing beliefs has predictive and explanatory power, then so too does attributing knowledge or lack of it. Finally, the suggestion that there is not even an approximate way of drawing the boundaries of the concept knowledge is just false. Perhaps it is difficult to draw the boundary between "knowledge of the external world" and "experiential knowledge" (where the latter is taken to mean knowledge *about* sensory experience), but this presupposes an unnecessary (and unworkable) Cartesian foundationalism. If children start out not with beliefs about that contrast, but simply about the contrast between knowledge (full stop) and *ignorance*, that is a perfectly acceptable basis for drawing the boundaries of the concept, even if it does not tell them everything they need to know right away.

So what I am suggesting is that even if philosophers' theories of knowledge have had unfortunate implications in the past, studying actual folk epistemological practices (as the subject naturalist tells us to do!) reveals that these implications do not exhaust the theoretical integrity of the folk concept of "knowledge." For that reason, it is really much more like "heat" than it is like "witch." As a result, there is more reason to think that difficulties in philosophers' theories of knowledge make

no difference to the fact that there is knowledge. As a consequence, Williams's case against epistemological realism is undermined, and the motivation for his deflationary naturalism about "knowledge" loses its motivation. This is why undermining the version of deflationary naturalism about belief that we have discussed is also relevant to undermining deflationary naturalism about knowledge—even if there are other possible versions of belief deflationism available that I have not yet considered.

Now that the first form of pessimistic naturalism has been called into question—and given that optimistic naturalisms have been discredited since chapter 2 and 3—we have no choice but to turn to the final version of naturalized epistemology: Quinean pessimistic naturalism. In the next chapter, I will examine what distinguishes this version of naturalism from others, by describing its aims and its roots. In the final chapter, I will argue that its roots grow from both scientific and philosophical errors, and that if we can conceive alternatives to them, we need not resort to Quinean pessimism—or any version of naturalism, for that matter.

Appendix: Gordon on reason explanations and counterfactuals

In the chapter above, I think I have called into question the possibility of a simulation-theoretic explanation of the predictive and explanatory power of folk psychology. There is, however, recent work by Gordon that might be taken to establish that simulation has greater explanatory resources than I might originally have thought. In more recent work, Gordon has put forth a view of simulation theory that might circumvent some of the problems I have raised above. In "Simulation and Reason Explanation: The Radical View" (2001), Gordon contends that radical simulation can provide important folk psychological explanatory power, provided that we retool our understanding of *explanation*. Perhaps, even if simulation does not have the *same* explanatory power as false belief attribution, it still has *some* explanatory power, and its function in our mental economy is worth considering after all.

Part of Gordon's project in "Simulation and Reason Explanation" is to defend his older view of simulation from a different objection: "explanations" produced by simulation appeal to *reasons for action*, not *causes*. Davidson (1963), however, has argued that there is an important difference between the first and the second, and that only the appeal to the causes can amount to an explanation. The presence of the grizzly bear in Tom's path may present a *reason* for him to run, a *justification* for his quick departure, but its presence does not necessarily explain his action. This last point is particularly salient for cases of overdetermination. When there is more than one reason for action present in a particular case, it is impossible to appeal to only one as the explanatory factor. It seems, therefore, that the only way to explain an action is nomological, to subsume it under some law of nature. To explain human action, the usual strategy is to subsume behavior under laws relating mental states and actions—not laws relating objects in the environment (like *bears*) and actions.

Of course the deductive-nomological view of explanation has suffered from numerous philosophical problems in recent years (see Salmon (1998)). Cognizant of this and of his need to find an alternative account of psychological explanation that overcomes the overdetermination problem, Gordon considers *counterfactual* explanation. An example of an overdetermined action would be Gordon's braking at an intersection. One reason to brake is that there is a red light; another is that he is driving ten miles over the limit. Both provide a good reason to brake, but which reason explains braking? Gordon says that it *is* the case that if there hadn't been a red light, he would not have braked. But it is not the case that if he hadn't been traveling over the speed limit, he would not have braked (for even if he hadn't been speeding, he might have had a red light). So appealing to the counterfactual differences between competing reasons for action does seem to dissolve the overdetermination problem, and explain actions.

Gordon says that this kind of counterfactual explanation is easily adopted in the context of simulation, for the purpose of explaining human behavior. We can see that already in the example concerning the explanation for Gordon's braking. Suppose that Sam wants to explain Gordon's action.

Sam can imagine himself in Gordon's situation, and begin to imagine counterfactual possibilities. In one iteration, he imagines himself driving in Gordon's car, over the speed limit and through a green light at the intersection. He does not brake. Then he imagines himself driving the speed limit, only this time the light is red. He does brake. Because of that counterfactual difference, in the context of his simulation, he is able to say that he brakes *because of the red light*, not because of driving over the speed limit.

Of course Gordon (2001) acknowledges that this kind of counterfactual explanation does not account for the kinds of behavior that false belief explanation might account for:

> Where the explanans . . . is a reason-in-favor, then for the explanation to be correct, I had to have known or been aware that there was smoke; therefore, I had to have believed there was. … In interpreting the counterfactual that corresponds to a reason explanation, we consider only worlds in which the counterfactual condition c, specified by the antecedent of the conditional, lies within the agent's epistemic horizon: the agent knows or is aware that c—and, therefore, believes that c. We don't allow the counterfactual 'facts' to vary independently of the agent's beliefs.

So it is clear that this counterfactual reason explanation adds nothing to overcome the "problem of adjustment" I have discussed in the body of the paper above. It does not furnish any new resources needed for turning simulation into full-bore false belief explanation. (Of course I have argued that even the simplest versions of simulation run up against the epistemological problem of adjustment, but I will leave that aside for the moment.) However, we are presently concerned with the question of whether counterfactual reason explanation can account for the explanatory value of *any* degree of simulation. If it could, then perhaps there is something to simulation theory after all. Perhaps it could provide for some simple explanations from which theories could be built, and the remainder of folk psychological explanation could be account for by theory-theory.

But once we remember that counterfactual explanation is supposed to provide a simpler type of explanation than false belief explanation, we are faced with a new problem. That is because, when we look at the literature in developmental psychology, it appears that young children have systematic difficulties reasoning with counterfactual conditionals. Indeed the ability to deal with counterfactuals

emerges at about the same time as the ability to pass the false belief test (Riggs et al. 1998). This has led numerous psychologists to believe that the abilities are related, though there is much debate about the particular nature of the relation (see Mitchell and Riggs (2000)).

Therefore, even if counterfactual reason explanation is efficacious as a form of simulation, it is unlikely that it accomplishes this without whatever resources are needed for false belief explanation. Since we already have reason to think that genuine, radical simulation cannot account for false belief explanation, it is also therefore likely that counterfactual reason explanation is a form of radical simulation. More likely, it involves some kinds of doxastic presuppositions that put it on par with theory-theoretic explanation.