Mind & Language, ISSN: 0268-1064
Vol. 12. Nos. 3/4 September/December 1997, pp 297–326.

Cognitive Penetrability, Rationality and Restricted Simulation

STEPHEN STICH AND SHAUN NICHOLS

Abstract: Heal (1996a) maintains that evidence of cognitive penetrability doesn't determine whether stimulation theory or theory theory wins. Given the wide variety of mechanisms and processes that get called 'simulation', we argue that it's not useful to ask 'who wins?'. The label 'simulation' picks out no natural or theoretically interesting category. We propose a more fine-grained taxonomy and argue that some processes that have been labelled 'simulation', e.g., 'actual-situation-simulation', clearly do exist, while other processes labelled 'simulation', e.g., 'pretence-driven-off-line-simulation' are quite controversial. We do concede that evidence of cognitive penetrability isn't decisive evidence against pretence-driven-off-line-simulation. Nonetheless, advocates of pretence-driven-off-line-simulation need to provide some explanation of the experimental evidence of penetrability. We argue that Heal's suggestion that simulation is restricted to 'rational' processes is unprincipled, and we offer an alternative proposal for restricted simulation.

1. *Introduction*

In a series of recent papers, Jane Heal (1994, 1995a, b, 1996a, b) has developed her own quite distinctive version of simulation theory and offered a detailed critique of the arguments against simulation theory that we and our collaborators presented in earlier papers. Heal's theory is clearly set out and carefully defended, and her critique of our arguments is constructive and well informed. Unlike a fair amount of what has been written in this area in recent years, her work is refreshingly free of obscurity; it generates more light than heat. While we have many disagreements with Heal, we also find much that we can agree with and learn from. In this paper we hope to advance the

Earlier versions of some of this material were presented to the Psychology Department at Northwestern University and the Humanities and Social Science Division at the California Institute of Technology. We are grateful for helpful comments from both of these audiences. Special thanks are due to Meredith Williams, Michael Williams and an anonymous referee for this Journal.

Address for correspondence: Stephen Stich, Department of Philosophy, Rutgers University, New Brunswick, NH 08903, USA.
Email: stich@ruccs.rutgers.edu, nichols@cofc.edu.

discussion by saying where we agree and how we think we can build on that agreement. We'll also explain where we disagree and why.

2. *Preliminaries*

In 'Simulation and Cognitive Penetrability' (1996a),[1] a paper intended primarily as a critique of our views, Heal writes:

> 'Theory theory' and 'simulation theory' are rival accounts of what is involved in possession of psychological concepts, in the ability to apply them to ourselves and others and the ability to use judgments invoking these concepts in arriving at predictions of the thoughts and actions of others. (p. 45)

She then quickly puts off to the side consideration of what is involved in the possession of psychological concepts and debates about applying psychological concepts to ourselves, and focuses in on differing accounts of 'our ability to predict others', because that is the domain in which the most important disagreements between her views and ours can be found. This strategy of dividing the issues is one we happily accept. In our earlier papers (Stich and Nichols, 1992, 1995; Nichols et al., 1995; Nichols et al., 1996) we have said very little about self attribution and nothing at all about what's involved in the possession of psychological concepts. We think questions about the mechanisms underlying self-attribution raise important and difficult issues, some of them empirical, others conceptual. Our silence on the matter has a simple explanation: We don't yet know what we want to say. Our reluctance to participate in the debate about what is involved in the possession of psychological concepts has a different explanation. We have no clear idea what the participants in this debate mean by concept mastery, and we rather doubt that they do either. So it is a debate that we find hard to take seriously.[2]

In setting out her account of 'the story so far', Heal offers the following brief characterization of the theory theory:

---

[1] All quotes are from this paper unless otherwise indicated.

[2] In his very useful overview of the debate over mental simulation, Martin Davies 1994 suggests that one crucial question in dispute is whether mental simulation can 'figure in a philosophically fundamental account of the nature of mental states, or the conditions for mastery of mental concepts' (p. 121). A bit earlier he asks whether simulation can be used in 'a constitutive account of the nature of mental states' (p. 120). For reasons that one of us has elaborated elsewhere (Stich, 1996, pp. 60–3) we are more than a bit skeptical about the notion of a 'constitutive account' of *anything*. We are equally skeptical about the notion of a 'philosophically fundamental account'. Indeed, we suspect that what counts as 'philosophically fundamental' tends to vary widely with the geographical location of those assessing the account.

> The core ideas of theory theory are that mastery of psychological concepts is grasp of a theory and that when I predict the thoughts and actions of another person using my competence with psychological concepts I do so by calling upon this theory, i.e. a body of information about the nature, causes and effects of psychological states. This theory may be explicitly known or (much more probably) in large part only tacitly known; it may consist of some structured set of nomological generalizations or just be a melange of low level rules of thumb; it may be internally represented in sentences or, alternatively, in pictures or rules. But whichever of these variants we go for, individual predictions, for example of the decision of another person, are arrived at by taking information about that person, integrating it with the information of the theory and deriving the prediction. (p. 45)

Unfortunately, as Heal goes on to observe, 'simulation theory is not quite so easy to summarize, since the differences between its versions range more widely' (p. 45). This is a claim with which we certainly would not quarrel. Indeed it is our contention that the diversity among the theories, processes and mechanisms to which advocates of simulation theory have attached the label 'simulation' is so great that the term itself has become quite useless. It picks out no natural or theoretically interesting category. Some of the processes to which the 'simulation' label has been attached patently do exist and do underlie some episodes in which people predict the thoughts or actions of others. The existence of many of the other processes to which the label has been applied, and of the psychological mechanisms on which they depend, is much more dubious. While a great deal more experimental work is needed, it is our view that neither the theoretical arguments that have been offered nor the currently available empirical evidence gives us any reason to believe that many of these processes and mechanisms exist.

In Section 3 we will sketch a number of crucial differences among the processes and mechanisms for which the 'simulation' label has been used, and explain why we think that the existence of some of these processes is uncontroversial while the existence of others is very controversial indeed. Our central theme in that section will be that it is time to stop thinking about these issues as a dispute between a pair of 'rival accounts'. There are lots of importantly different phenomena to be explained, and a wide range of theories have been proposed in an effort to explain them. Trying to partition these theories into rival camps so that we can ask, as Heal does, 'who wins, the theory theorist or the simulation theory'? (p. 47) strikes us as a pointless exercise. There is no natural or well motivated place to draw the boundary. It is much more productive to get clear about the distinctions among the very different sorts of processes and mechanisms that have been proposed, and then to ask which of them are actually used in the prediction of thought and action. In Section 4, we will consider Heal's analysis of the experimental results that we have used to argue against some of the hypotheses proposed

by simulation theorists, and in Section 5 we'll take a careful look at the theory she proposes to account for these results.

3. *The Many Varieties of Simulation (and Why It's Not Useful to Ask 'Who Wins?')*

In her attempt 'to extract an agreed core' among simulationist accounts of action and mental state prediction, Heal begins with the widely used analogy of a model aircraft in a wind tunnel:

> If we are convinced ... that a model aircraft will behave similarly to a real aircraft of the same shape, at least in a usefully wide range of circumstances, then we may test models with varying shape in varying wind speeds etc., measure the quantitative outcome in various respects and use those figures as a basis for the needed detailed predictions of actual aircraft. We use the model aircraft to simulate the real aircraft.
>
> The simulationist hypothesis is that something analogous to this takes place in many cases when we arrive at predictions about the thoughts and actions of other people. We call upon our similarity to other people, in particular the similar functioning of our minds. So we play a role *vis à vis* another whom we predict which the model aircraft plays *vis à vis* the real aircraft. (p. 46)

What Heal does not note, and what has been little noticed elsewhere in the literature, is that this idea of using one's own mind to simulate the functioning of another person (the 'target') can be elaborated in two fundamentally different ways, one of which is much more problematic than the other.

To see the less problematic elaboration suppose that you were shown an Ames Room like the one portrayed in Figure 1, and asked to predict how the things in the room would look to a particular target if she peered through the viewing window. One strategy would be to rely on theory—marshal whatever information you may have learned about how the human visual system works, combine it with what you know about this Ames Room and the situation the target will be in, and then derive a prediction. Another strategy is simply to peer through the viewing window yourself and note how things look. Then, because you think it's likely that the target's visual system works in pretty much the same way that yours does, you can safely predict that things will look the same way to her. So, if the dog looks much larger than the man to you, you can predict that the dog will also look much larger than the man to the target. In this case you have put yourself in a situation which is very similar to the target's situation and used your own reaction as a basis for predicting hers.

Much the same approach of quite literally putting yourself in the target's situation can sometimes be useful in predicting other people's beliefs. Sup-

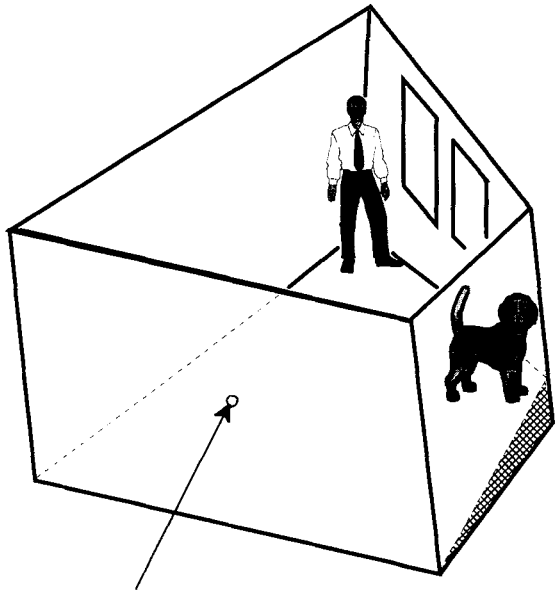

Figure 1

pose Stich knows that Nichols has paper and pencil handy and is about to add 123 + 456 + 789. If Stich wants to predict what Nichols will believe the answer is, the obvious strategy is for Stich to reach for a pencil and do the addition himself. Since Stich believes that he and Nichols are about equally competent at elementary math, after doing the problem himself he will predict that Nichols will believe the same thing he believes. Much the same strategy would work if Stich wanted to predict which of the sentences on a certain list Nichols will believe are grammatical when Nichols reads them. As we both speak very much the same dialect, all Stich needs to do is read the sentences himself and predict that Nichols will believe a sentence is grammatical if and only if he himself does. In some cases, the strategy can even be used to predict desires and emotions. If Nichols knows that Stich will soon be given a strong psychoactive drug and wants to predict its effect, he can simply take the drug himself and note what happens. If, after taking the drug, Nichols finds he has a strong desire to go to sleep, or that he feels irrepressibly cheerful for no apparent reason, he might well predict that Stich will have the same desire or the same emotion after he takes the drug.

In all of these examples, we do just what Heal suggests in the passage quoted earlier:

> We call upon our similarity to other people, in particular the similar functioning of our minds. So we play a role *vis à vis* another whom

> we predict which the model aircraft plays *vis à vis* the real aircraft. (p. 46)

Thus all of them might be taken to illustrate a strategy of prediction via mental simulation. Indeed, Heal herself uses an addition prediction case as an example of prediction via simulation (p. 51), and the grammar example was first proposed by Paul Harris (1995), another leading advocate of simulation theory. In each of the cases we've sketched, the predictor simulates the target after first putting himself in a situation which is very similar to the target's. We'll call this *actual-situation-simulation*.[3]

The vast majority of cases in which people predict the thoughts or actions of other people are not instances of actual-situation-simulation. In some cases it would be absurdly impractical to try to put oneself in the target's situation. (What will Martha do when she learns that her house has just burned to the ground?) In other cases it is near enough psychologically impossible. Often we want to predict the thoughts or actions of people whose beliefs and/or desires are different from ours in ways that are likely to affect their decisions. Actual-situation-simulation would require that we somehow contrive to have the same beliefs and desires they do, and typically that's just not a serious option. To explain how prediction is possible in these cases, simulation theorists have proposed an intriguing two part hypothesis.

The first part of the hypothesis is that some of the mental mechanisms or capacities that are invoked when we generate new beliefs or intentions or feelings can be driven not only by real beliefs and desires but also by imaginary or pretend beliefs and desires. So if we want to predict what the target would decide to do if she were offered a particular job, we can imagine (or pretend) that we have been offered that job, and that we have the desires and beliefs that we think the target has. We can then let our ordinary decision-making mechanism make a decision. If we have provided our decision-making system with the right imaginary input—beliefs and desires having the same content as the target's and, in the case of desires, the same strength—and if our own decision-making system is indeed similar to the target's, then this process driven by imaginary mental states will simulate the actual decision-making process in the target and reach the same decision. But, of course, the predictor can't treat this decision as he would a real decision. It is only an imaginary or hypothetical decision. Rather than acting

---

[3] Other examples of actual-situation-simulation are commonplace in the literature. Goldman (1995b) maintains, plausibly enough, that we typically use the process to predict which jokes other people will find amusing, and Gordon (1995) describes a case in which one person finally comes to understand his hiking companion's fright when he himself confronts the grizzly bear that frightened his companion. We should note that, as we are using the term, the target's 'actual situation' includes not only her physical and social circumstances but also her current beliefs and desires. Thus to put yourself in something close to the actual situation of a target who believes that ginseng increases sexual potency because his herbalist has just assured him that it does, you would have to do more than find an herbalist who will tell you this; you would have to *believe* him.

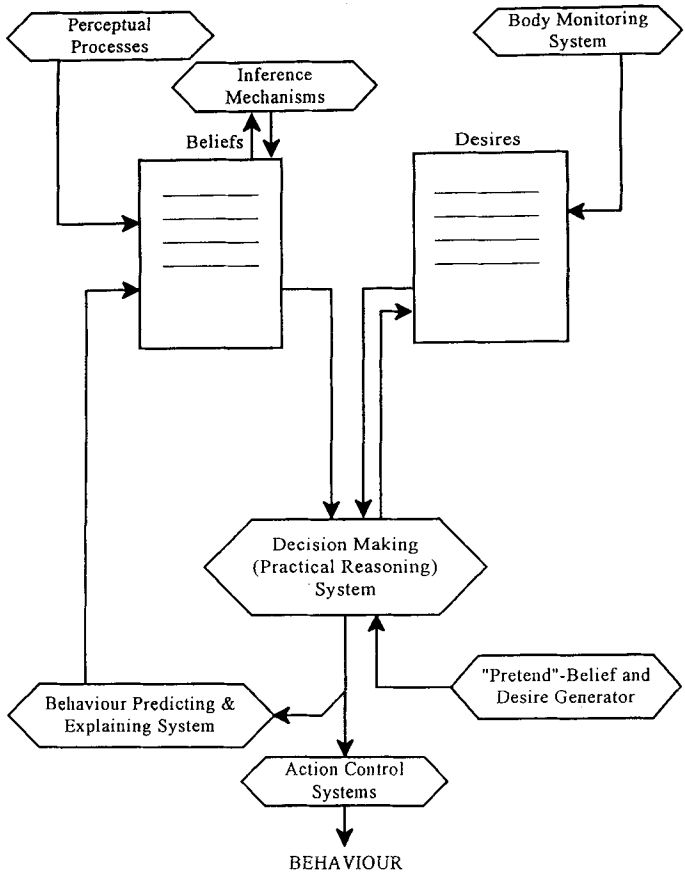

Figure 2

on that decision, the predictor uses it as a basis for predicting what the target will decide. This account of what the predictor does with the decision is the second part of the hypothesis offered by simulation theorists. In earlier papers we've described it as taking the decision 'off-line' and we've called the hypothesized two step process 'off-line simulation'. But to emphasize that the process is driven by imaginary or pretend beliefs we now propose to call it *pretence-driven-off-line-simulation*. As we'll use the term, pretence-driven-off-line-simulation is not restricted to the decision-making or practical-reasoning system. Rather it is a process which can, in principle, be linked with any mental mechanism that takes intentional states as inputs and yields any sort of mental state as outputs.

Figure 2 is a schematic rendition of the mental mechanisms underlying the pretence-driven-off-line-simulation of decisions. We first proposed this

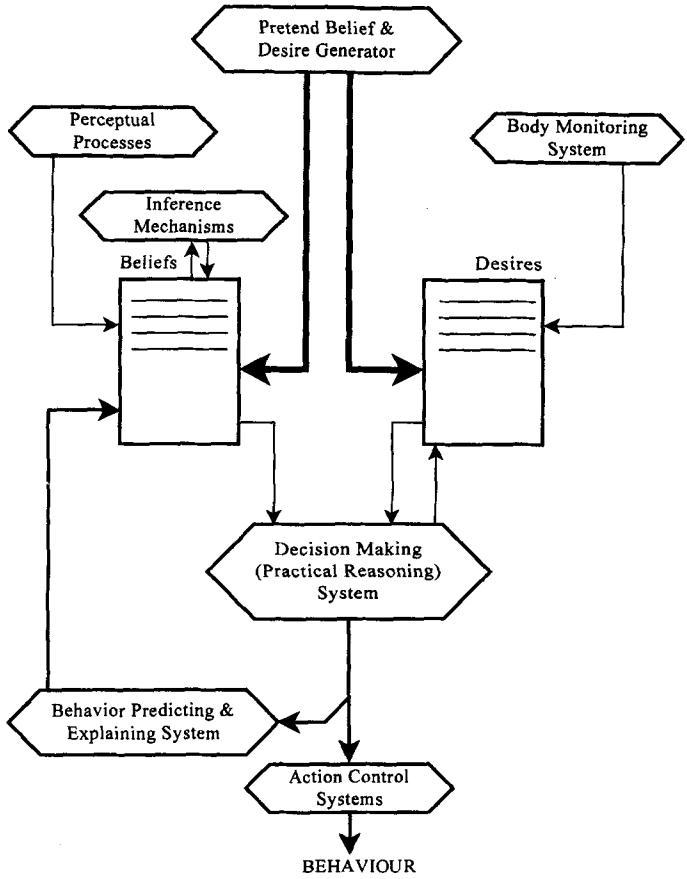

Figure 3

sketch in Stich and Nichols (1992), and Alvin Goldman, a leading advocate of simulation theory, has endorsed it with some enthusiasm and reprinted it in several of his own publications. (Goldman, 1993, 1995a). However, we're now inclined to think that Figure 3 may do a better job of depicting pretence-driven-off-line-simulation of decisions, because it makes clear that as part of the process of deriving a decision prediction, the pretend beliefs can be used by the inference mechanism to generate further pretend beliefs.

We're fairly confident that Heal thinks pretence-driven-off-line-simulation is the process underlying many cases in which we predict the thoughts or actions of other people. Here's what she says about it:

> [T]he idea is that we possess certain capacities to develop new thoughts (beliefs, feelings, intentions, etc.) from given thoughts, by

> reasoning and becoming aware of connections. These capacities we can exercise both on straightforward beliefs and desires and also (very importantly) on thoughts that are mere suppositions or imaginings. So, for example, I can infer an actual belief that p from existing beliefs that q and r or I can hypothesize q and r and then see that p would follow. Now we can harness these capacities to work through the implications of thoughts, not only on our own behalf but also to enable us to predict others. To do so we take the contents of the beliefs and desires of the other as material for our own reasoning and reflective capacities. When this occurs the mind of the would-be predictor and the mind of the person to be predicted (if all goes well) proceed through parallel evolutions and arrive at similar end states; each of these persons develops thoughts from thoughts, using his or her capacities. Of course in the simulated person the upshot will be a real belief, feeling or decision whereas in the simulator the upshot will be something of a merely imagined or hypothetical character. But the contents will be the same and the simulator can thus use the upshot, as it occurs in him or her, as a basis for prediction of the other. The prediction is derived by a simulation process, not by calling on some theory about how minds work. (p. 46)

Heal is less enthusiastic than we are about the utility of 'boxological' sketches in explaining hypotheses about mental capacities. She is concerned that unless the diagrams include 'a realistically large number of boxes and a realistically large number of arrows connecting them', one might be 'seduced' into thinking 'that the output of a box is determined by one sort of input only' (p. 52). We don't know whether anyone has actually been seduced by the diagrams. But Heal's warning is certainly well taken. The diagrams are intended only as crude sketches of *some* of the mechanisms and processes underlying various cognitive capacities, and it should be borne in mind that they do not pretend to depict all the mechanisms and processes that may affect people's actual performance.

On our view, pretence-driven-off-line-simulation is an ingenious and important hypothesis about how people *might* produce predictions about other people's decisions and actions when they can't or don't exploit actual-situation-simulation. However, it is important to note that actual-situation-simulation and pretence-driven-off-line-simulation are *very* different processes. Pretence-driven-off-line-simulation posits mental mechanisms (the pretend belief and pretend desire generators) and mental capacities (the capacity of the decision-making and inference systems to work more or less normally on 'pretend' inputs, and the capacity to take decisions off-line and use them as the basis of predictions about other people's decisions) which are not needed for actual-situation-simulation. *From the fact that actual-situation-simulation is sometimes used we can infer nothing at all about whether these*

additional mechanisms and capacities exist, or about whether pretence-driven-off-line-simulation is ever used to predict decisions or actions.

Now as Heal sees the debate between simulationists and theory theorists, particularly theory theorists like us,

> theory theory is imperialistic. It says 'Everything comes from theory.' Simulation theory is the contradictory of this. So it says only the more modest 'At least some important things don't come from theory but come from simulation.' (p. 48)

Heal is, of course, well aware that *some* simulation theorists are much less modest in their claims[4] and that sometimes theory theorists sound much less imperialistic. But she is 'fairly confident that Stich and Nichols understand the distinction in the way set out.' (fn 2, p. 48). Is she right? There is no easy answer here since Heal has failed to note the distinction between actual-situation-simulation and pretence-driven-off-line-simulation. As we mentioned earlier, she even includes an example of actual-situation-simulation (a math prediction case) as one illustration of the kind of mental simulation she advocates. Apparently, then, she intends to use 'simulation' in an inclusive way which covers both actual-situation-simulation and pretence-driven-off-line-simulation. And when the term is used in this way, Heal is right to endorse simulation theory, but wrong about us. On our view, there is absolutely nothing problematic or controversial about any of the examples of actual-situation-simulation that we've sketched. They illustrate a strategy of prediction that people obviously can use and sometimes do. If the imperialistic version of theory theory denies that some predictions come from *actual-situation-simulation*, while simulation theory 'says only the more modest 'At least some important things don't come from theory but come from simulation'' *broadly construed*, then we cheerfully concede that simulation theory wins and theory theory loses. Though, of course, we deny that we are, or ever have been, advocates of the imperialistic version of the theory theory.

Some might protest that this easy victory for simulation theory is a reason to be unhappy with Heal's preferred terminology. But, for two reasons, that's not a protest we'll endorse. First, we suspect that, from time to time at least, many advocates of simulation theory use the term with an extension as broad or even broader than the one Heal has in mind.[5] Second, we think it's silly to argue about terminology. Our inclination is to embrace the inclusive use of 'simulation' that Heal and other simulationists prefer, and to note that when the term is used in this way it turns out to be quite useless in posing interesting questions because the things that count as simulations are not a theoretically interesting category. What we really want to know is not whether *some* predictions about other people's thoughts or actions are

[4] See, for example, Gordon 1995, 1996.

[5] See, for example, Gordon 1995, Harris 1995 and Goldman 1995b.

subserved by *some* sort of simulation process—the answer, obviously, is yes—but which of the very heterogeneous class of processes that count as simulations are used in which kinds of predictions. To ask those questions, and to be sure they are not confused with the question of whether simulation theory (construed broadly) is true, we'll adopt a terminology bristling with hyphens. We've already distinguished actual-situation-simulation from pretence-driven-off-line-simulation and argued that while one of them clearly does exist, the existence of the other is far from obvious. In the next few paragraphs we'll elaborate a bit on a theme that Heal herself has very sensibly stressed: There are lots of different kinds of pretence-driven-off-line-simulation, and each of these must be assessed on its own merits.

In setting out her own version of simulation theory, Heal cautions her fellow 'simulationists' about a mistake that it is important to avoid:

> [T]he question of whether a kind of psychological state can be simulated, and if so how fully and with respect to what kind of property, must be looked at separately for each kind of state. We must also beware of supposing that we can just help ourselves to the idea that there are 'mental systems' subserving each kind of psychological state and there is such a thing as running these systems 'off-line'. To suppose that we make clear to ourselves what we mean by simulating some kind of psychological state by using this kind of talk is to be taken in by language and to put the cart before the horse. If we favour the 'off-line running' idiom, the right to use it must be established separately for each sort of state by specifying some kind of activity that is running the system 'off-line'. (pp. 56–7)

The warning is one we would certainly endorse. Indeed, we think there are two rather different warnings here, and though we agree with both of them it's important to pull them apart.

One of them starts with the observation that psychological states like beliefs or desires or emotions can play a role in a variety of different psychological processes subserved by a variety of different mental mechanisms or systems. When a theory posits the existence of simulated (or 'pretend' or 'imaginary') beliefs or desires or emotions, it can't be the case that the simulated state has *all* the same causes and effects as the real state being simulated. If it did it wouldn't be a simulation; it would be the real thing. So a theorist who posits simulated states must be careful to specify which properties of the real state the simulation has and which it lacks. Often the way to do this will be to specify those mental processes or mechanisms in which the simulated state plays the same role as the real one, and those in which it does not. This is a theme to which we will return in Section 4 when we consider Heal's account of the differences between real beliefs and desires and their mental simulations.

Heal's other warning is that the question of whether a kind of state can be simulated must be looked at separately for each kind of state. It is entirely

possible, for example, that while beliefs can be simulated, desires, or perceptions, or some of the emotions cannot. Another possibility, not explicitly mentioned by Heal, is that some mental mechanisms or processes can be exploited in one or another kind of mental simulation process, while other mechanisms or processes cannot, either because they have no access to pretend inputs or because they cannot process pretend inputs. So, for example, it might be the case that the inference mechanism in Figure 3 can be run on hypothetical (or 'pretend') beliefs, enabling a cognitive agent to reason about what would be the case if the hypothetical beliefs were true, but that the decision making or practical reasoning mechanism can't be run on hypothetical beliefs and desires. The crucial point here is that, as Heal would surely agree, claims that one or another mental mechanism can be run on pretend or hypothetical inputs must be evaluated separately. From the fact that one mental mechanism can be used in pretence-driven-off-line-simulation, nothing whatever follows about whether other mental mechanisms can be used in this way. The point will loom large in Section 4 when we consider Heal's very interesting proposals for explaining various experimental findings. And it becomes even more important in light of recent claims, by evolutionary psychologists and others, that the cognitive mind is 'massively modular' (Samuels et al., in press). If the mental processes that we think of as inferences are in fact subserved not by one mental mechanism (as suggested by the Inference Box in Figures 2 and 3) but by dozens of separate 'Darwinian modules', then the job of determining which of these (if any) can be exploited in pretence-driven-off-line-simulation will be truly daunting.

Now suppose it is the case that pretence-driven-off-line-simulation is invoked when we predict how a target's store of beliefs will be modified if she acquires a new belief. But suppose it is also the case that in order to derive a prediction about what the target will decide to do upon acquiring a new belief, our cognitive system makes use of a tacit folk psychological theory about how beliefs and desires combine to generate decisions and actions. Perhaps, as theorists like Fodor have suggested, this theory contains a number of interconnected *ceteris paribus* laws, along with a list of exceptions and special cases. The mental mechanism that produces decision-predictions and action-predictions has access to our beliefs about the target's desires and to the newly updated information about what the target will believe (generated by pretence-driven-off-line-simulation). In deriving a prediction, the mechanism integrates this information with the *ceteris paribus* laws and the list of special cases. This 'hybrid' account is surely a coherent theoretical possibility. If it is correct, who wins? Well, since we've embraced the inclusive interpretation of simulation theory, and since the hypothesis we're considering invokes pretence-driven-off-line-simulation to predict the beliefs that the folk psychological theory needs in order to produce predictions about action decisions, it looks like simulation theory wins again. Here as before, it might be thought that this easy victory is an indication that the line between theory theory and simulation theory had been drawn in the

wrong place. But we don't think there is any better place to draw it. On our view, the lesson to be learned from hybrid hypotheses like the one we've sketched is that it is a mistake to think that the class of plausible theories can be divided into two rival camps in any unique and theoretically interesting way.

Before bringing this section to a close, we want to offer one more example in support of our contention that the 'simulation' label has become quite useless because simulations are not a theoretically interesting kind. Suppose that the hypothesis proposed in the previous paragraph is only half right. We do exploit a tacit theory in making predictions about other people's decisions, but we don't exploit pretence-driven-off-line-simulation in making predictions about how their belief store will be updated when they acquire a new belief. That too, let us assume, is subserved by a tacit folk psychological theory which specifies how inference works and how beliefs are updated. Now let's tell a story about how such a process might work. Nichols is asked to predict what Stich will believe about who would become president of the United States if Bill Clinton were to die. On the hypothesis we're considering, the tacit theory about reasoning and belief updating is activated. What, it is asked, would Stich infer if he came to believe that President Clinton had died? The answer, of course, is going to depend on what else Stich believes. If, for example, he also believes that when the President dies the Vice-President immediately becomes President, and if he believes that Al Gore is the Vice-President, and if he doesn't have beliefs (about conspiracies or impending coups, for example) that might defeat the obvious inference, then, we might suppose, the tacit theory will churn out the prediction that Stich will believe that Gore will become President. But *does* Stich have these further beliefs? They weren't specified as part of the problem, and Nichols has never discussed these matters with Stich. We have more interesting things to talk about. So how can Nichols (or the hypothetical cognitive system subserving the inference) know whether or not to assume that Stich has these further beliefs? An obvious and appealing answer is that, in a wide range of cases, Nichols (or the relevant cognitive system) simply assumes that Stich believes what he himself believes. Indeed, the hypothesis that people use their own beliefs as the default value in assigning beliefs to a target is just about the only serious proposal we know of for how this sort of tacit-theory-driven account of belief prediction could be elaborated.

Now what shall we say about the process that uses the predictor's own beliefs as default values in figuring out what the target believes? Is that an example of simulation? Well, it's not pretence-driven-off-line-simulation, and it's not actual-situation-simulation. But no matter. It seems to accord well enough with Heal's rough and ready characterization of simulation as a process in which we 'call upon our similarity to other people, in particular the similar functioning of our minds'. So it looks like we've found yet another kind of simulation, and that this kind of simulation is embedded within what Heal herself would certainly take to be a paradigm case of a theory theory, since:

> individual predictions, for example of the decision of another person, are arrived at by taking information about that person, integrating it with the information of the theory and deriving the prediction. (p. 45)

It begins to look like it's no easy thing to elaborate a theory about the mechanisms underlying the prediction of other people's thoughts and actions that does not include *some* kind of simulation. So Heal is right. Simulation theory wins. Simulation is *everywhere*. What better indication could there be that simulation is a theoretically uninteresting category?

4. *Cognitive Penetrability*

Our central theme in the previous section was that 'simulation' is a hopelessly heterogeneous category of no theoretical utility. Thus there is little point in debating whether or not simulation theory is true. But we have no such qualms about more specific hypotheses which purport to explain a particular cognitive capacity by invoking one or another sort of ('hyphenated') simulation process and specifying which mental mechanism or mechanisms that kind of simulation is supposed to be running on. So, for example, we take the hypothesis sketched in Figure 3 to be at least the first step in developing an interesting, empirically testable account of the psychological processes underlying our capacity to predict other people's decisions. According to this hypothesis these predictions rely on pretence-driven-off-line-simulation in both the inference and the practical reasoning system. In several of our earlier papers we have argued that this hypothesis cannot readily account for various experimental results that can be handled easily by a competing explanation of decision-predictions which assigns a central role to a tacit folk psychological theory. In 'Simulation and Cognitive Penetrability', Heal presents an enlightening and sophisticated critique of that argument. Her critique has two parts. In the first part, she challenges a central theoretical claim invoked in our argument, a claim which links pretence-driven-off-line-simulation with 'cognitive impenetrability', though she then concedes that even if our claim is mistaken the experimental results we cite still pose a problem for hypotheses like the one sketched in Figure 3. In the second part of her critique, Heal sets out a new and interesting proposal about the sorts of mental processes that can (and cannot) be accessed by pretence-driven-off-line-simulation, and argues that when this proposal is incorporated into an account of decision predictions, the experimental results no longer pose a problem for the simulationist. In this section our focus will be on the first part of Heal's critique, and our conclusion will be that Heal's challenge is correct, and so too is her concession that this does not help the simulationist to explain the experimental findings. In the next section we'll take up Heal's proposed reformulation of hypotheses like the one sketched

in Figure 3. We'll offer no quick summary of the conclusions in that section, since the pattern of agreement and disagreement will be more complex.

The central theoretical claim in the argument that Heal criticizes is the contention that if your predictions about other people's decisions are generated by a process like the one sketched in Figure 3, then what you don't know (or at least what you don't know *about psychology*) can't hurt you. A bit less cryptically, the idea is that if a tacit folk psychological theory plays a central role in decision-predictions, and if there is some psychological process about which that theory either has no information or has mistaken information, then predictions about decisions affected by that process are likely to be wrong. By contrast, if decision-predictions are produced by pretence-driven-off-line-simulation as sketched in Figure 3, then even if the predictor has a tacit folk psychological theory, the shortcomings of that theory should be irrelevant to the accuracy of the predictions, since the theory (if it exists) plays no role in decision-predictions. On the hypothesis sketched in Figure 3, if the predictor's inference and decision making systems are similar to the target's, and if the predictor's pretend belief and desire generators provide pretend beliefs and desires that match those of the target, then it seems that the prediction has to be correct. In earlier papers, borrowing a colourful term from Zenon Pylyshyn, we summarized this by saying that, on the hypothesis sketched in Figure 3, decision-predictions are *cognitively impenetrable*—they are not affected by the predictor's beliefs or tacit theories about psychological processes, nor are they affected by the predictor's ignorance about psychological processes. Theory driven predictions, by contrast, are *cognitively penetrable*; ignorance or error about psychological processes can lead to mistaken predictions.

In the next step of the argument we assembled experimental evidence suggesting that people's predictions about other people's decisions *are* cognitively penetrable. Our strategy for finding appropriate examples was to look at some of the very surprising phenomena (or 'effects' in the jargon of the trade) discovered in recent years by cognitive social psychologists. The effects that interested us are those in which people's decisions are influenced in systematic though unexpected ways by apparently unimportant aspects of the circumstances in which the decision is made. For example, Langer (1975) reported an experiment in which subjects who have agreed to buy a ticket in a lottery are later offered an opportunity to sell their ticket. What Langer found was that subjects who are offered an opportunity to choose their own ticket set a significantly higher price than subjects who are simply handed a ticket when they agree to buy one.[6] A second example came from a number of studies exploring what has been called the 'endowment effect'. These studies compared the decisions of subjects who are given an object of

[6] In a recent paper, Kuhberger et al. 1995 question whether Langer's results are replicable. Though the issue is clearly an important one, Heal assumes for the sake of the argument that the effect is genuine (p. 64). We propose to do the same.

modest value (in several experiments it was a mug engraved with their school's logo) and offered an opportunity to sell it back, with the decisions of subjects who are offered a choice between being given the mug and being given a sum of money. Those who were given the mug typically held out for significantly more money than those who were offered the choice between getting the mug and getting the cash. These are surprising results, and the fact that they are surprising suggests that if people have and use a folk psychological theory in predicting these sorts of decisions, that theory contains no information about the effects or the processes that give rise to them. Thus, if people were asked to predict what a target would decide in a Langer style lottery, or in an endowment effect experiment, they might well get it wrong. Suppose, on the other hand, that people use a strategy like the one sketched in Figure 3 to predict these sorts of decisions. If the predictions are elicited under circumstances designed to make it easy for the predictors to generate the right pretend beliefs about the situation, and if there aren't any systematic differences between the predictor's inference and decision-making systems and the target's, then, we argued, we should expect the predictor's pretence-driven-off-the-line decisions to mirror the targets' real decisions and thus the predictions should generally be accurate.

One study that we cited in support of our contention that predictions like these are not subserved by pretence-driven-off-line-simulation used a pair of video tapes in which a target was asked to set a selling price for a lottery ticket in a Langer-style situation. The two tapes were identical except for the fact that on one the target got to choose his own ticket while on the other the ticket was simply given to him. Subjects were shown one or the other tape and asked to predict the target's decision. Those viewing the choice version of the tape did not predict higher values than those viewing the no-choice version of the tape. In another study, Loewenstein and Adler (1995) set out to determine whether people could predict *their own* decisions in endowment effect situations. The subjects were allowed to examine an engraved mug. They were then asked to imagine that the mug had been given to them and to decide on a minimum price for which they would sell it back. Shortly after that, the subjects actually were given a mug just like the one they had examined, and really were offered an opportunity to sell it back. What Loewenstein and Adler found was that subjects systematically underpredicted the price they themselves actually set. We argued that results like these are hard to explain on the assumption that decision predictions are subserved by a process like the one in Figure 3, and are much easier to explain if we assume that the predictions are subserved by a tacit theory—a theory which includes no information about the Langer and endowment effects.

The first part of Heal's critique takes aim at our contention that predictions produced by mechanisms like the one in Figure 3 are not cognitively penetrable. To begin, she notes that the claim that simulation driven predictions are not cognitively penetrable is actually a conjunction of two quite independent claims. One of these, which Heal calls 'Absence' is that '[i]f a predic-

tion is derived by simulation then ignorance (e.g. of a theory about the workings of what is simulated, of what to expect in this case etc.) will not prevent arrival at a prediction and arrival at a correct one, if the simulation is correctly set up and the particular process goes smoothly' (p. 50). 'This,' Heal agrees, 'must surely be accepted by a simulationist' (p. 50). The other half of the claim that a process is not cognitively penetrable Heal labels 'Presence'. It maintains that '[b]elief in positive mis-information about what will happen in a case could not affect a prediction arrived at by using simulation in that case' (p. 50). This claim, Heal argues, is simply mistaken. To make the point, she suggests a number of different scenarios in which a 'belief in positive mis-information' might affect a prediction process which, like the one sketched in Figure 3, relies on pretence-driven-off-line-simulation. Though Heal does not present the argument in this way, we think it is useful to divide the kinds of effects she has in mind into three categories which we'll call *downstream, upstream* and *midstream.*

Downstream effects are those that take place after the pretence-driven-off-line operation of the decision making and inference systems. At this point in the process the predictor's cognitive system has produced an off-line decision and passed it on to the behaviour predicting and explaining system. Now suppose that before any of this pretence-driven-off-line processing had begun, the predictor had formed one or more beliefs that strongly suggest the target would make a particular decision. In the example that Heal proposes, the issue at hand is whether the target will accept a certain job offer, and the predictor believes that a friend who knows the target very well thinks that the target will decide to accept.[7] Let us suppose the pretence-driven-off-line-simulation yields a decision not to accept the job, and that that's the right decision for it to reach, since the target will go through a parallel process and reach the same decision. Nonetheless, Heal observes, it is entirely possible that, quite unconsciously, the belief about the friend's prediction prevents the cognitive system from accepting the output of the pretence-driven-off-line-simulation. That output is, after all, only one relevant consideration available to the cognitive system as it goes about settling on a belief about what the target will do. And the pretence-driven-off-line-simulation account is certainly not committed to the view that the output is not defeasible. So here we have a case in which, although pretence-driven-off-line-simulation is being invoked, a belief clearly does affect what a predictor will ultimately come to believe (and report) about the target's decision. Because of this 'last minute contamination of the output of the simulation' (p. 53) the predictor gets it wrong.

Though Heal does not develop the point at any length, a largely analogous case can be made for upstream contamination of pretence-driven-off-line-

[7] As Heal tells the story, the predictor has already formed the belief that the target will accept, but wants to put that belief aside and run a simulation as 'an open minded check'. But as we see it, this is an unnecessary complication. Heal's point can be made cleanly and persuasively without it.

simulation decision predictions. Advocates of pretence-driven-off-line-simulation have been less informative than one would like about how the pretend belief and desire generators go about their work. But on any plausible story these devices are going to have to rely heavily on the predictor's beliefs about the target and the target's situation, and under appropriate circumstances just about anything the predictor believes might affect the beliefs that the pretence generators rely on. Suppose, for example, that the predictor mistakenly believes that all devout Catholics think they will go to Hell if they work in hospitals that perform abortions. If the predictor also believes that the target is a devout Catholic and that the job the target is considering is in a hospital that performs abortions, then the predictor's pretend belief generator may provide the pretence-driven-off-line-simulation with a misleading input that will lead to a mistaken prediction.

At this point it might be protested that, for two quite different reasons, these examples of upstream and downstream effects do not really constitute a refutation of the claims we made about cognitive penetrability. First, we claimed only that if a prediction is produced by pretence-driven-off-line-simulation then ignorance or misinformation *about psychological processes* could not affect the outcome of the prediction process, and in the cases we've described the misinformation is not about psychological processes. But this would at best be a debater's quibble, and it would be easy enough to cook up more outlandish examples in which a predictor's mistaken beliefs about psychological processes contaminated a pretence-driven-off-line-simulation either upstream or downstream. Second, neither upstream nor downstream effects show that predictions produced by pretence-driven-off-line-simulation are cognitively penetrable since, in the upstream case, the effect occurs before the simulation gets started and the simulation is run on the wrong inputs, while in the downstream case the prediction is *not* produced by pretence-driven-off-line-simulation because the output of the simulation process is ignored or overridden. But we are not much impressed with this move either. For the crucial point of Heal's critique is that experiments which show that decision predictions are cognitively penetrable do not constitute a 'quick empirical knock down for simulationism' (p. 44) and on this point she is surely right. Moreover, as Heal notes, upstream and downstream effects are not the only ways in which the predictor's beliefs can affect the outcome of pretence-driven-off-line-simulation. There are also midstream effects in which misinformation can alter the way in which a pretence-driven-off-line-simulation proceeds.

Perhaps the most obvious way in which a predictor's beliefs (whether they are mistaken or not) can affect the internal (or 'midstream') operation of a pretence-driven-off-line-simulation is via an indirect route through the predictor's moods or emotions. It is plausible to suppose that if a person is angry, or terrified, or depressed, or perhaps even if she is exceptionally happy, this will affect the way in which her inference and decision-making processes operate, and that this effect might be much the same whether the processes are operating on-line or off-line. And, of course, a person's beliefs

are among the principal factors affecting her moods and emotions. So if a predictor is livid because she mistakenly believes that someone has just intentionally erased all her computer files, her pretence-driven-off-line-simulations may produce outcomes that are very different from the ones that they would have produced had she been more calm. If the predictor is angry and the target is not, the result may be a mistaken prediction. Heal also suggests a way in which a predictor's mistaken beliefs can have a midstream effect on pretence-driven-off-line-simulation without first having an effect on the emotions. Decisions, she notes, can take varying amounts of time. One of the things that may be going on when a decision takes a lot of time is that the person making the decision (or her cognitive system) is searching memory for relevant information. It sometimes happens that the information that tips the balance in one direction or the other is only recalled after an extended search. Now if pretence-driven-off-line-simulation is used to predict such a decision, and if the predictor spends less time searching memory than the target did, it might well happen that the predictor would get it wrong. Moreover, Heal suggests, the predictor's beliefs may be one factor determining how much time her cognitive system will devote to memory search. Thus, for example, in the job decision case, if the predictor believes that a knowledgeable friend of the target's thinks that the target will accept the job, this may shorten search time, and lead to a mistaken prediction.

On our view, Heal's conjectures about the ways in which a predictor's beliefs might affect the outcome of a decision prediction subserved by pretence-driven-off-line-simulation—midstream, downstream and upstream—are all very plausible. What they show, we think, is that it was a mistake to claim that pretence-driven-off-line-simulation is not cognitively penetrable. Decision predictions can fail as the result of cognitive penetration, even when the predictor and the target are psychologically similar, and the predictor is in an excellent position to generate pretend beliefs and desires similar to the target's. What follows about the experiments that we've described above? On the one hand, these experiments certainly do not *prove* that a mechanism like the one in Figure 3 is not being invoked. For in each instance it *might* be the case that the predictors have some belief or emotion that is affecting the process and undermining the accuracy of the prediction. So Heal is certainly right that there is 'no quick empirical knockdown' (p. 44) to be found in these experiments. On the other hand, although it is *possible* that the experimental findings are the product of a systematic cognitive penetration of a pretence-driven-off-line-simulation process, we don't know of even one serious suggestion about an unsuspected belief or emotion (or a cluster of these) that might explain any of the results. If there were any *specific* proposals about contaminating beliefs, it would be easy enough to design experiments to test them. And we're betting that the experiments would rule them out. Moreover, we're pleased to find that Heal is betting with us. '[T]o be frank,' she writes, 'it seems to me unlikely that the kind of cognitive penetration outlined above is the explanation of the wrong predictions about others which have been discussed in the literature so far. To get

light on them I suggest that we need to look elsewhere' (p. 55). So it appears that on the topics discussed in this section, Heal's views and ours are very much in agreement. We all agree that predictions subserved by pretence-driven-off-line-simulation *can* be cognitively penetrated, and that this can result in mistaken predictions. But we also agree that this is not a plausible explanation of the mistaken predictions in the experiments described above or of other mistaken predictions that have been discussed in the literature. If the predictions were subserved by a tacit folk psychological theory, there would be a ready explanation of the mistakes—the folk theory is wrong or incomplete. However, because Heal thinks that decision predictions are subserved by pretence-driven-off-line-simulation she recognizes that she must offer some other explanation of the mistakes. And once again, we agree. In the explanation Heal proposes, the notion of rationality plays a central role. At that point our agreement ends.

5. *Rationality and the Restricted Simulation Hypothesis*

5.1 Heal's Hypothesis: The Rationality Restriction

The starting point for Heal's explanation of the mistaken predictions made by experimental subjects is the observation, noted in Section 3, that we should not expect pretend beliefs and desires to have exactly the same effects as real ones. If the pretence-driven-off-line-simulation account of decision prediction sketched in Figure 3 is correct, then pretend beliefs and desires will have to trigger some mental mechanism which ensures that pretend beliefs (and any inferences drawn from them) are not simply added to the 'belief box' and left there after the episode of pretence-driven-off-line-simulation is completed. Pretend beliefs and desires will also have to trigger a mechanism that makes sure the decisions reached at the end of the pretence-driven-off-line-simulation are not treated as real decisions and acted upon, but instead are shunted off to the behaviour predicting and explaining system. So obviously pretend mental states will have effects that their real counterparts do not have. There will also be effects that real beliefs and desires have and pretend ones do not—effects like generating real ('on-line') decisions that lead to real action. In order to explain prediction errors made by experimental subjects, Heal focuses on a particular class of effects that real mental states have but which, she hypothesizes, their imaginary counterparts lack. Real beliefs and desires sometimes lead to the formation of beliefs, desires, emotions or decisions in ways that are not *rational* (or, as she sometimes says, not *intelligible*); however it is Heal's hypothesis that pretend or imaginary beliefs and desires typically do not. When the pretend states that serve as the input to a pretence-driven-off-line-simulation lead to the formation of further pretend mental states, these new pretend states are rationally related to the input.

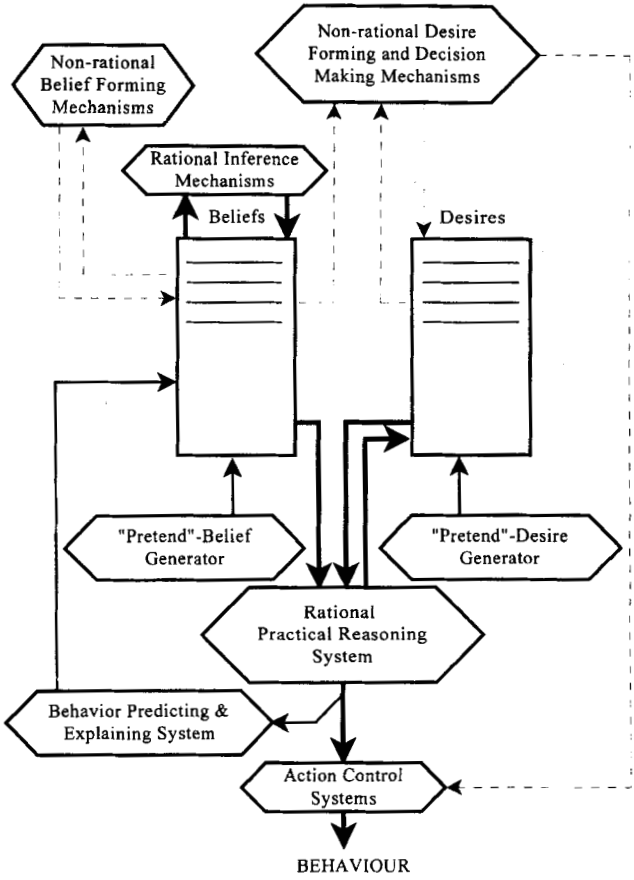

Figure 4

> The kind of simulationism I would like to defend says that the only cases that a simulationist should confidently claim are those where (a) the starting point is an item or collections of items with content, (b) the outcome is a further item with content, and (c) the latter content is rationally or intelligibly linked to that of the earlier item(s). (p. 56)

Since pretend states typically do not have the power to engage the mental mechanisms or processes that result in irrational outcomes, we should expect that pretence-driven-off-line-simulation would in general not be capable of predicting these outcomes. Though Heal is not enthusiastic about boxological diagrams, we think Figure 4 may be a useful way of capturing her hypothesis. In that Figure we've posited the existence of Rational Inference

and Rational Practical Reasoning systems, and we've also posited a cluster of non-rational belief, desire and decision-forming mechanisms. The darker lines indicate the causal links that pretend mental states share with their real counterparts, while the lighter lines indicate causal processes in which only non-pretend mental states can play a role; their pretend counterparts just do not engage these mechanisms.

If decision predictions are produced by pretence-driven-off-line-simulation in the way suggested by Figure 4, then we would expect that people would be much more likely to make a mistaken prediction when the target's actual decision is affected by non-rational processes like those indicated by the lighter lines. And that, according to Heal, is the most likely explanation for the mistaken predictions made by subjects in the experiments we have cited in this and earlier papers. In most of these cases, she maintains, the decisions that subjects are being asked to predict are heavily influenced by non-rational factors. Consider, for example, the lottery ticket buy back experiment. Why do the target subjects (the subjects whose decision is to be predicted) set a higher price for the tickets when they have been offered a choice?:

> Are they responding intelligibly to one of the few relevant factors of which they are allowed to be aware? Or are they in a situation where, precisely because of the lack of relevant information, non-rational factors come strongly into play?
>
> It looks at first sight as if the latter is the case. . . . if so, the failure of simulation to predict [the decision] is entirely to be expected. (p. 64)

Similarly, Heal maintains that in the endowment effect experiment there is no obvious way in which the mere fact that the subject now owns the mug makes it more valuable or 'better as a mug'. Since 'intelligible desires arise out of the perceived presence of value' it appears that the stronger desire to keep the mug which arises after one owns it is not rational or intelligible. 'Hence it is not a case which the simulation approach can predict' (p. 65). Heal offers much the same analysis of a case in which people fail to predict that targets who are offered a choice of qualitatively identical products (e.g. nightdresses) tend to prefer those on the right hand end of the array, and of another case in which people fail to predict the 'belief perseverance' effect in which subjects who have been given evidence in support of a particular belief (e.g. that they are good at detecting the difference between real and fake suicide notes) retain the belief even after they learn that the evidence was faked. In both cases, Heal maintains, the phenomena to be predicted are irrational, and thus the sort of restricted pretence-driven-off-line-simulation that she advocates should not be expected to anticipate them (p. 62).

5.2 Some Reasons for Skepticism About Heal's Appeal to Rationality

We are inclined to be more than a bit skeptical about Heal's attempt to explain predictive failures by restricting pretence-driven-off-line-simulation to rational processes. However, we also think that the problematic appeal to rationality can be stripped away from Heal's hypothesis, and that the resulting account of pretence-driven-off-line-simulation is perhaps the most interesting and empirically plausible one yet proposed. In this section we'll set out our reasons for rejecting Heal's appeal to rationality. In the section that follows we'll try to reconstruct her hypothesis in a way that does not invoke rationality.

As we see it, the principal reason to be dissatisfied with Heal's proposed explanations of predictive failures is that the notion of rationality to which she appeals is unexplained, obscure, and suspiciously elastic—it expands and contracts as required, to make the theory fit the facts. Heal insists that the contentful states which are the starting points and the outcomes of a simulation must be 'rationally linked'. But what, exactly, does this mean; what does the rationality restriction come to? Unfortunately, she does not offer a detailed explanation of her notion of a rational link, nor does she suggest any references to the large and rather contentious literature debating the nature of rationality. It is clear that rationality, for Heal, is a fairly broad notion which can apply not only to decisions and processes of belief formation, but also to emotions and the processes that give rise to them. It is also clear that Heal's notion of rationality imposes a much more permissive standard than we might expect if we think of logic, statistics and decision theory as providing the best accounts of rationality in various domains. 'Rational', she tells us:

> should not be interpreted in a narrow and demanding sense, where some linkage counts as rational only if it withstands leisurely scrutiny by logically acute and formally aware minds and thus conforms to, or even improves on, the practice sanctioned by current best accounts of deductive logic, statistics and probability theory, decision theory, etc. One reason that this notion is inappropriate for the simulation approach is that the approach recognizes that people do their reasoning, form their stances and take their decisions in real time, often under pressure and facing the need to handle a great amount of complex material. We, quite properly, rely on short cuts, on the analogies and saliences that strike us, on the intellectual habits encouraged by our communities and so forth. Hence not everything 'irrational' in the strict sense falls outside the domain of simulation. (pp. 57–8)

The problem with all of this is that Heal tells us almost nothing about *which* departures from rationality in the 'narrow and demanding sense' are rational enough to be handled by simulation and which are not. And without some

guidelines to help us decide how much irrationality ('in the strict sense') is too much irrationality, Heal's view is so flexible that it can explain any prediction that people might make about rationally problematic inferences, emotions or decisions, whether that prediction is correct or incorrect. If people's predictions fail, well, that's not a problem; the process they are predicting is irrational. But if people's predictions succeed, once again it's not a problem; the process they are predicting, though it may be ''irrational' in the strict sense' is not irrational enough. Since she does not suggest any way of determining in advance whether particular cases are rational in the wide and undemanding sense she commends, her hypothesis seems to be immune from any possible falsification from predictions people make about irrational inferences, decisions or emotions. If the predictions are wrong, it's a case of excessive irrationality; if the predictions are right, then the target has been rational enough.

To see the problem a bit more concretely, it is useful to consider some specific cases. In arguing for an account of simulation that explicitly does not invoke the notion of rationality, Goldman (1995b) cited an experiment by Kahneman and Tversky (1982) in which subjects were given the following text:

> Mr. Crane and Mr. Tees were scheduled to leave the airport on different flights, at the same time. They traveled from town in the same limousine, were caught in a traffic jam, and arrived at the airport 30 minutes after the scheduled departure time of their flights. Mr. Crane is told that his flight left on time. Mr. Tees is told that his was delayed, and just left five minutes ago. Who is more upset?

Kahneman and Tversky report that 96% of their subjects said that Mr Tees would be more upset and, while we know of no attempt to test the accuracy of such predictions experimentally, we suspect that the subjects are right. (Kahneman, Tversky and Goldman also assume that this is the case.) But it is hard to see why it would be rational (in the 'strict' sense) for Mr Tees to be more upset. The two travellers are in very similar situations. Both have missed their flights; neither could have done anything about it. The only difference is that Mr Crane missed his flight by 30 minutes, and Mr Tees missed his by 5. A small and apparently unimportant difference, but perhaps it is enough to make Mr Tees' reaction rational in the permissive sense invoked in Heal's hypothesis. But now compare this case with the lottery ticket case and the endowment effect case. In both of the latter, a small and apparently unimportant difference (being given a ticket vs choosing it; owning the mug vs not owning it) has a major effect on how strongly people want to keep something, or at least on the price they set for it. However, in these cases, in contrast with the missed flight case, predictors don't predict the effect. Heal is unworried by this because the targets are being irrational, though it's hard to see why they are being any more irrational than Mr Tees.

It appears that what's really motivating Heal to rule that the endowment effect and the effect operative in the lottery ticket experiments are beyond the bounds of rationality even when broadly construed, but that effects like the one manifested by Mr Tees are not, is that this is what's required to make her theory fit the facts. If we can predict it, it's rational enough; otherwise it isn't.

Much the same point emerges when we compare the belief perseverance case with a number of other examples of rationally dubious belief formation patterns like the conjunction fallacy or base rate neglect. One very familiar problem used to illustrate the conjunction fallacy is the following:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> Please indicate which of the following statements is more probable.
>
> (a) Linda is a bank teller.
> (b) Linda is a bank teller and is active in the feminist movement.

The standard finding is that most subjects judge that (b) is more probable than (a). To the best of our knowledge, there have been no published studies exploring how well observer subjects do at predicting the judgments targets will make when confronted with problems like this. However, we have done some preliminary and quite informal experiments which indicate, not at all surprisingly, that observers are quite good at such predictions. In the belief perseverance case, by contrast, the fact that people hold on to the belief when the evidence has been discredited is genuinely surprising. As Heal notes, it is an effect that people do not predict. To accommodate these facts, presumably Heal would claim that perseverance is more irrational than the conjunction fallacy. Retaining a belief when the evidence has been discredited is 'a striking case of irrationality' (p. 62), though judging that Linda is more likely to be a feminist bank teller than a bank teller is not irrational except in the 'narrow and demanding sense'. We find it hard to believe that there is anything more than special pleading going on here. What distinguishes the two cases is not that one is more irrational than the other, but that people are good at predicting one and bad at predicting the other. The only reason for suggesting that one case is more irrational than the other is that that's what Heal's theory has to say to avoid being falsified.

Heal herself is obviously a bit uncomfortable about using the term 'rational' to label the link that must obtain between input and output if a process is to fall 'within the remit of simulationism' (p. 58). Rather, she suggests, '[p]hrases like 'such that some intelligible sense or point can be seen in it' or 'such that some justificatory account of it can be given' might serve better to summarize the simulationist's notion' (p. 58). However, these locutions are, if anything, even more flexible and dodgy than 'rational—but not in the

strict sense'. There are lots of discussions in the literature that offer *some* justificatory account of both the conjunction fallacy and belief perseverance, so if that's all that is required, they should *both* be predictable by pretence-driven-off-line-simulation (Cohen, 1981; Goldman, 1986; Harman, 1986; Gigerenzer, 1991). In one quite revealing passage, Heal describes cases that fall within the remit of the sort of simulation she advocates as those for which we can 'see 'from the inside' so to speak, why such actions are done' (p. 58). We are not at all sure what 'seeing from the inside' comes to, but one plausible reading might be that, in these cases, it seems natural or intuitively obvious that the actions would be done. And what that amounts to, near enough, is that the ones whose intelligibility we can 'see from the inside' are just those that we can predict. This enables her to explain people's mistaken predictions, though it also makes her explanations entirely vacuous. In restricting simulation to cases involving 'intelligible' links, what Heal is really doing is restricting it to the cases where people typically make the right prediction. The notions of rationality and intelligibility are doing no real work at all.

5.3 Restricted Simulation Without the Rationality Restriction

What we argued in the previous section was that Heal's attempt to explain failures of prediction by invoking the notion of rationality to restrict 'the remit of simulation' is explanatorily vacuous, since the only real handle we have on her very flexible notion of rationality is that it applies when predictions come out right, and not when they come out wrong. In this section we want to propose a friendly amendment to Heal's restricted version of pretence-driven-off-line-simulation—an account which tries to turn the vacuity of her appeal to rationality from a vice to a virtue.

The starting point in our revision of Heal's theory is the observation that there are two parts to her idea of restricting pretence-driven-off-line-simulation to cases where the links are rational, and that these two parts are separable. The first part, and the part we think is really important, is that a defender of pretence-driven-off-line-simulation need not—indeed should not—claim that pretend beliefs and desires will have all the same effects as real ones. Rather, the simulationist should expect that there will be mental mechanisms or processes that are triggered by real beliefs and desires but not by their pretend counterparts. The second part of Heal's idea was to characterize mechanisms that could be triggered by both real and pretend mental states as 'rational' or 'intelligible'. It is this second part that we think Heal should jettison. So the pared down theory says only that the mind contains a variety of different mechanisms via which beliefs and desires can generate new beliefs and desires as well as decisions and emotions, and that some, but not all of these mechanisms can also be engaged by pretend beliefs and desires. Well, but which ones? you might ask. Heal's answer, the answer that got her into trouble, was: the rational ones. The answer we would urge in our friendly amendment to Heal's theory is: At this point, no one knows.

It is an empirical question, and there is simply no way of knowing in advance whether or not a given mental process can be triggered by pretend inputs as well as real ones. The only way to answer the question is to construct a detailed and empirically well confirmed account of the various mechanisms and processes that subserve the on-line formation of beliefs, desires, decisions and other mental states, and then to explore, for each of these mechanisms and processes, whether or not it can be engaged by imaginary mental states as well as real ones.[8]

The pared down version of Heal's theory that we've sketched can't, of course, say in advance whether or not people will be good at predicting the outcome of a given sort of mental process, though the theory has a readily available explanation for any systematic failure in people's capacity to predict. When people regularly get a certain sort of prediction wrong, it's because the decision (or other mental state) to be predicted is produced by a mechanism which, as it happens, cannot be driven by inputs which are merely pretend or imagined. The advocate of this sort of theory might even use systematic failures in prediction as a tool in attempting to characterize the on-line boxology of the mind. If there is independent reason to think that pretence-driven-off-line-simulation is invoked in predictions, then if predictions systematically fail for a coherent class of cases, that will count as evidence for positing a separate 'box' for the process subserving those cases—a box (or mechanism) that can be engaged by real mental states, but not by their pretend counterparts. It might be thought that this version of the pretence-driven-off-line-simulation-theory has two obvious shortcomings. It makes no predictions about the sorts of cases in which people will make systematic errors, and it can explain, albeit post hoc, any errors that happen to be discovered. But if these are shortcomings, then they are shortcomings shared by the hypothesis that people use a tacit theory in deriving their predictions. The theory theory doesn't predict any specific prediction errors either, though it too has a ready explanation for any systematic errors that are found. If people get it wrong, then the tacit folk theory they are using is mistaken or incomplete.[9]

At this point one might begin to worry that *no* empirical evidence will help us decide between the theory-theory and the restricted version of the pretence-driven-off-line-simulation-theory that we've been sketching. But that would be too pessimistic a conclusion. What is the case, however, is that empirical work aimed at deciding between these two accounts is going to have to explore some new strategies. In the experiments discussed in this paper, the strategy was to look for cases in which normal adult subjects

---

[8] There is reason to think that Heal might have some sympathy with this response: See, for example, the passage we quoted in Section 3, where she insisted that 'the question of whether a kind of psychological state can be simulated, and if so how fully and with respect to what kind of property, must be looked at separately for each kind of state' (p. 56).

[9] This point was first made clear to us by Meredith and Michael Williams.

failed to predict decisions or inferences that they and other normal adults would make, and to show that these prediction failures occurred even under circumstances when the opportunity for simulation seemed optimal. That strategy is a good one if the version of the pretence-driven-off-line-simulation-theory being considered is of the unrestricted sort that simulationists usually proposed prior to Heal's recent work. But if the simulationist follows Heal and recognizes that there may be lots of mental mechanisms that real mental states can engage but imaginary ones cannot, the strategy will be of little help, since the theory-theory and the restricted version of the pretence-driven-off-line-simulation-theory are equally good at explaining predictive failures.

In attempting to decide between these two alternatives, a strategy that might be more promising is to look at *abnormal* subjects, since when normal systems break down or fail to develop, the two theories suggest that we might find quite different patterns of dissociation. On the theory-theory account, an agent's own decision-making and her predictions about other people's decision-making are subserved by largely separate systems. Thus it might be possible for the former system—the on-line decision maker—to be impaired while the latter is quite normal. A subject with an impaired decision making system and an intact decision-predicting system might be an excellent folk psychologist—able to predict what other people will do in normal or near normal ways—though her own on-line decisions are made in quite bizarre ways which she herself is not able to predict; or she might be simply incapable of reaching any decision at all in many cases, though she can easily predict what others would do if they were in her situation. If there are subjects like this, they would pose a prima facie challenge for restricted versions of the pretence-driven-off-line-simulation-theory, since on that account the capacity to predict decisions and the capacity to make one's own on-line decisions are subserved by the same mental mechanism. So anything that does serious damage to the on-line decision making system should damage the decision predicting system as well. On the restricted pretence-driven-off-line-simulation account we might expect to find that subjects whose on-line decision-making capacity is impaired in some partial way would have parallel partial impairments in their capacity to predict other people's decisions. If that is the typical pattern of impairments, it would count as impressive evidence in favour of the restricted pretence-driven-off-line-simulation-theory and against the theory-theory. Recent brain imaging technology may provide yet another source of evidence that can help us decide between the two theories. The pretence-driven-off-line-simulation-theory suggests there should be one part of the brain that is active in both on-line decisions of a given sort and predictions about decisions of that sort, while the theory-theory says that on-line decisions and predictions about them are subserved by different mental mechanisms which may be located in different parts of the brain. We don't expect that these lines of inquiry, or others that may be proposed, will provide conclusive evidence for or against either account of decision prediction. We don't, and never

have, supposed that either side will score a 'quick empirical knockdown' (p. 44). However, we do think that the debate can be and should be cast as an empirical one, to be settled by the creative use of evidence deriving from many different parts of cognitive science. There's lots of exciting work to be done before we really understand how one mind can predict what another will decide.

Department of Philosophy
Rutgers University

Department of Philosophy
The College of Charleston

References

Cohen, L.J. 1981: Can Human Irrationality Be Experimentally Demonstrated? *Behavioral and Brain Sciences*, 4, 317–70.

Davies, M. 1994: The Mental Simulation Debate. In C. Peacocke (ed.), *Objectivity, Simulation and the Unity of Consciousness*. Oxford University Press, pp. 99–127.

Gigerenzer, G. 1991: How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases'. *European Review of Social Psychology*, 2, 83–115.

Goldman, A. 1986: *Epistemology and Cognition*. Cambridge, MA.: Harvard University Press.

Goldman, A. 1993: *Philosophical Applications of Cognitive Science*. Boulder, CO.: Westview Press.

Goldman, A. 1995a: Empathy, Mind and Morals. In M. Davies and T. Stone (eds), *Mental Simulation: Philosophical and Psychological Essays*. Oxford: Blackwell, 185–208.

Goldman, A. 1995b: Interpretation Psychologized. In M. Davies and T. Stone (eds), *Folk Psychology*. Oxford: Blackwell, pp. 74–99.

Gordon, R. 1995: The Simulation Theory: Objections and Misconceptions. In M. Davies and T. Stone (eds), *Folk Psychology*. Oxford: Blackwell, pp. 100–22.

Gordon, R. 1996: 'Radical' Simulation. In P. Carruthers and P. Smith (eds), *Theories of Theories of Mind*. Cambridge University Press, pp. 11–21.

Harman, G. 1986: *Change in View*. Cambridge, MA.: MIT Press.

Harris, P. 1995: From Simulation to Folk Psychology: The Case for Development. In M. Davies and T. Stone (eds), *Folk Psychology*. Oxford: Blackwell, pp. 207–31.

Heal, J. 1994: Simulation vs. Theory Theory: What Is at Issue? In C. Peacocke (ed.), *Objectivity, Simulation and the Unity of Consciousness*. Oxford University Press, pp. 129–44.

Heal, J. 1995a: Replication and Functionalism. In M. Davies and T. Stone (eds), *Folk Psychology*. Oxford: Blackwell, pp. 45–59.

Heal, J. 1995b: How to Think About Thinking. In M. Davies and T. Stone (eds), *Mental Simulation: Philosophical and Psychological Essays*. Oxford: Blackwell, pp. 33–52.

Heal, J. 1996a: Simulation and Cognitive Penetrability. *Mind and Language*, 11, 44–67.

Heal, J. 1996b: Simulation, Theory and Content. In P. Carruthers and P. Smith (eds), *Theories of Theories of Mind*. Cambridge University Press, pp. 75–89.

Kahneman, D. and Tversky, A. 1982: The Simulation Heuristic. In D. Kahneman, P. Slovic and A. Tversky (eds), *Judgment Under Uncertainty*. Cambridge University Press.

Kuhberger, A., Perner, J., Schulte, M. and Leingruber, R. 1995: Choice or No Choice: Is the Langer Effect Evidence Against Simulation? *Mind and Language*, 10, 423–36.

Langer, E. 1975: The Illusion of Control. *Journal of Personality and Social Psychology*, 32, 311–28.

Loewenstein, G. and Adler, D. 1995: A Bias in the Prediction of Tastes. *The Economic Journal: The Quarterly Journal of the Royal Economic Society*, 105, 929–37.

Nichols, S., Stich, S. and Leslie, A. 1995: Choice Effects and the Ineffectiveness of Simulation. *Mind and Language*, 10, 437–45.

Nichols, S., Stich, S., Leslie, A. and Klein, D. 1996: The Varieties of Off-Line Simulation. In P. Carruthers and P. Smith (eds), *Theories of Theories of Mind*. Cambridge University Press, pp. 39–74.

Samuels, R., Stich, S. and Tremoulet, P. in press: Rethinking Rationality: From Bleak Implications to Darwinian Modules. In E. LePore and Z. Pylyshyn (eds), *Rutgers University Invitation to Cognitive Science*. Oxford: Blackwell.

Stich, S. 1996: *Deconstructing the Mind*. Oxford University Press.

Stich, S. and Nichols, S. 1992: Folk Psychology: Simulation vs. Tacit Theory. *Mind and Language*, 7, 29–65.

Stich, S. and Nichols, S. 1995: Second Thoughts on Simulation. In M. Davies and T. Stone (eds), *Mental Simulation: Philosophical and Psychological Essays*. Oxford: Blackwell, pp. 87–108.