Empathy, Neural Imaging, and the Theory versus Simulation Debate[1]

Frederick R. Adams

[Mind and Language 16 (4):368-392 (2001)]

1. **Empathy Enters the Debate**

A debate is raging over how we attribute intentional states. We do it pretty much without notice. We realize that the waiter standing at our table is growing impatient for us to order. We notice the cab driver's amusement, as we fumble through change for a tip. We detect the person over there staring as if she thinks she recognizes us. None of this is hard to do (though it can be, when the intentional states ascribed seem somewhat irrational or alien as say, in the subjects of Milgram's experiment or autistic children or children prior to the age of four in the psychology lab). The debate is not about the fact *that* we do it. The debate is about *how* we do it. How do we so effortlessly ascribe beliefs and desires?

The Theory of Mind Theory (hereafter TT for 'theory-theory') says that we use a theory of mind to make attributions of intentional states to others (Churchland, 1998, pp. 3-15, Gopnik, 1995, Stich & Ravenscroft, 1994). The Simulation Theory (ST) says that we use our cognitive capacities to simulate and pretend to be in the situation of others. In thinking about the thought of another 'we take the subject matter of that thought, whether we believe the same or not, and think directly about it (Heal, 1995, p. 35).' We utilize the outputs of our cognitive mechanisms in order to make attributions of intentional states to others, but we don't employ *a theory* of mind, when doing so (Heal, 1995, Goldman, 1995, Gordon & Barker, 1994, Gordon, 1995, Barker, manuscript). On the surface these two theories appear different enough. On TT, one employs a theory—a theory of what it is to have a mind and how minds operate. On ST, one uses cognitive, intentional mechanisms of one's mind, but one does not employ a theory. Instead, one imaginatively identifies with the target or pretends to be the target. One then takes one's perceptual inputs of the situation of the target 'off-line' into a 'pretend module,' as it were, and cognitive mechanisms of

[1] An earlier version of this paper was presented at the Third European Congress of Analytic Philosophy in Maribor Slovenia in June of 1999. In addition John Barker, Gary Fuller, Bob Gordon and Thomas Pellathy, Joel Pust, the editors and three anonymous referees of this journal provided helpful comments on earlier versions of this paper.

Address for correspondence: Department of Philosophy, University of Delaware, Newark, DE, 19716.
Email: fa@udel.edu

one's own mind produce 'as-if' beliefs, desires, intentions and other intentional states. One then brings this output of the pretend box back on-line and attributes the outputs to the target. Supposedly, all of this is done without employing a theory of mind or folk-psychological platitudes (Lewis, 1966, 1970, 1972).

Worries creep into the debate over the crispness of the line that divides these theories (Heal, 1995, Stich & Nichols, 1995, Davies & Stone, 1996). Subjects who are simulating and pretending may be utilizing bits of a theory of mind--bits that are stored in the reaches of the mind that are not easily (White, 1987), nor reliably (Kornblith, 1998) accessible by consciousness. The folk-psychological theory may exist tacitly in the mind of the ascriber. If so, the boundaries between TT and ST begin to blur. With the boundaries of the two opposing theories threatening to blur, any sound attempt to provide evidence to tease apart TT and ST (or components of each) is welcome in the debate. The truth may exist in a hybrid account on which our ability to ascribe beliefs may involve a 'complicated interplay between theory-driven and simulation-driven moves' (Heal, 1995, p.50).

In this paper, I wish to consider an attempt to help distinguish TT from ST by looking at our ability to empathize. Ian Ravenscroft (Ravenscroft, 1998) thinks that consideration of our ability to empathize helps both to distinguish TT from ST and to support ST over TT. Ravenscroft recounts a case of watching a rock climber struggling with an overhang just prior to a fatal fall. From below observers watched in horror as the climber's strength began to ebb. He was unable to reverse the traverse he had made bringing him into his present position. He could not go forward to the good hand holds ahead that were just beyond his reach. The rope was positioned where it would offer no protection from a fall. Tiring and desperate, the climber sought a way out of his predicament. He searched every nook and cranny to find a new hold or a way to relieve the terrible strain on his already unbearably tired arms and fingers. Ravenscroft recalls: 'Looking on, I vividly experienced what it was like to be him, and not only because, as a climber, I had been in similar situations myself: any non-climber looking on could also experience what it was like to be that poor soul (Ravenscroft, 1998, p. 171).'

Let us grant that we do know what it was like to be the climber. Let's also grant that we have the ability to empathize and that this ability or skill was used by observers (and perhaps by readers of the above passage). How does this help separate TT from ST? It does if, in knowing what it was like to be the climber, we exercise simulation skills or abilities that give us empathic knowledge--knowledge that could only be gleaned from an implausible or uneconomical theory of mind. Were that the case, then there would be good reasons to believe that those who know what it was like to be the climber exercised these simulation skills, but did not rely primarily on knowledge of a theory of mind. Thus, additionally, the ability to empathize would provide evidence in favor of the more plausible and simpler[2] ST, over TT, in the debate over how we attribute intentional states to others. Indeed, this is Ravenscroft's claim about our empathic knowledge of what it is like to be the climber.

There is more to Ravenscroft's argument in favor of ST than just his appeal to this example and to the skills of simulation and pretence. He also argues that any attempt on the part of TT to do justice to our ability to empathize with the climber resorts to unrealistic and implausible additions to TT. So I will offer ways to make sense of our skills and abilities to empathize which are neither implausible nor unrealistic additions to TT. In this paper I take up the following matters. First, I challenge the basis upon which Ravenscroft claims that ST can, but TT cannot plausibly account for our empathic abilities with respect to the ill-fated climber. I suggest that every theory, not just a theory of mind, must bridge a type of theory/experience gap to which Ravenscroft appeals. Therefore, Ravenscroft's appeal to such a gap to distinguish TT and ST and settle the debate may be a red herring. I will show a respect in which ST too must bridge this gap. Second, I suggest that, in principle, TT can give a plausible account of how to bridge the theory/experience gap. I suggest a plausible way in which TT can account for our empathic abilities. I show that my addition to TT is better able to handle empathy than the revision to TT that Ravenscroft considers (but rejects as implausible). Third, I suggest that still more needs to be done to pull apart TT and ST as theories in this debate. Appeal to a single phenomenon of empathy will not be sufficient to differentiate TT and ST, nor be sufficient to determine a winner in the debate. At the very least it will take a syndrome of features that one or other side cannot handle and that the other side can, in order to determine a victor. I close by considering possible evidence from neuroscience that may help advance the discussion in the debate--evidence that may help shape the future empirical dimensions of the debate.

[2] Ravenscroft also argues that ST is simpler than TT. However, it is not certain that he can make this stick. For, as Stich and Nichols (1995) point out, ST makes use of cognitive mechanisms for taking processing

2. **Basis for Difference: Experience versus Theory**

How does Ravenscroft attempt to insert a wedge between TT and ST, and ultimately argue for the failure of TT? Specifically, how does Ravenscroft argue that TT lacks the capacity to account for our ability to empathize with the climber? He appeals to ideas stemming from the work of Nagel (Nagel, 1974) and Jackson (Jackson, 1982, 1986) in the literature on qualia. Specifically, he appeals to the principle that '…no amount of theorising about a possible experience E will yield an experience as of E' (Ravenscroft, 1998, p. 171). Readers familiar with this debate will recall Jackson's claim that Mary (a neuroscientist) may know all the facts about color, color vision, and neurophysiology, yet if barred life-long from seeing colors, Mary won't know what it is like to see red. There is what we might call a 'theory/experience' gap.[3]

There is much going on in the arguments of Nagel and Jackson (Churchland, 1998), but for my purposes we may ignore all but the following apparent kernel of truth (and, thereby, we may circumvent their controversial attack on physicalism). I intend to grant that knowledge by description of property R is not the same as knowledge by acquaintance with R--is not experiencing R (having one's nose rubbed in R, as it were). Knowing the wavelength and emission and reflectance properties of things that are red cannot take the place of having one's eyes trained on a red object under white light under good conditions of visibility and acuity. Knowing what it is like to see red requires getting oneself into a situation such that an object looks red to the person.[4] Without this, one may know everything there is to know about red *except* what it looks like (what it is like to experience something red). If this is the 'theory/experience' gap, I fully acknowledge it, but point out that ST must solve it as well. Whether Mary is simulating or theorizing about a target, she must know what it is like to experience mental states similar to those of the target in order to empathize with the target.[5]

off-line and bringing it back on line. This and other complications that do not exist on TT may need to be factored into one's metric for simplicity.

[3] Similarly, Nagel (Nagel, 1974) claims that we might know all the descriptive facts about what goes on in a bat's brain, but we still would not know what it is like to be a bat (we'd lack the bat's qualia). There are many variations on the theme—super smart Martains may know all about our brains but not know what it is like to be us.

[4] In a non-inferential sense of 'looks' (perhaps in the non-epistemic sense of Dretske, 1969).

[5] I say 'similar' because even though one has no antecedent knowledge of what it is like to experience a particular emotion, perhaps a description of the situation would call forth something similar. Producing the description may allow one to rely upon imagining being in the situation and may generate a similar

In the climber example, the theory/experience gap is supposed to work in a similar way. An observer may know all there is to know about a theory of mind (know the complete psychology or neurophysiology of the climber), but none of this knowledge of descriptive facts and theory translates into knowledge (from the inside) of the experiences going through the mind of the climber (the qualia). So, if observers from the ground are equipped only with a theory of the climber's mind (TT), they would not be able to have empathic knowledge of the experiential states of the climber—not be able to know what it is like to be him[6]. But we do know what it is like to be him. So our empathic knowledge is not coming from a theory of mind. It is coming from empathic cognitive simulation mechanisms.

Notice that what TT misses, according to Ravenscroft, are the qualia of the climber's experience--just the sort of thing missed by Mary. Mary knows that Fred is having an experience of red caused by a red object, say. Mary just doesn't know what that's like. Similarly, armed only with theory, Al may know the climber is having a bad time of it in his cognitive economy with respect to rock climbing. Al just may not know what that is like. Here is how Ravenscroft describes the deficit of TT (in contrast with our empathic knowledge):

> Seeing the climber's situation and noting his behaviour, we reason (via folk psychology [i.e. TT]) to claims about his mental states: he is afraid; he wishes he had not embarked upon such a dangerous climb; he believes that if he can just gain the next good hand hold he will be saved; and so on. But now the theory-experience gap is very apparent. When we empathise with the distressed climber we do not merely hold a series of propositions about his mental life. We personally experience states very much like his. Our situation is not at all like that of a scientist who, in spite of possessing a true and complete theory of echolocation, does not have an experience as of echolocation. We experience what it is like to be the distressed climber (Ravencroft, 1998, p. 172).

Ravenscroft is claiming that TT will give only a list of propositions about the climber's mental states. It will not give the climber's mental states from the first person perspective. It will not give the qualitative character of his experiences. And it won't give us similar qualitative states in the here and now. TT won't tell us how the climber feels—won't tell us what it is like in his shoes. And it won't allow us to

response in one providing some understanding of what it would be like to have that emotion. Thanks to an anonymous referee for this possibility.

[6] The deficit is not that we do not know all of the mental states of the climber. To know all of what it is like to be him would require this, of course. But that is beyond what is at issue. The question is only whether we know the qualitative character of his experiences associated with this specific climb in these specific (types) of circumstances. This seems to require going beyond the third-personal nature of a theory of mind. Below I will suggest a way to incorporate a more first-personal understanding of some sentences of a theory of mind—what I will call 'Kaplanesque-sentences.'

share similar experiences as we observe the climber undergoing his experiences. But we do seem to know this. So we must be acquiring this knowledge in some other way—viz. via cognitive simulation mechanisms. Were Ravenscroft correct that TT has this deficit, it would give ST the upper hand over TT in accounting for our empathic abilities. But is Ravenscroft correct?

3. **Bridging the Theory/Experience Gap**

Ravenscroft thinks that TT cannot explain our empathic knowledge because TT cannot bridge the theory/experience gap. Before we consider whether TT can be revised so that it can explain our empathic knowledge, let's review how ST is supposed to explain our empathic knowledge. How does ST bridge the gap, so to speak[7], according to Ravenscroft?

As with all versions of ST, the observer (let's stick with Al) is 'imaginatively identifying' with the climber or 'pretending to be the climber'. Al introduces the climber's perspective (perceptual and other inputs) into Al's own cognitive economy. Al then lets his cognitive machinery ('off-line') process the information that is being fed through the climber's cognitive system. Since Al shares sufficient similarity with the climber[8], Al waits for his cognitive mechanisms to produce beliefs, feelings, and other qualitative states. Then Al takes the outputs of his imagination (pretend) box back 'on-line' to attribute cognitive states to the climber.

It is important that a portion of this process be 'off-line' because Al's inputs do not have their typical outputs:

> …if, as a result of imaginative identification with the climber, I form the desire to cry for help, I may not actually cry out: my usual outputs. Nevertheless, simulation theorists hypothesise that the desire to cry out thus engendered is available to the mental state attribution mechanism: I attribute that desire to the climber (Ravenscroft, 1998, p. 178).

[7] ST won't literally 'bridge the theory/experience gap' because, if true, Al consults no theory when simulating the climber. Still, Al does have to get from his inputs to his pretence box to the outputs of qualia constitutive of what it is like to be the climber. Furthermore, these qualia have to be linked up with concepts of the mental states that are to be attributed to the target.

[8] What counts as sufficient similarity is vague. Can we simulate our pets and thereby ascribe beliefs and desires to them? Can they simulate us and thereby ascribe beliefs and desires to us? How similar must we be? What happens on the fringes where we share very little in common with our 'target' of simulation? These are important questions that have so far been inadequately addressed in this debate. Otherwise ST could be vacuously true (no two people are similar enough, but if they were, ST would enable us to attribute cognitive states). Or it could be empirically false (we do attribute cognitive states, say to Martians, Vulcans, and other interesting extraterrestrials or to our pets, but we are not similar enough to them in all the ways that count for our attributions to be the products of our 'similar' mental mechanisms).

It is also important that, when off-line, the inputs to Al's cognitive mechanisms can still cause, in Al, the qualia that they might cause on-line, were Al actually in the shoes of the climber (not just pretending to be in them).

> … the psychological process set in motion by the introduction of pretend perceptual and cognitive states can terminate in an experiential state such as fear. On this view, empathy involves a process of re-enactment. When I empathise with the distressed climber I re-enact a fragment of his mental life, and so come to occupy psychological states very close to his (Ravenscroft, 1998, p. 178).

Notice that there are a variety of ways in which off-line mechanisms may produce the relevant experiential states in Al (qualia of the very type that are on-line in the climber). The fear that Al detects (in simulating the climber) may be a result of pretending to be the climber and imaginatively identifying with the climber. We still need to know how *that works*. It may involve Al's *perceiving* the predicament of the climber and *recognizing* the similarity of the climber's purely physical situation to a type of situation that Al has previously been in. This may call forth memories and extrapolation mechanisms in Al and produce the relevant qualia. This may call forth the memories and their related qualitative nature and call them forth in the here and now because Al is perceiving the situation of the climber in the here and now (Nichols, et.al., 1996, 59-67). For Al to simulate the climber and thereby cognize the qualia of the type the climber is experiencing, Al's cognitive mechanisms may have to match the climber's *situation.* The climber is hanging there in peril, distraught, in despair, and exhausted, and so on. Al must experience the appropriate *qualia* for that situation (qualia of peril, of being distraught, of being in despair, and of being exhausted, and so on). That Al's cognitive mechanisms do this on-line or off-line, is no small achievement. How Al's simulational mechanisms do this is what is at issue. Surely it will require memory.[9] Surely it will require perceptually recognizing the climber's current situation. Surely it will require matching the climber's current situation with qualia relevant to those situations. And this is the bare minimum that it will require. Surely it will require a lot more (but this will be enough for now). For ST to explain how Al bridges the relevant gap, Al's cognitive system will have to make use of at least these features (and probably a lot more)[10].

[9] Even to hold an imagining or a pretending in mind long enough to elicit a qualitative feel will require holding the pretense or imagining in short term memory storage.

[10] Both TT and ST may also require some 'metacognitive' knowledge of the climber's mental states to be able to empathize with him. We must 'profile' him as a typical climber who prefers to live (not as a climber on a suicide mission with a wish to die or a masochist with a desire for a type of pain associated with experiences had just prior to death). For unless we had this knowledge profile of the climber, we

How might TT attempt to explain our empathic abilities? To do so, TT must bridge the theory/experience gap. Ravenscroft suggests a possible fix for TT, a fix which he claims '…is the only avenue open to the theory-theorist (Ravenscroft, 1998, 173).' Yet it is a fix that he ultimately rejects as being 'extraordinarily far-fetched.' The suggestion is that when we use folk psychology (i.e., a theory) to make attributions of mental states, each mental state attribution brings about in us the sort of experiences which someone in the attributed state would actually have. I won't attempt to spell out too carefully how this might go because I agree that it is extraordinary and far-fetched. That is, I agree that it is far-fetched that a *mechanism of making attributions would cause in us the qualia* (or other mental states) of the type attributed.[11] I do not agree that it is the only avenue available to TT, however.

There may be some plausibility in thinking that if we see someone cut her finger, we might wince at the attribution of the pain she must feel from the cut. There is a definite queasiness that accompanies the sight of such an event, at least in some of us. But the idea that all mental state attributions tap into some folk psychological cognitive mechanisms that then generate qualia in us does seem a bit extreme.[12] It

would hardly have empathic knowledge of his mental states, if he were taking joy in his predicament and looking forward to the adrenalin rush prior to his impending death. Thus, I cannot resist pointing out that even on ST, in order to get the empathic knowledge right, we have to be correct in our metacognitive profile of the character type of the climber. For example, to use TT to have empathic knowledge of a suicide climber one might need to consult (1) 'People who are engaged in committing suicide before a chosen audience are typically experiencing elation rather than depression' and (2) 'Jones is engaged in committing suicide before a chosen audience.' Consulting this might help Al know what it is like to be Jones the suicide climber. Even if Al were to simulate Jones, on ST, he may need to consult (2) and then take the rest off-line, as for the normal climber. So even ST has to utilize these 'causal pathways' from the metacognitive profile attributed to the climber to the to the empathic knowledge of what it is like to be the climber. The issue of causal pathways becomes central in what follows.

[11] While I think the mechanism Ravenscroft suggests (and rejects) is far-fetched for most of us, it may not be so far-fetched for others of us. John Barker has suggested that for some socially functioning adults with autism, the far-fetched may be the way they possess empathic knowledge. The usual causal pathways of empathic knowledge may not be available to them. (I will suggest that the usual pathways may go from perception of a situation that is ripe for causing a set of qualia to folk-psychological theorizing, to attribution to others of states with those qualia.) Socially functioning adults with autism may have to learn to exploit a connection from a theory of mind: the climber is in a terrorizing situation, one in a terrorizing situation normally feels terror, terror typically feels something like this ___, so the climber must feel something like this ___ too. Here, the causal pathway to the qualia goes via the folk-psychological theory (not the other way around). Simply seeing the climber in distress may not be enough to cause the appropriate fear states in the adult with autism.

[12] Although in a paper on the nature of concepts, Barsalou (Barsalou, 1999) argues that all concepts are modal…tied to specific modalities. Indeed, Barsalou maintains that all concepts are multi-modal criss-crossing modalities, proprioception, and introspection. If true, there may be *constitutive* ties to perceptual experiences in the attribution of any property to anything, not only attributions of mental states to the climber. We are not prepared to accept this theory of concepts. However, if we were, there would be a very direct reply to Ravenscroft, viz. any attribution of any state to the climber activates states of a multitude of perceptual mechanisms. Perceptual mechanisms are responsible for qualia (what it is like to

might entail that if Al thinks you have a headache, Al tends to have a headache as a result. This is pretty strange, to say the least[13]. And besides such a claim could not possibly be true in general. Remember that Mary, the color-quale-less scientist, *can* attribute to Fred the experience of seeing red. After all, Mary can know all of the theoretical facts about what is going on in Fred's head. But Mary need experience no red sensations at all, when attributing these cognitive states to Fred. Mary has no stored memories of what red things look like (for she has never experienced red). Thus, there just *is no mechanism to produce in her sensations as of red, when she attributes such sensations to Fred.* So there is no reason to think there has to be such a mechanism in order for Mary to make attributions of red qualia to Fred.[14] If it is not already obvious, I should point out that this is true whether Mary is consulting a theory of Fred or simulating Fred. In imaginatively identifying with Fred, Mary may know that Fred is seeing a red object. Mary may then pretend that she too is seeing a red object. She still will not know what red qualia are, though she would be able to know Fred is having them. This helps indicate why the theory/experience gap problem should be neutral with respect to ST or TT. Even while simulating Fred, Mary might lack the relevant connections between attributions of 'seeing red' and red qualia. Mary might simulate Fred as seeing a fire engine and know that fire engines are red and still not know what it's like for Fred to see the red fire engine.

Similarly, there is good reason to think there must be connections between mental state concepts and affective states in Al in order for Al to be able to make cognitive attributions to the climber. Returning to Ravenscroft's suggestion of how TT might bridge the gap, we can tell that it is not a very attractive attempt to rescue TT from the problem of explaining our empathic knowledge. I won't bother recounting his tweaking of the 'fix' that he considers, for it is clear that Ravenscroft never takes it seriously himself.

However, I do think there is a possibility that Ravenscroft does not consider—a possibility that would explain our empathic knowledge, if TT were correct. Think again of Mary. What does she lack?

be anyone). So, there *is no theory/experience gap* because concepts are housed in the very mechanisms that produce qualia. The 'belief box is located in the 'feel closet,' as it were.

[13] This would make processing pretty top down from high level cognition to low level sensing. There are good reasons to think that the brain is more modular and less penetrable than that between modules going top down (Fodor, 1983).

[14] Of course, Mary's concept of what Fred is experiencing will be impoverished in respect of 'what it is like' to experience red first hand, but this does not mean she cannot attribute to Fred her theoretical understanding of the mental state. She has the concept of experiencing something red. She does not have the concept of experiencing 'that color' where the 'that' refers to a subjectively accessed red. However, she could fill in the 'that' with a description of the emission and reflectance properties of red objects. So Mary does have a concept of 'experiencing red.'

She lacks knowledge by acquaintance of red things. Freed from her colorless environment and allowed to experience red things, she would not lack this knowledge. But prior to that time she lacks a type of knowledge that we might express thusly: ‘So, that ___ (referring to an experience of a red object *as red*) is what red looks like.’ Mary lacks knowledge of a set of truths that relate observation terms of a theory (terms such as ‘red’) to actual experiences designated by Kaplanesque ‘d-that’ demonstratives (Kaplan, 1968). These might be called ‘verification sentences,’ ‘protocol sentences,’ or whatever term is in fashion for sentences that relate a theory to the way an observation should look when one tries to confirm an observation sentence of a theory.[15] Such sentences will need a slot for the experience of the world itself. ‘Red looks like this ____(place nose in front of red thing under white light now please).’ ‘Roses smell like this___(lift fresh rose to nose now and sniff).’ ‘Precipitates look like this ___ (mix sodium nitrate and silver chloride in a solvent and have a look at the silver nitrate).’ You get the picture. Some sentences will be required to link ANY theory to reality, in an attempt to verify the theory, not just with respect to a theory of mind or folk psychology. In order for someone to use the theory to recognize things and make attributions, some of the sentences of the theory that will need to be consulted will include these: sentences of the form ‘F’s look like this___(the qualia of a perception of an F goes here).’[16]

At this point, I repeat that the theory/experience gap issue may be a red herring in the ST/TT debate. Suppose that we are wondering whether Mary empathizes with Fred according to TT or ST. Also, suppose that Mary is affect-impoverished (instead of color-experience-impoverished). She lacks the sorts of qualia associated with emotions. Now whether Mary consults sentences of a theory or tries to put herself in Fred’s shoes by imaginatively identifying with Fred, Mary will still not know what it is like to be Fred afraid or Fred angry. She will not because she will lack the ‘d-that thoughts’ that link qualitative states of fear or anger to her knowledge of mental attributions. The mere fact that Mary simulates an angry Fred or

[15] Observation sentences are neutral with respect to observers. So if a theory says the litmus paper will turn red when dipped into the acid, this doesn’t tell one how red looks, nor how it may look to us versus to some other cognitive creatures. Still, to verify that the litmus paper turned red, someone must know what red looks like. So theories have to provide for sentences which record the link between the observation sentence and the attempt to verify the observation. If I am correct, theories will also make room for sentences which bridge the gap between the theory and the qualitative experiences one would have under conditions appropriate for verifying those observations. Naturally, by the way, this does not commit me to a false verificationism about meaning. For I am NOT claiming that the meaning of the observation terms, like red, are exhausted by the verification conditions (of something’s looking red). Indeed, I would insist that ‘red’ means *red* not *looks red*.

a fearful Fred will not give her empathic knowledge all by itself. Mary also needs a way to connect her mental concepts of anger or fear with qualia, and that needs to be done *on ST or TT*. And Mary needs a way to get all of this into the here and now.

Now back to Ravenscroft's climber. If Al employs a theory of mind in order to possess empathic knowledge of what it is like to be the climber, Al will have to consult observation sentences of the above form. Al would note that the climber is straining to hold on. To have empathic knowledge of this aspect of the climber's experience, Al would also have to access truths such as 'straining to maintain one's grip feels like this____(insert experience of what it feels like from suitably relevant experience from Al's past).' Al would also note that the climber is in a situation likely to produce exhaustion. Thus, to know what that is like for the climber, Al would have to consult a sentence of TT of the form 'Extreme physical exhaustion feels similar to this ____(Al consults stored episodic memories from Al's past about extreme exertion).' Not to put too fine a point on it, Al will at some time realize the despair the climber must be feeling. Thus, to know what this is like Al will consult an episode of high anxiety or despair from his past thinking 'Despair feels something like that ___(consulting his episodic memory).' Furthermore, Al's consulting these sentences will be done in the here and now because Al's theorising is triggered by perceiving the predicament of the climber in the here and now. These perceptions of the situation of the climber in the here and now will trigger and bring forth memories of the relevant qualia into the here and now (Nichols, et.al., 1996).

Now surely Al won't have exactly the same past as the climber (not being a climber himself). Al would have to possess cognitive mechanisms that can extrapolate and generate experiences of qualia in Al right now from Al's innate 'similarity spaces' (not unlike 'quality spaces' Quine, 1969, p. 123f). Al will know what it is like to be the climber not because he has exactly the feelings the climber has, but because Al has enough experiences with enough similarity to those of the climber that Al and the climber share the same quality spaces. They share a significant number of psychological capacities and mechanisms that they are in the same empathic ballpark. There will be some similarity in qualia when they both entertain 'Despair feels like this____(enter an episode from your past, please).' And so on.

[16] The theory implies that there are Fs. Fs look like this ____. Go test the theory. See if anything looks like Fs are supposed to.

My view of how Al bridges the theory/experience gap is different from the ST account of Ravenscroft. My account is closer to an information-based account of empathy (Nichols, et.al., 1996), which also differs from the account of Ravenscroft. It is also different from the TT account suggested (and rejected) by Ravenscroft. On Ravenscroft's modification to TT, there were causal pathways going from the folk-psychological mechanism of ATTRIBUTION of cognitive states to the experienced QUALIA of the attributor. On his suggested view, *the attribution actually causes the qualitative state*. On my view, the causal order of mechanisms involved is not the same as on ST, nor is there the duplication of causal pathways suggested by Ravenscroft's implausible version of TT. On my account, Al must use perceptual inputs that he sees the climber receiving. Al must use his recognitional ability to access stored memories about his own qualia experienced in physical situations similar to that of the climber. Finally, Al can reconstruct and attribute states to the climber. The climber is in a situation that generally causes despair. When in despair, people tend to feel the qualia associated with despair. Al knows what despair is and can access, from memory, recollection of how it feels. Despair feels like this ____(Al plugs in episodic memory of the qualia of despair). Al is then able to attribute despair to the climber. The climber feels something like this___ (only to a more intense degree). This cognitive activity in Al in the here and now would call forth memories of similar qualia of despair in Al in the here and now, and therefore, Al would infer that the climber probably feels despair.[17]

This is clearly a TT account rather than an ST one. For example, on my account, none of Al's reasoning need be taken off-line. On ST the simulating usually is presumed to take place off-line. Also, on ST, memories of qualitative states are not accessed in the same way. On my version of TT, Al's access to

[17] My account is similar to the 'information-based' account by Nichols, et. al., p. 61. My account is also similar to that of empathy by Sober & Wilson (1998), p. 234. They suggest that 'S empathizes with O's emotion E if and only if O feels E, S believes that O feels E, and this causes S to feel E for O.' They do not maintain that S's feeling O is caused solely by S's belief that O feels E. It is caused in part by O's feeling E. So there must be some perceptual mechanisms via which S perceives O's situation. These perceptual mechanisms and memories from the past are available to produce E in S and support S's feeling E 'for O.' Nor do they maintain that what S feels for O is empathy without the belief that O feels E. The mere matching of emotions by O and S would be what they call 'personal distress' not empathy. (Their example is that if one newborn may be in distress and cause crying and distress in another newborn. The second experiences E, but does not experience empathy with the first, since it lacks belief about the first and projection of E 'for' the first.) Hence, I classify their theory with TT not ST, because they claim an empathizer must be a 'psychologist' to experience empathy. Sober & Wilson's distinction of empathy vs. personal distress also counts against Ravenscroft's appeal to just such an example of crying newborns as 'capable of empathetic responses, but …not capable of attributing mental states….' (Ravenscroft, p. 181).

qualitative states from memory together with his recognition of the type of situation the climber is in jointly cause Al to entertain attributions to the climber. The attributions to the climber do not cause qualitative states in Al. They are caused by perceptions of the climber's situation and memories from the past in the mind of the observer.

It is true that Ravenscroft does consider that a TT account 'may have to be augmented by a theory of memory (173),' but unlike Ravenscroft's implausible TT account, my account does not have Al's attribution to the climber cause the qualia in Al via accessed memories of qualia. For the attributions still cause the qualia on his proposal. Rather, on my account Al is caused, by his perception of the climber's situation to recognize that such situations normally produces qualia of despair and to recognize how despair feels via memory. In addition, there are no implausible or duplicated causal pathways in this account. On my account, qualia are linked with mental state terms, but not by being caused via attributions from the folk psychological mechanisms. Qualia are caused through sensory experiences. Memories of qualia types are stored and can be retrieved by perceiving situations that normally produce them or by remembering events where they were produced via perception or by thinking about them. Al simply accesses these qualia that are produced in the normal ways and are stored in memory. The qualia are not caused *by attribution*. They are caused by memories of qualitatively similar experiences and the memories are triggered by seeing (or hearing about) the situation of the climber. The qualia can be an aid to attribution when they are accessed by thinking about the types of situations that cause them in self or others. There is nothing implausible or uneconomical in this. If Al is asked to think about what it is like to be afraid, he can easily access thoughts and memories of what it was like to be afraid. He need not be attributing fear to himself or others. He can use this knowledge of what it is like to be afraid to aid him in making attributions to the climber and in empathizing with the climber. However, this does not mean that his attribution to the climber is implausibly duplicating causal pathways to qualitative states.

Also, there is every reason to think that intense memories of some experiences can invoke experiences themselves in the here and now. A vivid account of a traumatic experience might send shivers down one's spine. One may also re-live the original emotional experiences to a fainter, lesser extent.

There is nothing implausible about this[18]. So while Ravenscroft said that TT could introduce memory into an account of knowing what it is like to be the climber only with implausibility, there is nothing implausible about my account of how memory is accessed or how TT connects memories of qualia with concepts of mental states.

I surmise that Al would be able to run a similar set of inferences for his knowledge of what it is like to be in the climber's state of fear, exhaustion, angst, and so forth (for the various qualities of mind that Al can know, with respect to the climber). In all cases, Al uses his abilities of perceptual recognition, folk psychology, and memory of the Kaplanesque 'd-that'-sentences to reconstruct what it is like for the climber. The causal pathways in our account go from perceptual inputs, to memories, to folk-psychology, to 'd-that'-clauses linking with qualia from memories, and then finally to attributions to the climber. So my account does not require that Al's making attributions to the climber cause qualia in Al (qualia of what it is like for the climber). On my account, Al already has that link to the qualia by the time he makes the mental attribution to the climber.

This new account also helps to understand why it aids Al's abilities to know what it is like to be the climber to have had acquaintance with similar situations. Ravenscroft says:

> …pretend inputs which are close to our own experience will be easier to generate than those quite outside our experience. Thus we find it easier to empathise with those in situations of which we have first hand knowledge. Moreover, directly observing another's situation may facilitate imaginative identification more than hearing or reading about it. The simulation model thus explains the tendency to experience empathetic states more vividly when we actually witness—rather than, say, read about—another's plight (Ravenscroft, 1998, p. 179).

This proximity to our own first hand experiences is one of the hallmarks of my addition to TT. It is because Al can then store the relevant types of 'd-that'-sentences about the relevant qualia that Al is better able to empathize. Since there will be an extrapolation mechanism that one can consult, the exact same type of qualia will not be necessary to make such attributions. The more similar the experiences of Al and

[18] Lane, et.al., 1997, suggest that 'the evaluation of the emotional significance of a novel stimulus necessarily involves reference to previous experience' (3972). My Kaplanesques sentences may be instantiated here. They also say that anterior cingulate cortex and prefrontal cortex play a specific role in representing subjective emotional responses. In their studies, perceptual inputs from pictures and selective attention to features of the pictures caused emotional responses within subjects. There was nothing unusual or implausible about the causal pathways involved. These pathways (or ones like them) through cingulate and prefrontal cortex could be used by the brain to represent emotion on a TT account such as mine.

the target, the better Al's ability to identify the qualia of the target. It is easier to empathize[19] when one has first hand experience because that is just the type of acquaintance knowledge stored in the 'd-that'-sentences.

My account also fits nicely with some of the developmental evidence that is normally claimed to support ST. It is widely known that in their first year children learn to track their gaze with that of an adult. If an adult looks at an object, the infant will look at the same object. When an adult looks at an object and simultaneously expresses an emotion toward it, the infant will tend to adopt a similar response to the same tracked object. Also, by age two, children will re-direct the gaze of an adult seemingly to get the adult to have an experience or perception which may produce in the adult an emotion similar to one of the child (Harris, 1995, pp. 212-215). While I do not deny that this involves mimicry and other mechanisms generally consistent with ST, I happily point out that, by tracking perceptual inputs of the adult, on the assumption of shared perceptual apparatus, it is likely that the child (and adult) will be storing similar 'd-that'-sentences during this process. Those will be very useful for the child to have empathic knowledge of what it is like for the adult. They will also be useful for the child to try to redirect the gaze of the adult so the adult will know what it is like for the child.[20]

4. **Looking to Neural Imaging to Advance the Debate.**

Ravenscroft went for the slam-dunk, but the debate between ST and TT cannot be won with a single shot. The ability (alone) to have empathic knowledge of what it is like to be the climber cannot decide the debate. Empathic knowledge can come from more than one pathway or mechanism. ST can supply it. And, if I am right, TT can supply it (without resorting to highly implausible causal pathways of the type Ravenscroft considers and rejects).

[19] If Ravenscroft is correct that it is easier to empathize when one has shared experiences of the type had by the target, then surely TT can explain this ability to empathize in the way I describe. Barker (personal communication) suggests that ST does not require shared experiences. Little Mary may attend a play and for the first time in her life, by imaginatively identifying with the protagonist, Mary may experience contempt. Mary's cognitive mechanisms plus the inputs from the other actors in the play, may produce this experience in Mary (in her imaginative identification with the protagonist). Now we suspect that TT can explain this ability too, but may do so via the extrapolative mechanisms that we mention. However, if TT could not explain it or if the mechanisms used in theory consultation (TT) were distinct from those of ST, this would provide another way to attempt to pry apart the implications of the two theories.

[20] Compare Nichols, et. al. 1996, pp. 63-7, for a discussion of motor mimicry and emotional contagion and deliberate role playing. It is also sometimes suggested that this tracking of gaze is something that autistic children are not good at and that this may be partially explanatory of why autistic children may not have

In stead of looking for a winner in the debate, I suggest instead that we turn toward looking for neural evidence of simulation or of theorizing. As I have indicated above, I think that the explanation of how we make sense of one another will be a hybrid theory. It will involve some cognitive processes that include simulating or imagining and pretending, and it will also involve connecting those activities up with the understanding of what beliefs, desires, intentions and other psychological states are. So for this last section of the paper, I will turn my attention to how we might go about looking for evidence of the mind in ST mode or the mind in TT mode, as it were.

John Barker (Barker, manuscript), following the lead of Jane Heal (Heal, 1995), makes suggestions that I find intriguing. Barker claims that computational processing for minds implementing TT vs. minds implementing ST may give reasons to favor ST. His suggestion is that by looking only at how ProtoThinker (Barker, 1998) works, we can tell that there will be a genuine difference between what is actual and what is virtual in the minds of a cognizer adopting the intentional stance (Dennett, 1987). Barker constructs two models in his computer simulation of a thinker. In one model the proto-thinking agent PT, uses TT. In another model, PT uses ST. What is the difference? Barker shows how when PT is using ST, PT makes 'inferences' about beliefs of targets without appealing to folk-psychological principles. We know PT doesn't appeal to them because we know what sentences PT consults in the program (and what she doesn't consult). Nowhere in the information PT consults (in ST mode) are there any sentences of the form 'If someone believes that all bankers are rich, and believes that Smith is a banker, then that person will tend to believe that Smith is rich.' Indeed, there are no principles about beliefs or believers in the program at all. Rather, in ST mode, PT simulates believing the propositions that all bankers are rich and that Smith is a banker and PT infers that Smith is a banker. Then PT attributes this belief to Jones, by virtue of 'imaginatively identifying with Jones.' Basically, it is an unwritten and non-consulted assumption that PT is similar to Smith and that what PT believes[21], Smith will believe. So PT attributes the target belief to Smith. While Barker does not claim any genuine thinking is going on inside of PT, he does claim that we can tell (by considering how PT works) how TT and ST may come apart. Cognizers might

empathic knowledge (Gordon & Barker, 1994, Harris, 1995a, Coltheart & Langdon, 1998). See also Currie & Ravenscroft, 1996 for a discussion of imagery and simulation.

[21] Barker doesn't think PT *really* believes. But there is something playing the role of a proto-belief in PT.

work in the way PT does in ST mode[22]. If so, this would give us an understanding of a clear difference in TT and ST. This difference in real cognizers would be modeled by the difference in how PT works in ST mode versus TT mode. In TT mode, of course, PT does store *and deploy* a sentence about what people tend to believe (a principle like the one above about belief, bankers, and wealth).

We need to know if human cognizers are like PT in ST mode (or TT mode). I am encouraged by the fact that it is at least a research project to try to find out which we are. A relevant feature that might help us find out is new technology: CAT scans, PET scans, and MRI scans (Posner & Raichle, 1994, Churchland, 1995, Solso, 1997). We know many things about where the brain stores information from new imaging technology about what parts of the brain are active during certain cognitive tasks. We are learning amazing things such as that second languages (Perani, et.al., 1998, Kim, et.al., 1997)are stored in different parts of the brain than first languages, and names are stored in different locations than predicates[23].

It is well known that when subjects are asked to imagine walking along a familiar street while observing the landmarks en route, PET scans show that their parietal and temporal lobes 'light up' with activity that is consistent with their scanning mental images (Posner & Raichle, 1994, p. 95). Since the subjects are not actually walking, this processing is done 'off-line'[24]. Thus, we are now already able to detect changes in blood flow and processing activity in the brain that is consistent with off-line processing. The subjects are pretending to be walking along the route. They might as well be imaginatively identifying with someone else who is walking along the route (someone whom they are simulating to be walking along the route). So they are simulating what someone would be seeing and believing about the environment, along the way, so to say (Currie, 1995, Currie & Ravenscroft, 1997).

[22] Jane Heal makes the same kind of case for ST as Barker '…in thinking about another's thought what we do is take the subject matter of that thought, whether we believe the same or not, and think directly about it' (Heal, 1995, p. 35).

[23] There may be neuroscientific reasons why learning propositional logic is different from learning predicate logic.

[24] Nichols & Stich (manuscript) map out a theory of pretense in which they compare the benefits of thinking that pretense is done in a kind of possible worlds box that is kept separate from one's belief box (to speak metaphorically). Nichols and Stich explain why something like this helps to understand pretense. We are encouraged by the possibility that there may be a correlate in the type of processing given to representations that are syntactically like beliefs, but are not beliefs. There may well be different processing associated with these differences. In what follows, I shall try to link up the brain's attempt to keep track of its pretences with 'source monitoring' in false memory syndrome experiments.

As we've seen, it is a central feature of ST that one takes perceptual inputs off-line, when imaginatively identifying with the target one is simulating (the climber). This kind of pretence and off-line reasoning might well take place in a separate region of the brain than on-line reasoning. Since off-line reasoning is divorced from its usual on-line outputs, we might even expect this result. And, with a bit of neural imaging technology (brain snooping), we might well discover that in ST mode, as it were, different regions of human brains 'light up' than when in TT mode.

In search of such empirical evidence of different processing in the brain, I suggest that (among others) the debate look at neural imaging studies of false memories and false recognition (Loftus, 1997, Schacter, 1997). It is now clear how easily false memories can be induced in subjects. For example, Loftus (Lofuts, 1997) screened subjects[25] to be sure that they had never been lost in a shopping mall as a child. Then subjects were read accounts of events that were supposedly in their pasts. Three of the events recounted actually happened, the fourth (being lost at the mall, crying, being found and aided and ultimately reunited with family) did not. Subjects were instructed to read about the accounts and then write what they remembered about each of the events. If they did not remember, they were to record 'I do not remember this.' At two-week intervals after responding to the four events, subjects returned to be asked again about the events. After the second visit, twenty five percent of the subjects 'remembered' the false event fully or partially. Incredibly, false memories may be induced this easily.

Neuroscientists have begun studying the role of processing in the frontal lobes via neural imaging during false recognitions. Schacter and colleagues (Schacter, 1997) looked at both true and false recognition via PET Scans. They found evidence of temporal lobe activation during both true and false recognition, but found activity in the right anterior frontal region tended to be greater during false recognition. A follow-up study using ERP techniques ('event related potentials') found similar results to the PET results. To offer a possible explanation of the findings, Schacter surmises that the right frontal regions may play an important role in strategic monitoring of processes that are required to determine whether a piece of information was experienced previously (a true recognition) or is novel. The increased activity associated with a false recognition may represent increased search processing for stored information from the past. Of course, this does not explain why these monitoring processes fail to override a false recognition. The monitoring may be overridden if there is damage to the brain (as in some subjects tested). I suspect that when there is no damage, subjects are processing representations in ways that they normally reserved for representations that the brain stores as true recognitions. Then, for reasons still unknown, subjects lose track of which representations are of actual events and which are not. As with vivid hallucinations, false recognitions are stored representations that seem just like representations of actual events, from the point of view of the brain which contains them (and loses track of their etiology).

We will have to await new information from neuroscience to be able to draw any further conclusions about false recognitions or simulation. The reason I find this avenue to be promising is that I see a plausible connection between false recognition and simulation. On the theory of pretense of Nichols and Stich (Nichols & Stich, manuscript), the English sentence 'I was lost at a shopping mall when I was five' goes into the possible world box, if we are simulating one who was lost as a child. The same sentence goes into that work-space when one reads the false account in Loftus's experimental paradigm. Now somehow subjects who do not make false recognitions are able to keep track of when a sentence belongs in the possible world box only (the pretend box, if you will) and not in the belief box. Subjects who suffer false recognitions are not able to keep track of this. It is worth pursuing the question whether the results of experiments by Schacter and his colleagues are detecting the cognitive processing associated with this attempt at monitoring passage of representations from one's possible world box into one's belief box. Subjects unable to keep track or monitor origin of the sentence 'I was lost in a mall as a child' may be susceptible to false recognitions (Currie, 2000). Subjects who can keep track are not susceptible.

The connection with ST is that when we simulate a target, our pretend box will be populated with many false sentences (I was lost at a mall when I was five,' 'I am hanging here in exhaustion and despair on this ledge,' and so on.) Suppose there are parts of the brain that have the job of monitoring and tracking representations that should not go into one's belief box because they are false.[26] If there is an increase in

[25] By talking with relatives, it was possible to determine that subjects had not had the relevant type of experience so there was no actual memory of the experience for the subjects to recall.

[26] Currie (Currie, 2000) notes the very interesting work of Chris Frith. Frith (Frith 1987, 1992) points out that during perception images are produce in us independently of our will. However, in imagining, images are produced in us at will. Both Frith and Currie suggest that schizophrenic delusions about thought insertion or thought removal may be the result of the brain's inability to monitor which imagery the world caused and which imagery the subject's own brain caused at will. Frith suggests that a tracking signal not unlike efferent copy feedback may be lost in cases of schizophrenia. There may well be a similar loss of monitoring in the case of false belief.

activity in parts of the brain that do such monitoring, then it may actually be possible to detect the brain making these cognitive checks. If so, it may be possible to witness a difference in a brain that is in ST mode from one that is in TT mode. Now I admit that this is highly speculative and not likely to bear fruit in the immediate future. Still, the studies of false recognitions, neural imaging, and ST are converging. I maintain that if there is a difference between ST and TT, then there will be a detectable difference between a brain in ST mode and one in TT mode. Persons engaged in simulation will need to 'track' the sources of the stored representations that are pretenses. If Nichols and Stich are correct, subjects will need to track whether these sentences are in their 'possible world box' (so to say). This should involve more activity in the brain's prefrontal regions—activity of the sort that we find in subjects with false memories. Therefore, this is one more feature to look for in the debate (along with Barker's appeal to computational differences).

Another place to look is at visual vs. sentential processing. Theories (including folk psychology) are typically considered to be sets of laws and sentences recording those laws. So folk psychology is a set of platitudes about how beliefs, desires, intentions, and other mental states tend to interact to cause behavior. Given the right initial conditions and background conditions (sensory inputs, standing mental states and characteristics, etc.), folk psychology plus a set of stimulus conditions deductively imply (or inductively make probable), a set of sentences about what someone will believe, do, feel, say, etc. The point is that consulting a theory (TT) may involve language processing. ST, in contrast, may involve some language processing, but it may largely involve the processing of visual, tactile, or other forms of sensory processing of images. When Al simulates the climber, Al pictures himself in the situation of the climber. This may involve various forms of visual and other sensory processing. But it may involve little or no sentential processing and no consultation of theory of mind whatsoever. At least, this is one possible way in which TT and ST may be distinguishable for some cognitive tasks. So if Al tries to attribute mental states to the climber by consulting a theory, this may involve some form of cognitive consultation with the sentences of the theory of mind. While if Al is only simulating the climber, Al may devote most of his cognitive processing to imagining the sensory inputs that would be impinging upon the cognitive apparatus of the climber.

Why is this relevant? Again it is relevant because there is empirical evidence that for at least some cognitive tasks involving working memory, visual processing takes place in different regions of the brain than verbal processing. Jonides & Smith (Jonides & Smith, 1997) devised a paradigm in which subjects had to be able to remember and match the spatial positions of cards. The target card had to be matched to the spatial position of a previous card shown three cards back in a series (during the spatial memory condition). Or the subjects had to match the letter on the target card to the letter on a card shown three cards back (during the verbal memory condition). This forced the subjects to hold either spatial or verbal information about the series of cards and the target card in working memory. Their findings were that 'Although there was …evidence of bilateral activation in both tasks, in the spatial task there was more activation in posterior and anterior areas in the right than in the left hemisphere. Conversely, in the verbal task, the activations in the left hemisphere were higher than those in the right hemisphere' (Jonides & Smith, 1997, p. 253).

In Johnson & Raye's work on source monitoring (Johnson & Raye, forthcoming) there is further evidence that the frontal regions of the brain are responsible for monitoring origin of information. There is also evidence that cross modal processing and aging (Henkel, Johnson & DeLeonardis, Forthcoming) can each have effects both on functioning of these areas of the brain and on accuracy of source monitoring.

Maguire & Mummery (1999) found that context rich episodic memories (of the type that veridical memories would be) require processing provided by the hippocampus (p. 60). Though this is not a new result, their research focuses on complex, real world memory retrieval. Many parts of the brain are involved in memory, but their study confirms evidence of the role of the hippocampus in processing episodic memories. Their study is important for self-relevant memories. The medial frontal region of the brain had previously been implicated in personal memory and in other studies about attributions of mental states generally. Maguire and Mummery's results support an *episodic* view of personal memories that also implicate involvement of the hippocampus, saying 'we have found that the left hippocampus is differentially responsive to personally relevant memories with specific temporal contexts, as are the left temporal pole and medial frontal cortex' (p. 60).

If these findings comport to broader types of cognitive localization of processing symptoms, then it is possible, in principle, that we will discover a way to detect a brain in ST mode vs. one in TT mode[27].

[27] I am aware of an obvious protest from prototype theorist who might run a version of folk psychology (TT) in that form (say, Churchland, 1995) rather than in a language of thought form (say, Fodor, 1988). This would mean that we would be looking for different contrasts with neural imaging. We'd be looking

The difference would be detectable by the location of the cognitive processing. For example, based on the research of Johnson and Maguire & Mummery, it may turn out that one in TT mode will be doing more processing of actual worldly events (to which theories are being applied). Whereas, one in ST mode may be doing more processing of pretended events and simulated ones). These processing differences may implicate different regions of the brain.

I am not so naïve as to think that there is a TT module and/or an ST module and that all we have to do is use PET scans or some other method of neural imaging to detect where these modules are located. Rather, my claim is that *if* the kind of cognitive processing required for a brain in TT mode involves more actual worldly memories or more sentential processing than required of a brain in ST mode, *then* this difference in processing in principle will be detectable[28]. There will be some reflection of the differences involved due to the differences in type of cognitive processing. These differences will be detectable due to the differences of contribution by different portions of the brain to these overall cognitive functions, though there is no simply story about how differences of function align with neuro-anatomical differences. So the differences may not exactly be localized. A brain in TT mode may light up in as many as 6 to 10 locations. A brain in ST mode may light up in 6 to 10 locations as well. But there may be a different emphasis or weight on the regions they light up in common. And there may even be significant differences—some regions lit up when consulting a theory (TT) may not light up at all when imaginatively identifying with the situation of another (ST).[29]

for prototype processing typical of high-level conceptual processing where fine semantic discriminations are possible (rather than lower level processing where visualizations alone may be taking place), for example. Though we would be looking for different things associated with different types of prototype processing, I still firmly believe there would be differences in the brain when processing in TT mode vs. ST mode.

[28] Computer simulations of the kind that Barker describes may provide further support of this.

[29] In neural imaging studies of subjects who are asked to process stories where comprehension of the stories requires processing information about the mental states of those in the story, Fletcher et. al., 1995 found neural activation in the temporal poles bilaterally, the left superior temporal gyrus, the posterior cingulate cortex, and the left medial frontal gyrus. Strikingly, when subjects comprehended stories requiring no processing of information about mental states of protagonists, there was no activation of the left medial frontal gyrus causing them to conclude “our data pinpointed the medial dorsal region of the left frontal cortex as being critically involved in mentalising.” (p.121). There is further neural imaging support for frontal lobe involvement in modeling of other minds in Goel, et. al., 1995. This location is consistent with the findings of Johnson, et.al. Barker (Barker, manuscript and personal communication) has pointed out that he views ST and TT as using many of the same cognitive mechanisms. I suspect that Nichols and Stich would agree, given that on their theory of pretense, the pretense box has access to the contents of almost ‘the entire’ belief box, if necessary. Still, I think that having access to the contents of the belief box and actually using all of its contents are quite distinct. While TT and ST exploit many of the same

I suspect that taking the intentional stance requires a bit of both modes. If so, ST and TT would never be entirely pulled apart and there would be no overall victor. But by considerations such as Barker's, and by new evidence from neural imaging, we would at least be able to tell when the brain was in which processing mode. I suggest that this is much of what should be of interest in the ST/TT debate, anyway. Which of our abilities in taking the intentional stance are due to ST mode and which are due to TT mode. It is most unlikely that a single theory fits all intentional stance phenomena.

In closing, I will note that there is very exciting new research offered in support of ST. Gallese and Goldman (1998) have written about the discovery of mirror neurons in macaque monkeys. When one macaque is observing another perform a goal-directed activity, such as picking up and rotating objects, the mirror neurons in the observer fire exactly as (i.e., mirroring) those neurons causing the activities of the actor. (Of course, there is some kind of inhibition. The observer does not necessarily imitate the movements of the agent, while watching.) Gallese and Goldman argue that what is happening in the macaques constitutes an instance of brain simulation. It offers a mechanism via which the observer knows what it is like to be the actor—viz. they share a brain mechanism and neural firing patterns.

There is good reason to think there are mirror neurons in humans as well. And if these can be shown to fire in cases such as that of Ravenscroft's climber, then ST may have this neurological support for how we gain empathic knowledge from such brain mechanisms (Gallese & Goldman, 495). Still, I will offer two factors to consider that suggest problems for thinking that mirror neurons alone would account for empathic understanding of others. First, apparently mirror neurons only fire in response to goal-directed behavior on the part of the actor. Not just any bodily motion on the part of the actor produces the result in the observer's mirror neurons. This suggests the following problem for ST. If the observer has to interpret the activities of the actor *as goal-directed*, there may be an *intentional stance* overlay that is used to filter incoming perceptual information. This would again make it difficult to drive a wedge between TT and ST. Second, following Sober & Wilson's (1998)[30] account of empathy, empathy involves not only feeling the emotion of the target, but believing that the target has that emotion and projecting one's

mechanisms, they need not exploit them to the same degree. This is why I say that the same regions of the brain may 'light up' when in TT or ST mode, but *not* light up to the *same degree.* We clearly need more data but we do seem to be homing in on the locations in the brain through which the processing of information about mental states must pass.

[30] See footnote #17.

empathetic affect to the target. They call this feeling the emotion 'for' the target. The mere firing of motor neurons alone may produce the affect, but not necessarily the belief nor the projection of the emotion felt 'for' the target. Sober & Wilson contrast feeling empathy for someone like Ravenscroft's climber with 'personal distress.' In personal distress, one may feel the despair the climber feels but not feel it *for the climber*. So even if seeing the climber's situation, I feel despair and I feel this because my mirror neurons are firing (my brain is simulating his brain with respect to affective states), that alone does not mean that I am empathizing. I may be feeling a bit desperate myself. I may not be projecting my emotional and cognitive states onto the climber. I may not be feeling despair 'for the climber' (in the words of Sober & Wilson). To do this may require a theory of mind and how mental states connect with these affective states both in me and in others. Sober & Wilson claim that even to feel empathy requires being a psychologist (taking the intentional stance).

Again, we will have to await new research. However, I think these sorts of results show that if there is hope of finding empirically detectable differences between TT and ST, neural imaging and other neural studies point the way to future of the debate.

Department of Philosophy
University of Delaware

References

Barker, J. Manuscript: The Fate of Folk Psychology. [Computer Modeling and the Fate of Folk Psychology. In Moore, J. and Bynum, T. (eds) *Cyberphilosophy: The Intersection of Philosophy and Computing*. Blackwell 2003, 26-44.]

Barker, J. 1998: *ProtoThinker: A Model of the Mind*. Belmont, CA: Wadsworth.

Barsalou, L. 1999: Perceptual Symbol Systems. *Behavioral and Brain Sciences*, 22, 577-609.

Cahill, C. and Frith, C.D. 1996: False perceptions or false beliefs? Hallucinations and delusions in schizophrenia. In P. Halligan and J. Marshal (eds), *Methods in Madness*. Brighton: Lawrence Erlbaum Associates.

Carruthers, P. & Smith, P. (eds): 1996: *Theories of theories of mind*. Cambridge: Cambridge University Press.

Coltheart, M. & Langdon, R. 1998: Autism, Modularity and Levels of Explanation in Cognitive Science. *Mind and Language*, 13, 138-152.

Churchland, P. 1995: *The Engine of Reason, the Seat of the Soul*. Cambridge, MA.: MIT/Bradford.

Churchland, P. & Churchland, P. 1998: *On the Contrary*. Cambridge, MA.: MIT/Bradford.

Churchland, P.: 1998. "Knowing Qualia: A Reply to Jackson." In Churchland & Churchland (eds.), 1998, 143- 157.

Currie, G. 1995: Visual Imagery as the Simulation of Vision. *Mind and Language*, 10, 25-44.

Currie, G. 1998: Pretence, pretending and metarepresentation. *Mind and Language*, 13, 35-55.

Currie, G., 2000: Imagination, Delusion and Hallucination. *Mind & Language*, 15, 168-183.

Currie, G. and Ravenscroft, I. 1997: Mental Simulation and Motor Imagery, *Philosophy of Science,* 64, 161-180.

Davies, M. & Stone, T. 1995a: *Folk Psychology*. Oxford: Blackwell.

Davies, M. & Stone, T. 1995b: *Mental Simulation.* Oxford: Blackwell.

Davies, M. & Stone, T. 1996: The mental simulation debate: a progress report. In Carruthers & Smith (eds), *Theories of theories of mind.* Cambridge: Cambridge University Press, 119-137.

Dretske, F.1969: *Seeing and Knowing*. Chicago: University of Chicago Press.

Dennett, D.1987: *The Intentional Stance*. Cambridge, MA.: MIT/Bradford.

Fletcher, P., Happe, F., Frith, U., Baker, S., Dolan, R., Frackowiak, R. & Frith, C. 1995: Other minds in the brain: a functional imaging study of 'theory of mind' in story comprehension. *Cognition*, 57, 109-128.

Fodor, J.A.1983: *The Modularity of Mind*. Cambridge, MA: MIT/Bradford.

Fodor, J.A.1988: *Psychosemantics*. Cambridge, MA: MIT/Bradford.

Frith, C. 1987: The positive and negative symptoms of schizophrenia reflect impairments in the perception and imitation of action. *Psychological Medicine*, 17. 631-48.

Frith, C. 1992: *The Cognitive Neuropsychology of Schizophrenia*. Hove, E. Sussex: Lawrence Erlbaum Associates.

Goel, V., Grafman, J., Sadato, N., & Hallett, M. 1995: Modeling other minds. *NeuroReport*, 6, 1741-1746.

Harris, P. 1995: From Simulation to Folk Psychology: The case for Development. In Davies & Stone, 1995a, 207-231.

Gallese, V. & Goldman, A. 1998: Mirror Neurons and the Simulation Theory of Mind-Reading. *Trends in Cognitive Science*, 3, 493-501.

Goldman, A. 1995: Empathy, Mind, and Morals. In Davies & Stone, 1995b, 185-208.

Gopnik, A. & Wellman, H. 1995: Why the Child's Theory of Mind Really is a Theory. In Davies & Stone, 1995a, 232-258.

Gordon, R. 1995: Folk Psychology as Simulation. In Davies & Stone 1995a.

Gordon, R. & Barker, J. 1994: Autism and the Theory of Mind Debate. In G. Graham & G. Stephens (eds) *Philosophical Psychopathology*. Cambridge, MA.: MIT/Bradford, 163-181.

Heal, J.1995: How to think about thinking. In Davies & Stone, 1995b, 33-52.

Henkel, L., Johnson, M. & De Leonardis, D. Forthcoming: Aging and Source Monitoring: Cognitive Processes and Neurological Correlates. *Journal of Experimental Psychology*.

Jackson, F. 1982: Epiphenomenal Qualia. *Philosophical Quarterly*, 32, 127-136.

Jackson, F. 1986: What Mary Didn't Know. *Journal of Philosophy*, 291-295.

Johnson, M.K. & Raye, C.L. Forthcoming: Cognitive and Brain Mechanisms of False Memories and Beliefs. In D.L. Schacter & E. Scarry (Eds.) *Memory and Belief*. Cambridge,MA: Harvard University Press.

Jonides, J. & Smith, E. 1997: The Architecture of Working Memory. In Rugg, (1997), 243-276.

Kaplan, D.1979: Dthat. In French, et.al. (eds) *Contemporary Perspectives in the Philosophy of Language,* Minneapolis: University of Minnesota Press, 401-412.

Kim, K., et.al. 1997: Distinct cortical areas associated with native and second languages. *Nature*, 388, 171-174.

Kornblith, H. 1998: What is it like to be me? *Australasian Journal of Philosophy*, 76, 48-60.

Lane, R., Fink, G., Chau, P., & Dolan, R. 1997: Neural activation during selective attention to subjective emotional responses. *NeuroReport,* 8, 3969-3972.

Lewis, D. 1966: An Argument for the Identity Theory. *Journal of Philosophy*, 63, 17-25.

Lewis, D. 1970: How to define Theoretical Terms. *Journal of Philosophy*, 67, 427-446.

Lewis, D. 1972: Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy*, 50, 249-258.

Loftus, E. 1997: Memory for a Past That Never Was. *Current Directions in Psychological Science*, 6, 60-65.

Maguire, E. & Mummery, C. 1999: Differential Modulation of a Common Memory Retrieval Network Revealed by Positron Emission Tomography. *Hippocampus*, 9, 54-61.

Nagel, T. 1974: What is it like to be a bat? *Philosophical Review*, 83, 435-450.

Nichols, S. & Stich, S. Manuscript: A Cognitive Theory of Pretense.

Nichols, S., Stich, S., Leslie, A., & Klein, D. 1996: Varieties of off-line simulation. In Carruthers & Smith, 39-74.

Perani, D., et. al. 1998: The Bilingual Brain. *Brain*, 121, 1841-1851.

Posner, M. &Raichle, M. 1994: *Images of Mind*. New York: Feeman & Co.

Quine, W.V. 1969: *Ontological Relativity and other Essays*. New York: Columbia University Press.

Ravenscroft, I. 1998: What is it like to be someone else? Simulation and Empathy. *Ratio*, XI, 170-185.

Rugg, M. 1997: *Cognitive Neuroscience*. Cambridge, MA: MIT Press.

Schacter, D. 1997: False Recognition and the Brain. *Current Directions in Psychological Science*, 6, 65-70.

Sober, E. & Wilson, D. 1998: *Unto Others: The Evolution of Psychology of Unselfish Behavior.* Cambridge, MA: Harvard University Press.

Solso, R. 1997: *Mind and Brain Sciences in the 21st Century*. Cambridge, MA: MIT/Bradford.

Sorenson, R. Forthcoming: Self-Strengthening Empathy. *Philosophy and Phenomenological Research*.

Stich, S. & Nichols, S. 1995: Folk Pshychology: Simulation or Tacit Theory? In Davies & Stone, 1995a, 123-158.

Stich, S. & Nichols, S. 1997: Cognitive Penetrability, Rationality and Restricted Simulation. *Mind and Language*, 12, 297-326.

Stich, S. & Ravenscroft, I. 1994: What is Folk Psychology? *Cognition*, 50, 447-468.

White, S. 1987: What is it like to be an homunculus? *Pacific Philosophical Quarterly*, 68, 148-174.