Stich & His Critics
Edited by Dominic Murphy & Michael Bishop

Replies to Critics

I was both delighted and flattered when Mike Bishop and Dominic Murphy first told me that this volume was in the works. And I was quite thrilled when I got the list of distinguished philosophers who had agreed to contribute, though I quickly came to realize that the job of preparing responses to such an outstanding group of critics would be a daunting one indeed. Before plunging in, I want to express my gratitude to Mike and Dominic, to all the contributors, and to the series editor, Ernie LePore. Thanks are also due to Boris Yakubchik, who helped assemble the references for my replies.

Reply to Egan

Egan calls her rich and interesting paper "Is There a Role for Representational Content in Scientific Psychology?" and she tells us that "the question that forms the title of this paper is … addressed directly to Stich." It's an important question, certainly, and an appropriate one since, as she notes, my current position on the matter is less than clear. In *From Folk Psychology to Cognitive Science* I offered an account of our ordinary practice of ascribing content to mental states like beliefs. On that account, content ascriptions have what Egan dubs "the R properties" – they are vague, context sensitive and observer-relative; they are influenced by the ideological similarity between the attributor and the target and by the similarity between the reference of the terms that the attributor uses to express the belief and the reference of the terms that the target would use to express it. I went on to argue that a taxonomy of mental states which relies on commonsense content attributions is ill suited for use in a scientific theory whose ultimate aim is to explain behavior. If that's right, I concluded, then content has no role to play in scientific psychology. But in *Deconstructing the Mind*, published 13 years later, my view was much more guarded. "The jury," I said, "is still out on the question of whether successful science can be constructed using intentional categories." (p. 199) What prompts Egan's question is not that I changed my mind, but that I haven't offered any clear and focused explanation of *why* I changed my mind. I haven't said why I am no longer convinced by the arguments in *Folk Psychology* aimed at showing that the answer to Egan's title question is *no*.

Some philosophers might reject those arguments because they believe that *Folk Psychology* mischaracterized the commonsense concept of content, and that on a correct account a content based taxonomy would not have the R properties.[1] That's a concern that I think should be taken seriously; I'll return to the topic later. But it is not a concern that Egan shares. She thinks that my arguments aimed at showing that a content taxonomy will have the R properties are persuasive (ms 3) and that *Folk Psychology* "established" the point (ms 9). She also says that the case against content in *Folk*

[1] See, for example, Fodor 1987, 1990, 1991.

Psychology is a "tour de force" (ms 3). (Wow! A tour de force! Thanks, Frankie.) But one of the many valuable things I've learned from Egan is that "tour de force" is not a success term. Egan thinks that the arguments aimed at showing that content has no role to play in scientific psychology fail, and that content does play a crucial role in the explanatory strategy of scientific psychology.

To make the case, Egan focuses on computational psychology and sketches an account of the explanatory strategy of that central branch of psychology in which content does play an important role. On her account, content does not play an *individuative* role in computational cognitive theories. Rather, content or semantic interpretation is necessary to explain how, in a given context, the "abstractly characterized" processes of a computational theory "constitutes the exercise of a *cognitive* capacity." (ms 15, emphasis in the original) "[T]he semantic interpretation forms a bridge between the intentionally characterized explananda of the theory and the abstract, mathematical characterization of the mechanism that constitutes the explanatory core of a computational theory." (ms 15-16) Though she does not make the point explicitly, I think it is clear where Egan thinks that the anti-content arguments of *Folk Psychology* fail. Those arguments were aimed at showing that the laws or principles invoked in the explanatory core of a computational theory should not be couched in terms of content, and on this Egan agrees. But what the *Folk Psychology* account failed to see, Egan suggests, is that appeal to content is necessary if computational psychology is to offer a satisfying account of the link between those abstractly characterized processes and "the questions that define a psychological theory's domain," – questions that are "typically couched in intentional terms." (ms 15)

Egan's account of the explanatory strategy of computational cognitive theories, which she elaborates at much greater length elsewhere (Egan 1992, 1995, 1999, 2003), is an important contribution to the philosophy of psychology. It is well informed, insightful, carefully defended, and vastly more sophisticated than the story about computational psychology that I told in *Folk Psychology*. It isn't the only game in town, however. Other writers, most notably Cummins (1989) and Ramsey (2007) have offered competing accounts that are also far more sophisticated than anything to be found in *Folk Psychology*. However, I can't appeal to these competing accounts to buffer me from Egan's critique, for while Cummins and Ramsey differ with Egan on many important details, they *agree* that content has an important role to play in the explanatory strategy of computational psychology. So one might be tempted to conclude that if any of the accounts in this vicinity are right, then one of the core conclusions of *Folk Psychology* is wrong: contrary to what I argued there, intentional categories *do* have an important role to play in cognitive science. But I think that this would be too hasty. The issues here are rather more complicated.

To see why, it will be useful to take a closer look at Cummins (1989). In that book, as I've noted, Cummins gives an account of the explanatory strategy of the computational theory of cognition, an account in which content plays an important role. However, Cummins does not assume that the notion of content that is of use in computational psychology is the same as the one invoked in commonsense psychology. Indeed, in the first chapter of the book he warns against that assumption.

> To suppose that "commonsense psychology" ("folk psychology"), orthodox computationalism, connectionism, neuroscience, and so on all make use of the same notion of representation seems naïve. Moreover, to understand the notion of mental representation that grounds some particular theoretical framework, one must understand the explanatory role that framework assigns to mental representation. It is precisely because mental representation has different explanatory roles in "folk psychology," orthodox computationalism, connectionism, and neuroscience that it is naïve to suppose that each makes use of the same notion of mental representation." (Cummins, 1989, pp. 13-14)

Later in the book, Cummins argues that the assumption is not just naïve, it is false.

> The kind of meaning required by the CTC [the computational theory of cognition] is, I think, not Intentional Content [the kind of content invoked in commonsense psychology] anymore than entropy is history. There is a connection, of course, but at bottom representation in the CTC is very different from intentionality. (Cummins, 1989, p. 88)

If Cummins is right, then neither Egan's account of the explanatory strategy of computational psychology nor competing accounts like Cummins' and Ramsey's, pose a challenge to the arguments in *Folk Psychology* that were aimed at showing that content had no role to play in scientific psychology. For those arguments were quite explicitly focused on the sort of content that plays a role in commonsense psychology, *not* on some perhaps related but importantly different notion of content invoked in computational theories of cognition. So if Cummins is right, the question that Egan poses in her title is ambiguous. On one reading it is asking about the sort of representational content invoked in commonsense psychology, and on the other it is asking about an importantly different sort of content. It is entirely possible that the answer to her question is *no* on the first reading and *yes* on the second.

Of course this irenic outcome presupposes that Cummins is right about the difference between representational content as it is invoked in commonsense psychology and representational content as it is invoked in computational psychology. Egan might not grant that the two notions of content are different, but I am inclined to think she should. "A semantic interpretation of a computational system," she tells us, "is given by an *interpretation function* that specifies a mapping between equivalence classes of physical states of the system and elements of some represented domain…. In computational models of perception, the content ascribed to internal states of the device will be determined by the distal properties tracked by these internal states." (ms 14 & 16) On this account, the ascription of content to states in the visual system will not have some of the central features that, I argued in *Folk Psychology*, characterize the ascription of content in commonsense psychology. Consider, for example, the role of what I called "ideological similarity". In *Folk Psychology*, I proposed a thought experiment inspired by the real case of Mrs. T, a woman suffering from a degenerative brain disease. (Stich 1983, pp. 53-60) For much of her life, Mrs. T, who was born in 1880, had an avid

interest in politics and was well informed on the topic. She was deeply shocked by the assassination of President William McKinley in 1901. However, in her seventies her illness began to manifest itself. To make the point as cleanly as possible, I assume that the only effect of the disease was a progressive loss of memory. In the early stages of her illness, Mrs. T had trouble remembering recent events like who had been elected in a Senate race she had been following or where she had left her knitting. As the affliction got worse, she had trouble remembering who the president was, who George Washington was, or where the White House was located. Shortly before her death, she was asked "What happened to McKinley?" and immediately responded, "He was assassinated." But she could not say whether assassinated people die, or what death is, nor even whether she herself was dead. I maintained that commonsense psychology would no longer attribute the content *McKinley was assassinated* to the mental state underlying Mrs. T ability to *say* "McKinley as assassinated," and argued that this was because commonsense content attribution is sensitive, in a more or less holistic way, to similarity between the attributor's entire stock of beliefs and the target's entire stock. Not everyone agrees with this interpretation of the thought experiment, but Egan is not among the dissenters. She believes that the arguments in *Folk Psychology* have established that content ascriptions are sensitive to ideological similarity. What makes this important for current purposes is that, on Egan's account, attribution of content to internal states in computational models of perception is *not* ideologically sensitive. Recall that according to Egan, in these models "the content ascribed to internal states of the device will be determined by the distal properties tracked by these internal states." And Mrs. T's loss of memory would have little or no effect on that tracking relationship. Thus a computational modeler of Mrs. T's visual system would characterize the content of states in that system in exactly the same way before the onset of her illness and after it had become very severe. So it looks like the irenic response to Egan's title question is indeed a live option. The answer is *no* if the representational content in question is the sort presupposed by commonsense psychology and *yes* if it is a notion of representational content like the ones sketched by Egan or Cummins.

I don't think it would be wise to leave the matter here, however, since there is another line of argument in Egan's paper that can't be escaped by invoking the ambiguity of "representational content." Thus far we've focused on Egan's account of computational psychology, which occupies much of her paper. In her final section she argues that while beliefs, desires and the other propositional attitudes that loom large in commonsense psychology do not now, and likely never will, play any role in *computational* psychology, "*other* branches of scientific psychology *do* invoke beliefs and desires." (ms 21, italics in the original) If that's right, then it seems that the answer to Egan's title question is *yes*, even when "representational content" is understood as the kind of content that plays a role in commonsense psychology. The two branches of psychology that Egan mentions are attribution theory and the part of developmental psychology that "attempts to characterize the commitments that infra-linguistic infants bring to their interactions with the world" (ms 21). I don't think that either of these is good example to make her point. Attribution theory has hardly been flourishing in recent years, and while developmental psychology most certainly has been flourishing, Egan's claim that "developmental theories attribute to infants beliefs … about how objects move

in space" is hardly uncontroversial. But I'm not inclined to quibble here, since if Egan's examples are not optimal, there are lots of other examples that she might invoke. In cognitive social psychology, work in the "heuristics and biases" tradition has uncovered startling and important facts about the ways people form beliefs and preferences;[2] in developmental and abnormal psychology, work on how normal children acquire an understanding of mental states, and why children with disorders like autism don't, is up to its ears in talk of beliefs and desires;[3] the list would be easy to expand. I knew relatively little about this work when I wrote *Folk Psychology* – indeed much of it had not yet been done – and I rather naïvely assumed that all of "serious, scientific psychology" would gradually evolve into computational theories. But clearly that has not happened, and I see no reason at all to suppose that it will.

Where does all this leave us? Well, if the relevant parts of cognitive social psychology, developmental psychology and abnormal psychology count as scientific psychology – and I most emphatically maintain that they do – and if, as seems to be the case, these branches of psychology invoke representational content, then on either reading of Egan's title question, the answer is *yes*.

Let me close my reply to Egan's paper with a caveat, a puzzle and a concern. It certainly *seems* to be the case that various flourishing branches of psychology invoke familiar mental states like beliefs and desires, and exploit the commonsense representational content of these states in a variety of ways. But appearances can be deceiving. One of the lessons of Cummins' penetrating analysis of the computational theory of cognition is that, while it might seem that these theories are invoking the commonsense notion of content, they are actually relying on an importantly different notion. Might something similar be true in cognitive social psychology, developmental psychology, etc.? It strikes me as unlikely, but without the sort of careful analysis of the explanatory structure of these sciences – analyses of the sort that Cummins, Egan and Ramsey have given us for computational psychology – it is hard to be sure. Unfortunately, while philosophers have been avid consumers of work in cognitive social psychology and developmental psychology, there has been almost no work on the explanatory structure of these sciences. These are important projects for a variety of reasons, and it is to be hoped that philosophers will soon pursue them. But until that work is done, it would, I think, be wise to hedge our bets on whether these sciences are making substantive use of the commonsense notion of representational content. That's the caveat.

The puzzle is posed by a pair of facts. First, a number of branches of scientific psychology that apparently invoke commonsense content are clearly flourishing. Second, if the account in *Folk Psychology* is even roughly on the right track, commonsense content attribution has properties that would seem to pose serious obstacles to a fruitful science. It is vague, observer-relative and context sensitive. Moreover, folk practice sometimes classifies together mental states that differ dramatically, attributing the same

[2] See, for example, the essays collected in Kahneman, Slovic & Tversky (1982) and Gilovich, Griffin & Kahneman (2002).
[3] Nichols & Stich (2003).

content to psychological states in Einstein's head, in a brain-damaged person's head and in a dog's head.[4] Somehow the sciences in question succeed in sidestepping the problems that a content-based taxonomy of mental states might engender. How do they do it? As I note in my reply to Godfrey-Smith, his work offers some suggestive hints. But we won't have a completely satisfying answer until these parts of psychology attract the sort of detailed and insightful philosophical analyses that Cummins, Egan and Ramsey have provided for computational psychology.

Finally, the concern. Throughout this response I have been assuming, along with Egan, that the analysis of commonsense content attribution offered in *Folk Psychology* is more or less on the right track. But I think that's a dangerous assumption. In that analysis I used a philosophical method which was standard at the time and had been widely used in philosophy since Plato. I conjured imaginary scenarios, asked what "we" would say about them, and relied on my own intuitions (and the intuitions of a handful of colleagues and graduate students) as the sole source of evidence. In recent years I have become increasingly skeptical of that method. A growing body of literature suggests that the intuitions of mostly white, mostly Western, mostly male academics, all of whom have survived the curious selection practices that lead to graduate school fellowships and jobs at universities, often do not coincide with the intuitions of people who are not members of the philosophers' club.[5] If our project is to understand how ordinary folk go about the process of content attribution, then these findings raise a serious concern: Why should we think philosophers intuitions are a good source of evidence?

Reply to Godfrey-Smith

The aim of Godfrey-Smith's paper is "to cast representationalism within a different overall philosophical framework, supplied by recent philosophy of science" (ms 14) – more specifically by work aimed at understanding model-based scientific theorizing. In much of his paper, Godfrey-Smith lets "what remains of [his] hair down a little," and presents some ideas that seem to him to be "promising, even though they are unorthodox and in some places disconcerting." (ms 3) I am not in the least disconcerted by his ideas. Indeed, I find many of them quite congenial. However, given the constraints on length in this volume and the need to cover a lot of material in order to put the pieces of his new framework in place, Godfrey-Smith had no choice but to paint with a very broad brush. Since many important points are discussed very briefly, I think that an in depth response to his innovative ideas would best be postponed until the view has been set out in more detail. What I can offer instead is an endorsement of one of the main conclusions that Godfrey-Smith defends and an expression of hope that, when more fully developed, an intriguing part of his framework might help to resolve the puzzle I raise at the end of my reply to Egan.

[4] *Deconstructing the Mind*, pp. 26-27.
[5] See, for example, Nichols, Weinberg & Stich (2003) and Swain, Alexander & Weinberg (forthcoming).

The conclusion that I am happy to embrace is that "[f]ifty years from now … most of the late 20th century literature on mental representation will appear to have tried to lay an overly simple and regimented framework onto a very complicated and mixed reality." (ms 18) Cummins (1989), whose work is discussed in my reply to Egan, hints at much the same prediction when he suggested that it is naïve to suppose that folk psychology, orthodox computationalism, connectionism and neuroscience all make use of the same notion of representation. At the end of my reply to Egan, I expressed the hope that cognitive social psychology, developmental psychology and other branches of psychology get the same sort of careful and informed scrutiny that Egan and others have lavished on computational psychology, and I raised the possibility that these parts of psychology might turn out *not* to be using the commonsense notion of representational content. What Godfrey-Smith would likely add – and I would concur – is that it might also turn out that each of these branches of psychology is invoking a different notion of representation. Or as he might prefer to put it, they might each be emphasizing different features of his "basic representationalist model". In *Deconstructing the Mind*, where my focus was not on representation but on the closely related notion of reference, I suggested that linguistics, anthropology, evolutionary biology and the history and sociology of science might *each* require a somewhat different notion of reference, and that in some of these areas of inquiry it might turn out that two or more distinct kinds of reference are explanatorily useful. (*Deconstructing*, p. 45) So while Godfrey-Smith and I may have gotten there via rather different routes, I share with enthusiasm his pluralism about scientifically useful notions in the representation family.

The puzzle that I raised at the end of my reply to Egan began with the observation that a number of branches of contemporary psychology that are clearly flourishing appear to invoke the commonsense notions of belief and desire and the commonsense notion of representational content. However, if I am right about commonsense content attribution, it is vague, context sensitive, observer-relative, and attributes the same content to a very heterogeneous collection of beliefs and a very heterogeneous collection of desires. One might think that these features would pose a major obstacle to the construction of a fruitful science. What is puzzling is that they don't. Perhaps Godfrey-Smith's idea of viewing representationalism as an instance of model-based theorizing in science can point the way toward a solution to this puzzle. According to Godfrey-Smith,

> This kind of scientific work operates by constructing and exploring hypothetical, usually simple, systems that are intended to have some relevant *resemblance* relation to a real "target" system that we are trying to understand. All the quirkiness, vagueness, and context-sensitivity of the notion of "resemblance" are supposed to be in play here. Part of the point of model-based work in science is that one can try to develop a model system that has useable similarities to a target system while being unclear, indefinite, and changeable about exactly *which* features of the model are supposed to resemble features of the target, and unclear or changeable about the degree and kind of resemblance intended. (ms 6, emphasis in the original)

So perhaps it is just a mistake to think that the features of commonsense content that I used to generate the puzzle are an *obstacle* to productive scientific theorizing. If the quirkiness, vagueness and content-sensitivity that is endemic to model-based theorizing usually do not pose a problem when models are used in physics, chemistry and biology, there is no reason to suppose they will pose a problem in psychology either. Godfrey-Smith repeatedly stresses the *flexibility* of models in science and touts it as one of their main virtues. Perhaps that flexibility is facilitating, not hindering, the impressive progress witnessed in those parts of psychology that appear to invoke commonsense content. If Godfrey-Smith is right, science (or at least model-based science) is a lot messier and a lot less explicit than many philosophers of science have supposed, and scientists are much more adept at coping with the mess – indeed flourishing in it – than I supposed when wrote *From Folk Psychology to Cognitive Science*.

Reply to Devitt & Jackson

Both Devitt and Jackson focus on my critique of ontological arguments that invoke the strategy of semantic ascent and appeal to a theory of reference. Though they agree that the critique raises some important problems, they disagree sharply on how to react to those problems. Devitt shares my view that we should reject the arguments, though he thinks the reasons I offer are misguided, and offers some reasons of his own. Jackson, by contrast, thinks that the arguments can be retained provided that they invoke the right theory of reference, and he offers a criterion for choosing the right theory – "the one which is of interest when we do ontology." (ms 16) In this reply, I'll discuss Devitt's paper first, then turn to Jackson's.

Devitt

The reasons I set out for rejecting ontological arguments that use the strategy of semantic ascent and appeal to claims about reference rely heavily on my discussion of what the theory of reference is trying to do, and on this issue Devitt and I agree on some points and disagree on many others. To focus in on this complex pattern of agreement and disagreement, I'll sketch the path that led me to an exploration of the foundations of the theory of reference. The ontological debate that first raised these issues for me was, of course, the debate over eliminativism.

The story starts with the observation that while arguments for eliminativism vary significantly in detail, they all have the same structure. The first premise of the argument maintains that commonsense mental states like beliefs and desires can be viewed as posits of a widely shared commonsense theory – "folk psychology" – which underlies our everyday discourse about mental states and processes, and that terms like 'belief' and 'desire' can be viewed as theoretical terms of this folk theory. The second premise maintains that folk psychology is a seriously mistaken theory. This premise has been defended in many ways, with different authors focusing on different putative defects. The conclusion that eliminativists draw from these two premises is that beliefs, desires and the other posits of folk psychology do not exist. But, as Bill Lycan (1988) pointed

out with characteristic verve and clarity, the conclusion does not follow from these premises taken alone. To fill the gap, Lycan noted, most eliminativists either explicitly or tacitly rely on some version of the description theory of reference for theoretical terms. Description theories are not the only game in town, however. Inspired by the influential work of Putnam (1975) and Kripke (1972), many philosophers, including Devitt himself (Devitt, 1981; Devitt & Sterelny, 1999), have embraced one or another version of the causal-historical theory of reference. And if a theory in that vicinity provides the correct account of reference for the theoretical terms of a commonsense theory, then the argument for eliminativism fails even if the premises are true. With a single caveat, I think Devitt is in agreement with all this. The caveat is that description theories and causal-historical theories do not exhaust the options. Information-theoretic theories like Dretske's (1981) and teleological theories of the sort developed by Millikan (1984) and Papineau (1987) are also in the running. Devitt is quite right about this. So thus far we are in complete agreement.

If eliminativists' arguments explicitly – or more often tacitly – invoke an assumption about the reference of theoretical terms in a folk theory, then even if we grant that the first two premises of the eliminativists' argument are true, we will have to determine which theory of reference provides the correct account for these terms in order to assess the soundness of the argument. But prior to plunging into the debate over the merits of competing theories of reference, it is important to be clear about just what counts as getting a theory of reference right. What facts is a theory of reference attempting to describe or account for? What job is the theory expected to do? It is curious and puzzling that contemporary analytic philosophers, some of whom have offered detailed and insightful accounts of the goals, methods and explanatory strategies in other disciplines, have been notably unreflective about their own. While there may be some exceptions to this – some areas of philosophy in which philosophers have been concerned with their own methodology – the theory of reference clearly was not one of them. At the time *Deconstructing* was written, one could not go to the library and read what theorists in this area had said about their goals, since, to a good first approximation, they hadn't said anything at all. So I set out to describe some of the projects that people might have in mind when they put forward a theory of reference or debate the virtues of competing theories. At this point, Devitt is still on board, I think. Like me, he believes that it is important to get straight on what the goals of a theory of reference are, and he has offered his own account of what they should and should not be trying to do.

In *Deconstructing*, I distinguish two families of answers to the question, "What is a theory of reference trying to accomplish?" The first family of answers – which I call the *folk semantics* family – aims to come to grips with the fact that appeal to *intuitions* plays a central role in debates about the virtues of competing theories of reference; more often than not these intuitions concern what various terms would denote in bizarrely counterfactual situations. So, for example, in *Naming and Necessity*, Kripke (1972) asks us to imagine that the incompleteness theorem was actually proved by a man called 'Schmitt' and that after some skullduggery the theorem was attributed to Gödel. On a description theory, an "ordinary man" whose only interesting belief about Gödel is that he is the discoverer of the incompleteness theorem would be referring to Schmitt when he

uses the name 'Gödel'. But our intuitions tell us otherwise, and this, it was widely agreed, is a major embarrassment for the description theory for proper names. Similarly, Putnam introduced us to the imaginary planet of Twin Earth where the liquid that falls as rain and fills the lakes is XYZ not H_2O, and this spawned an enormous literature exploring our intuitions about the reference of 'water' when used by Earthlings, Twin Earthlings and a bewildering variety of travelers who are magically transported from one planet to the other. (Pessin & Goldberg, 1996) The correct theory of reference, it is widely assumed, must accommodate most of these intuitions.

What sort of project could make sense of all this intuition mongering? What goal could be served by it? One obvious idea turns on the notion of a tacit theory that has played a large role in cognitive science. People, cognitive scientists tell us, have a significant number of tacit theories – bodies of mentally represented information (or misinformation) which they cannot explicitly state but which manifest themselves in their intuitive judgments and behavior. No doubt the most widely discussed of these is grammar, a tacit theory which, according to Chomsky and his many followers, is exploited in the production and comprehension of natural language and manifests itself in people's intuitive judgments about grammatical properties of utterances. Folk psychology and folk physics are also widely thought to be tacit theories. It is plausible to hypothesize that there is also a tacit theory, call it *folk semantics*, which guides people's intuitive judgments about what terms refer to. If the goal of a theory of reference is to describe this tacit theory, then collecting intuitions about a wide range of cases, and insisting that the right theory of reference must be compatible with most of these intuitions, would be a perfectly sensible methodology. This gives us one account of what it is to get a theory of reference right. The right theory is the one that correctly describes the tacit theory that (*inter alia*) guides reference intuitions.

That's not the end of the story about the folk semantics account, however. In motivating the idea that reference intuitions might be subserved by a tacit theory, I mentioned three suggestive analogies: folk psychology, folk physics and grammar. But there is an important difference between grammar and the other two. This difference emerges when we ask whether the hypothesized tacit theory is true. It is important to keep in mind that in each of these three cases, there are actually two theories under discussion. On the one hand, there is the theory constructed by a psychologist or linguist or philosopher that purports to describe a tacit theory. That theory is true if it correctly describes the tacit theory represented in the relevant people's heads. But what about the tacit theory itself. What determines whether or not that theory is true? In the case of folk physics and folk psychology, the basic outlines of the answer are straightforward enough. Folk physics is a theory about the principles governing the behavior of middle-sized physical objects, and folk psychology is a theory about states and processes underlying behavior. So the tacit theory is correct if it correctly describes those principles or processes. In the case of folk physics we know from the work of McCloskey (1983), Clement (1983) and others that important parts of the theory are false. In the case of folk psychology, the jury is still out; whether or not folk psychology is largely true is a central issue dividing eliminativists and their opponents. But now what about grammar? Here the dominant view – the one advanced by Chomsky and his followers – is that a sentence

is grammatical in a dialect if and only if it is classified as grammatical by the grammar represented in the minds of speakers of that dialect. On this account, while a linguist can certainly be mistaken in her characterization of the grammar of a dialect, the grammar itself – the tacit theory she is trying to describe – cannot be wrong in what it entails about the grammaticality of sentences, since what makes a sentence grammatical in a dialect is simply the fact that the grammar tacitly known by speakers of that dialect entails that it is grammatical.

With this distinction in hand, let's return to the theory of reference. On the folk semantics account, the goal of a theory of reference is to correctly describe the tacit folk semantic theory represented in the heads of some group of speakers. But what is the status of the claims made by that tacit theory? Here, it seems, there are two possible answers. If we pursue the analogy with grammar, then folk semantics is *constitutive* for reference: a term (in a dialect) refers to an object if and only if the folk semantic theory entails that the term refers to the object. So there is an important sense in which folk semantics can't be wrong. On the other hand, if we pursue the analogy with folk physics or folk psychology, then folk semantics might well be mistaken. Whether or not it is mistaken depends on how well it matches up with the facts about reference, and folk semantics itself does not create these facts. So to know whether or not folk semantics is correct, we'll have to pursue some other inquiry. Just as the science of physics is the standard by which we decide whether folk physics is correct, and the science of psychology is the standard by which we decide whether folk psychology is correct, it will be the job of some appropriate branch of science to tell us whether folk semantics is correct.

Now if the job of a theory of reference is to describe a tacit folk semantic theory, then so long as we are focused on accomplishing that job it matters little which of these options is correct. The issue is important, though, when the theory of reference is pressed into service in metaphysical or ontological arguments like the argument for eliminativism. For what we need to know in assessing that argument is whether or not terms in a seriously mistaken theory refer. On the grammar analogy, the pronouncements of an internally represented folk semantics settle the matter. But on the folk physics and folk psychology analogy, the pronouncements of folk semantics might be quite mistaken. It's not a folk theory but an appropriate branch of science that tells us about the reference of various sorts of terms. So if we are concerned to assess the eliminativists' argument, then we're left with only two options: adopt the grammar analogy, or forget about folk semantics and explore what the appropriate science has to tell us about reference. Having reached this point in *Deconstructing*, I went on to argue that neither of these options is viable, and thus that the eliminativists' argument is fatally flawed. Devitt has a much more sanguine view of the second option, and that is, I think, by far our most interesting and important disagreement. I'll turn to that disagreement shortly. But first, let me say where I think Devitt and I agree and disagree with respect to the idea that the goal of a theory of reference is to describe a folk semantic theory. The pattern of agreement and disagreement here is rather complicated.

Let's start with the following question: What do philosophers who debate the virtues of competing theories of reference think that they are up to? – what do they aim to accomplish? In *Deconstructing*, I speculated that, if pressed, most philosophers would take themselves to be engaged in some version of the folk semantics project; they are trying to "capture the details of a commonsense theory about the link between words and the world." It is particularly gratifying that Devitt agrees about this, since while I dabble in the philosophy of language from time to time, Devitt is a pro who knows both the literature and the folks who produce the literature far better than I. Another point on which Devitt and I agree is that if a theory of reference is pursuing the folk semantics project, and if the analogy with folk physics is the right one, then there is no reason to think that the results of the project will give us the right account of reference. So while it might be of psychological interest, the project is of no help in evaluating the eliminativist argument. What about the other analogy, the one that takes its inspiration from the Chomskian account of grammar. What's Devitt's view on that option? Here, I confess, I am not entirely sure. In recent years, Devitt has been a trenchant critic of the Chomskian account of grammar, and has argued that "the idea that the grammar is represented in the minds of speakers is implausible and unsupported."[6] This is not the place to debate the merits of Devitt's challenging and controversial critique of Chomsky. But even if he is right, it is not clear that it would tell us much about the version of the folk semantics project that opts for the Chomskian analogy. For even if there is no *grammar* mentally represented in the minds of speakers, it might well be the case that speakers have a tacit folk theory of reference. People do, after all, have reference intuitions of the sort that Kripke, Putnam and many other philosophers explore and exploit. And, as Frank Jackson often says about intuitions of this sort, surely they are not random. So something is guiding people's judgments when they offer intuitions about reference, and a tacit theory is certainly a prime candidate. Does Devitt disagree? I don't know the answer to that.

The idea that there is a tacit theory of reference subserving reference intuitions is common ground between the linguistics analogy and the folk physics analogy. What is distinctive about the linguistics analogy is the contention that a person's tacit theory of reference can't be wrong because the word-world relation specified by that tacit theory is constitutive of reference for that person. This, I'm sure, is an idea that Devitt would reject, since it leads pretty directly to the claim that knowledge of some semantic facts is *a priori*, and Devitt thinks that no knowledge is *a priori*. But I am not much inclined to follow him here. For it seems entirely possible that an internally represented theory or an internally represented body of rules could sustain a sort of *a priori* knowledge. Suppose, for example, that I invent a variant on the standard game of chess by adding a few new rules. I dub the new game 'Stich-chess' and I memorize the rules. One of the new rules specifies a new way of check-mating your opponent: if a capture leaves your opponent with only 3 pieces on the board, then he has been check-mated. It seems entirely plausible to say that the new rules are partly constitutive of what it is to check-mate someone in Stich-chess, and that I can know *a priori* that a move which leaves an opponent with only 3 pieces on the board is a check-mate in this game. If this story is coherent, and surely it is, then I can't see why a theorist could not adopt a similar view

[6] Devitt's arguments for this highly contrarian claim is set out at length in Devitt (2006). And he calls me an *enfant terrible*! The quote in the text is from section 2 of Devitt's paper in this volume.

about the way in which an internalized theory of reference leads to *a priori* knowledge about reference.

In *Deconstructing*, I did not argue that there was anything incoherent or implausible about the folk semantic project when paired with the linguistics analogy. What I did argue is that if that's what the theory of reference is up to, then the results would be of no help in assessing the soundness of the eliminativists' argument, or in resolving other sorts of ontological issues. The argument I offered was long, complicated and, many readers have told me, more than a bit obscure. A few years after the book was published, Mike Bishop and I came up with a much shorter and more elegant way of making the point. (Bishop & Stich, 1998) To use a theory of reference and the strategy of semantic ascent to draw ontological conclusions, one must assume that the word-world relation, R, that the theory characterizes satisfies a "disquotation" principle like one of the following:

(1) (x) Fx iff 'F_' stands in the R relation to x.

(2) ¬(∃x) 'F_' stands in the R relation to x → Fs do not exist.[7]

But that is not an assumption that one gets for free. There are, after all, endlessly many word-world relations that *don't* satisfy these disquotation principles. So some *argument* is needed to justify the claim that the relationship defined by an internalized folk semantics does satisfy them. I have no idea how to produce that argument, and as far as I know, no one else does either. And without such an argument, the folk semantics project is of no help at all in determining whether the eliminativists' argument is sound.

Let's turn now to the other family of answers to the question: "What is a theory of reference trying to accomplish?" – the family that invokes some version of what I've called the "proto-science" project. Here the central idea is that just as we look to physics to tell us what matter and energy are, and we look to biology to tell us what genes are, perhaps we can look to science to tell us what reference is. The problem, of course, is that there is no science to turn to. No empirical inquiry has offered us anything like the sort of account of reference that would be of use in assessing the eliminativist argument. However, taking a cue from Cummins (1989), we might try to get some indirect help from the sciences. While no existing science tells us what reference is, some branches of science seem to invoke reference, or at least some reference-like word-world relation, in setting out their explanations and theories. Linguistics is the most obvious candidate, though cognitive psychology, anthropology or even history might be thought to make use of such a relation. If they do, then we can try to do for those sciences what Cummins, Egan, Ramsey and others have tried to do for *representation* – the head-world relation invoked in various branches of cognitive science.[8] That is, we can try to describe in

[7] (1) and (2) should be viewed as schemas which can be turned into sentences by substituting a suitable predicate for 'F'. A relation, R, satisfies a schema if each substitution instance is true.

[8] For more on the work of Cummins, Egan and Ramsey, see my response to Egan in this volume.

detail the sort of relation that is presupposed by these sciences, to say what features that relation must have for the theories and explanations that invoke the relation to be convincing or at least plausible. In *Deconstructing*, I expressed skepticism "that any of these areas of inquiry make genuinely explanatory use of a reference-like word-world relationship." (44) But in light of Devitt's comments, I'm inclined to back off on this point. While Devitt has not convinced me that linguistic semantics invokes a substantive (as opposed to a deflationary) notion of reference, he has convinced me that the issues involved are contested and complex. A second point I made in my brusque dismissal of the proto-science project in *Deconstructing* was one suggested by Cummins' work on representation. It is naïve, Cummins insists, to suppose that each branch of cognitive science invokes the same notion of representation. Indeed, if Cummins is right, it is worse than naïve, it is false. Similarly, if linguistics, cognitive science, anthropology and history all invoke a reference-like word-world relation, it would be naïve to assume that they all invoke the same relation. And if, as the proto-scientific project progresses, it turns out that different relations are in play in these different disciplines, then it looks like the philosopher who wants to use the results of the proto-scientific project to assess the eliminativists' argument is beset with an embarrassment of riches. For if there are different reference-like word-world relations used in different sciences, which one should the philosopher rely on in assessing the tacit third premise of the eliminativists' argument? Without some well motivated answer to this question, the proto-science project may turn out to be of no help at all.

Devitt proposes a different way of understanding the proto-science project, a way which, he thinks, sidesteps this problem. Before turning to Devitt's version of the proto-science project, I want to note an important point that I missed in *Deconstructing*. Suppose that the embarrassment of riches problem can be addressed in some principled way; suppose, for example, that Devitt is right and that only linguistic semantics invokes a reference-like word-world relation. So there is only one candidate for a Cummins-style proto-scientific explication. That by itself would not show that the word-world relation invoked by linguistic semantics is the one to be used in assessing the plausibility of the eliminativists' arguments and other, similar, ontological arguments. For as Bishop and I argued, to be useful in ontological arguments, a word-world relation must satisfy disquotation principles like (1) and (2). And neither the fact that a word-world relation is the one exploited in linguistic semantics, nor the fact that researchers in that area call the relation "reference," provides any obvious reason to suppose that the relation satisfies the disquotation principles. Here, just as with the word-world relation favored by folk semantics, *we need an argument*. And as far as I can see, no one has the faintest idea how to provide that argument.

In introducing his version of the proto-science project, Devitt begins with what he thinks is an obvious answer to the question that has been center-stage in our exchange: "What makes a theory of reference true or false? Well, *the nature of the reference relation* does. What is a theory of reference supposed to do? Well, *characterize that nature*." (ms 3) As he notes a bit later, however, "the simple and obvious claim that the task of the theory of reference is to characterize reference strikes people as naïve because there is thought to be a special problem about *identifying* reference. *Which* relation is it the task to characterize?"

(ms 5) Devitt thinks there is no problem in specifying which relation to characterize, or at least there is no special problem here. All we need do is identify some obvious and uncontroversial examples. And to do that, it is perfectly OK to rely on our intuition, with the clear understanding that later theorizing can lead us to reclassify some of the examples we thought were intuitively obvious. Here's how Devitt develops the idea:

> The first stage of a theory of any property *F* or relation *R* involves identifying some apparently uncontroversial examples where *F* or *R* is instantiated and some apparently uncontroversial examples where it is not instantiated. These examples can then be examined in the second stage to discover what is common and peculiar to *F* or *R* in the hope of determining its nature. In that first stage we should consult those most expert at identifying cases of *F* or *R*. If we are concerned with, say, *being a gene*, *being an echidna*, or *being an isotope of*, then we can look to scientists for the identification. But when our concern is with *being referred to by*, it is doubtful that anyone is more expert at identification than the folk. So these most basic folk intuitions about reference, intuitions that identify paradigm cases of it, play a role in the first stage of the theory of reference. …. Even these basic intuitions are not infallible. Theorizing at the second stage can lead to the rejection of results at the first stage: apparently uncontroversial examples turn out to be controversial; whales are not fish after all; tomatoes are not vegetables; unacceptable strings of words turn out to be grammatical. There is even less reason to think that any richer folk intuitions or theories about the nature of reference *must* be true. (ms 4-5)

So there is, Devitt thinks, no need worry about *which* relation a theory of reference is trying to characterize, and no need to be concerned that different branches of science might invoke different word-world relations. The sciences have no role to play in the first stage of the proto-science project, as Devitt construes it, though they do come in at the third stage where the task is to show "that the word-world relation that is reference is scientifically useful." (ms 5)

Is this account of what a theory of reference is trying to do naïve? Far from it. Rather, I think, it is straightforward, sensible and initially quite appealing. But I also think it is seriously under-described, and that when we start asking about some of the missing details much of the appeal of the idea dissolves. Let's begin by focusing on the "apparently uncontroversial examples" that we identify by relying on "intuitions that identify paradigm cases." Which intuitions are these? Devitt mentions the relation between 'Jemima' and Jemima and between 'cat' and cats as paradigm cases of the reference relation, and the relation between 'Jemima' and Fido and between 'cat' and dogs as clear cases of non-reference. So presumably the intuitions that give rise to these judgments are among the "basic folk intuitions" with which the project begins. But what about the sorts of intuitions that Kripke and Putnam relied on in their critique of description theories of reference – intuitions about Gödel and Schmitt or about Twin Earth. Do they count as "basic folk intuitions"? There is good reason for Devitt to say *yes* here. For it was intuitions like these that enabled philosophers to tease apart the description theory and the causal-historical theory – to see clearly that there were cases where the theories had quite different implications – and to argue that description theories were problematic for a range of cases.

So if intuitions like these are excluded in the first stage of Devitt's project, then there is a real risk that the second stage of the project will not be feasible. The job of that second stage, recall, is to "discover what is common and peculiar" to all the examples picked out in the first stage. But if in the first stage we are restricted to intuitions like those about the relation between 'cat' and cats (in the actual world but not, for example, in worlds where there are animals just like cats with very different evolutionary histories) then we are likely to find that there are *lots* of relations that all the examples instantiate. For any finite set of examples, there will, of course, be an infinite number of relations that they instantiate. I assume that Devitt would appeal to methodological criteria – criteria which are hard to state but easier to use – in order to rule out most of these. However, those criteria surely will not rule out either relations characterized by standard versions of the description theory or relations characterized by standard versions of the causal-historical theory. We'll need intuitions like those about Gödel and Schmitt and those about Twin Earth to rule out these alternatives. Without them, the second stage simply isn't viable.

So it looks like Devitt will have to say that Gödel / Schmitt intuitions and Twin Earth intuitions will count as "basic folk intuitions" in the first stage of his project.[9] That, however, poses problems of a different sort. For there is now good reason to think that, despite the important role they played in the emergence of the causal-historical theory, the folk are hardly in agreement on these "basic folk intuitions." Rather, it seems, the intuitions are culturally local. Machery et al. (2004) gave a pair of vignettes closely modeled on Kripke's Gödel / Schmitt case to two groups of fluent English speakers. One group were Americans whose cultural background was European, the other group were Hong Kong Chinese at the University of Hong Kong. The majority of Americans in this study, like the overwhelming majority of analytic philosophers (whose cultural background is also European!), had intuitions compatible with the causal-historical theory and incompatible with the description theory. The majority of Chinese, however, had intuitions compatible with the description theory and incompatible with the causal-historical theory. This is, to be sure, only a single study and much more work remains to be done. But if these results are robust and generalizable, then Devitt's project is in trouble. Without Gödel / Schmitt intuitions and their ilk, he won't have enough cases to pick out a unique relation whose nature will be studied in the second stage of the project. But if he wants to include Gödel / Schmitt intuitions, he'll have to offer some principled reason for excluding the intuitions offered by one cultural group or the other. And it is, to put it mildly, less than clear what that reason might be.

A second way in which Devitt's project is under-described emerges when we ask what the theorist is supposed to do if it turns out that the "basic folk intuitions" pick out quite different word-world relations for different sorts of terms. This is hardly an unexpected outcome. Indeed, Devitt himself conjectures that description theories will turn out to be correct for some words (including 'bachelor' and 'vixen' [ms 7]) and he has

[9] There is also some textual evidence that suggests this is what Devitt would say. In his book, *Coming to Our Senses*, Devitt describes "a methodology for descriptive tasks in general" in which we are encouraged to attend to what the appropriate experts would say when presented with "descriptions of phenomena" set out in "thought experiments." (Devitt 1996, pp. 72-76) The method that Devitt sketches in his paper in this volume is clearly intended to be a special case of that general methodology.

argued at length that causal-historical theories are correct for many other terms (Devitt, 1981). At first blush, it might seem that there is no real problem here. If reference turns out to be two relations (or three or four), so be it. Isn't that entirely analogous to what happened in the case of the that old philosophical chestnut, jade? In the first stage of the inquiry into the nature of jade, the experts (jewelers and gem cutters, I suppose) got to pick out paradigm cases of jade. But when scientists inquired into the nature of the stones picked out, it turned out that there were two quite different minerals in the sample: jadeite and nephrite. So there are two kinds of jade, but this creates no particular problem. If it turns out that there are two (or more) kinds of reference, the analogy suggests that this will be equally unproblematic. The analogy breaks down, however, when we ask how the theorist is supposed to deal with *contested* examples. If the jewelers and gem cutters disagree about a particular stone, we can give it to the relevant scientists and let them determine what it is. If it's jadeite or nephrite then it's jade. If it is some other mineral, it's not jade. But now what are we supposed to do about contested cases of reference? Some people think that many terms used in past theories do refer, and that "a causal theory (or perhaps a part-causal part-description theory) of reference is often more plausible for the words of past theories." (ms. p. 2). Others think that some version of the description theory is the right one for these terms, and that the terms do not refer at all. Given the controversy, obviously intuitions about the reference of these terms can't be used to pick out "paradigm cases." So presumably we'll have to let the theory decide, just as we did in the case of contested cases of jade. But how, exactly, is the theory supposed to help us? In these cases, the term in question typically *does* have the sort of causal (or causal-historical) links to things that really exist that were found in uncontested cases. Historical tokens of 'phlogiston', as Philip Kitcher points out, had a rich web of causal relations with oxygen (Kitcher, 1993). However, it is also the case that these terms are associated with descriptions of the sort that play a central role in fixing reference in other uncontested cases (cases like Devitt's 'bachelor' and 'vixen' perhaps). But those descriptions aren't satisfied (or even approximately satisfied) by anything. So if the causal or causal-historical relation is the one which determines reference *for these terms* – if that's the kind of reference they have, so to speak – then they do refer. If the description-theoretic relation is the kind of reference they have, then they don't refer. But which kind of reference *do* they have? As far as I can see, Devitt's account doesn't even begin to address that question. And without an answer, the project Devitt describes is of no help at all in characterizing "the nature of the reference relation" in just those cases where such a characterization was thought to be philosophically important.

As we saw earlier, Devitt agrees that most people actually engaged in debates over which theory of reference is the right one seem to be doing folk semantics. The 3-stage project he describes is not the one he thinks these people are engaged in, but the one he thinks they should be engaged in. It's the one that will tell us "the nature of the reference relation". Here we disagree sharply. I don't think the project he has described has any serious prospect of telling us "the nature of the reference relation" and I doubt that there is any way of modifying the account that will enable it to do that . Moreover, I'm inclined to agree with deflationists like Field (1986, 1994) and Horwich (1990) who maintain that the quest for the nature of the reference relation is misguided, since reference is not a substantive relation at all. Obviously this is not the place to offer a defense of that view and,

truth be told, I have little to add beyond what Horwich and Field have already said. While Devitt and I have very different views about reference, we agree that using the strategy of semantic ascent and then appealing to a theory of reference is not a good way to make progress in the eliminativism debate or in debates about other ontological issues. Here we both disagree with Frank Jackson. So it is time to take a look at Jackson's view.

Jackson

Jackson endorses my account of how the eliminativism debate became "embroiled" in the theory of reference. (I wish I had thought of using that word!) He also endorses the conclusion I once wanted to draw:

> [I]t matters for the truth of eliminativism *which* theory of reference is correct. Ergo, if that question has no one 'correct' answer, there is no one 'correct' answer to whether or not eliminativism is true or false. (ms 6)

Moreover, as Jackson was instrumental in persuading me years ago, the problem is not restricted to eliminativism: "if [the] argument works for the propositional attitudes, it works for the big bang, cures for AIDS, Venus, and so on." (ms 1) For the reasons I was groping toward in *Deconstructing*, and managed to express more clearly (I hope) in Bishop & Stich (1998), I am no longer inclined to think that the truth of eliminativism, or of any other ontological claim is linked in any interesting way to the quest for the correct theory of reference. But here Jackson remains unconvinced. He still maintains that the two projects are importantly connected.

Jackson thinks that "that the best policy is to be easy-going about what counts as representation," (ms 9) and he councils the same attitude toward debates about which theory of reference is correct, as these debates have been pursued in much of the philosophical literature. "Be the theory of reference proto-science, or the theory that best captures reflective intuitions, or a bit of both, there is no reason to hold that there is one theory of reference." (ms 11) But this casual attitude does not extend to the notion of reference used in debates over ontology. "Perhaps … there are many notions of reference, but there had better be a preferred notion in play for the purposes of raising ontological questions." (ms 13) And in the last third of his paper, Jackson tells us how to find that theory. He doesn't actually offer a theory, since "that's a job for a book;" what he does offer is "a criterion for choosing among a range of candidate word-world relations, the one which is of interest when we do ontology." (ms 16)

The picture of language that is central to Jackson's story is inspired by Lewis, Grice and ultimately by John Locke. And the version that Jackson sketches is, as one would expect, subtle, sophisticated and richly informed by the ongoing philosophical literature on these matters. The "fundamental point" is that "language is a learnt, convention-based system of representation" and that "our possession of language is how we are able to make public and communicate the contents of our mental states." (ms 17) "Very roughly," Jackson tells us, "we *start* with mental states that represent how things are, most especially beliefs and thoughts that things are thus and so, and somehow hit on

an implicit understanding that certain noises and marks are to be given the task of making public these representational contents."(ms 17, italics added) This requires a coding system, and "a term's reference is a key part of this coding system." (ms 18) Jackson then sketches how we can use facts about this coding system to zero in on the right reference relation. To put the matter very crudely indeed, the right reference relation is the one that will enable the coding system to do its job.

There is much to admire in Jackson's account, and much to debate. But, as I see it, even if Jackson is right, his account will not address the problem that motivated the project in the first place, for his account does not help at all in trying to decide whether the eliminativists' ontological conclusion follows from their two explicit premises. Rather, the bump in the rug just moves to another place. Perhaps the easiest way to see the point is to note that if one accepts Jackson's general view of the relation between language and thought, then there is no need to say or write anything at all in posing the eliminativists' argument. One could simply run through the argument in thought – *in foro interno* as Ernie Sosa might say. I am painfully aware that this is possible, since I have done it myself many, many times. *Let me suppose that beliefs and desires are the posits of a tacit, folk-psychological theory,* I think to myself, *and that the theory is seriously mistaken for the following reasons …. Would it follow that beliefs do not exist? Or should I conclude that beliefs do exist, but that folk psychology is wrong about many of the claims it makes about them?* Having a detailed account of the word-world relation picked out by Jackson's criterion would do me no good at all in attempting to answer this question, since that account tells me about words in a public language – "a learnt, convention-based system of representation" that we use "to make public and communicate the contents of our mental states." It tells me what "certain shapes on paper or sound wave patterns in the air"(ms 12) refer to. But I haven't scribbled any shapes on paper or emitted any sound wave patterns. I'm sitting quietly, thinking about eliminativism.

At this point I can imagine someone protesting that I'm being unfair to Jackson. To be sure, I can think about the eliminativists' argument without saying or writing something, the critic concedes, but I am still using words in a public language; I am using them *in thought*. There is no way of thinking the italicized thoughts in the previous paragraph without those words (or the words of some other natural language) running through my mind. So I can use Jackson's account of the reference of public language words to determine the reference of my thoughts about beliefs. And that will enable me to determine which conclusion follows from the premises.

Though the issues raised by this protest are complex and controversial, I am inclined to think the critic is right. Indeed, one can't think *that beliefs and desires are posits of a tacit, folk-psychological theory* unless one thinks it in some natural language or other. But the crucial point here is that while I can, and do, agree with the critic, Jackson *can't*. On the Locke-Grice-Lewis picture of language that Jackson embraces, "we start with mental states [most especially beliefs and thoughts that things are thus and so] that represent how things are."(ms 17) These thoughts "have (reasonably) determinate contents which we can report using sentences", where contents are taken to

be "states of affairs or propositions". (fn. 18) It is the content of our thoughts about beliefs that determines what the natural language term 'belief' refers to. So if Jackson were to appeal to his account of the reference of the term 'belief' to determine the content of thoughts about beliefs, he'd be trying to pull himself up by his own bootstraps. To oversimplify a bit, on Jackson's account, the reference of terms in natural language is determined by the reference thoughts. But if he is right about this, it does not help us at all, since we now need to know what determines the reference of thoughts. And about this, Jackson's theory of reference tells exactly nothing.

Reply to Cowie

In a memorable passage in Quine's *Word and Object* – one of many – Quine reminds us of "the immortal words of Adolf Meyer, where it doesn't itch don't scratch." (Quine, 1960, p. 160) For many years, I succeeded in following Meyer's advice in dealing with the concept of innateness. I was aware, of course, of the difficulties that philosophers, cognitive scientists and biologists were encountering in their efforts to give an account of innateness. And I had considerable sympathy with Paul Griffith's suggestion that the notion of innateness was a muddled bit of folk biology that should be banished from serious science. Indeed, on two occasions I invited Griffiths to make the case in my Rutgers graduate seminar, which he did, with great learning and enthusiasm. At the same time, as Cowie notes, I acted as though I was "quite committed to the concept's remaining part of the discourse of cognitive science" (ms 3) – sponsoring conferences on nativism, editing volumes on *The Innate Mind* (Carruthers, Laurence and Stich, 2005, 2006, 2007), and writing books and papers in which innate mental mechanisms were invoked to explain important features of mindreading (Nichols & Stich, 2003) and moral cognition (Sripada & Stich, 2006). There was clearly a tension there, and now and again I worried about it a bit. But it wasn't until I read Cowie's rich and interesting paper that the tension rose to the level of an itch.

Fortunately, Cowie is a full service philosophical interlocutor; in addition to generating an uncomfortable itch, she also provides a comforting scratch. By examining the recent history of the gene concept and the not so recent history of the concepts invoked in classificatory systems in organic chemistry and its predecessors, she makes a compelling case for the conclusion that "good things can come from bad concepts." (ms 35) She goes on to argue, convincingly I think, that "modern discussions of innate traits look a lot like 18th century discussions of plant materials…. And just as late 18th century chemists knew that their taxonomic principles were wrong ones, yet couldn't burn the boat they were fishing from, we can know that the innate/not innate distinction is not quite the right one to make, yet keep on making it nonetheless." (ms 43) Philosophical prudishness, she urges, should give way to "vulgar pragmatism" since "premature elimination can be disastrous." (ms 44) So, with the itch slaked, I'll go on invoking the notion of innateness since, as Cowie eloquently argues, "you can't investigate something if you don't have a way of thinking about it," and "the concept of innateness enables us to think about developmental phenomena that we don't yet fully comprehend." (ms 44)

The vulgar pragmatist strategy that Cowie advocates can, of course, be used to deflate eliminativist worries focused on other concepts, including those of intentional psychology. On the last page of *Deconstructing the Mind*, I argued that "what 'legitimates' certain properties (or predicates, if you prefer) and makes others scientifically suspect is that the former, but not the latter, are invoked in successful scientific theories." I went on to say that "the jury is still out on the question of whether successful science can be constructed using intentional categories." (Stich, 1996, p. 199) But that was then, and this is now. As I note in my reply to Egan, it now strikes me as undeniable that a number of different branches of psychology are producing exciting and important discoveries that we have no idea how to describe or explain without invoking intentional categories. So, to respond to my former eliminativist self, and to those who are still tempted by views in that vicinity, I am delighted borrow a sentence from Cowie. "It's simply not possible to design, conduct and interpret the very experiments one needs to perform in order to show the way forward without using some concepts – and if the bad old concepts are the only ones available, well: it's put up or shut up." (ms 43)

Reply to Sosa[10]

Sosa's topic is the use of intuitions in philosophy. Much of what I have written on the issue has been critical of appeals to intuition in epistemology, though in recent years I have become increasingly skeptical of the use of intuitions in ethics and in semantic theory as well.

In the first half of his paper, Sosa discusses my critique of *analytic epistemology*, which I use as a technical term for epistemological projects in which conceptual or linguistic analysis is taken to be the ultimate court of appeal for many disputes in epistemology, and intuitions are used to support or challenge the conceptual analysis. I maintain that projects of this sort are widespread in philosophy, and in several publications I have cited parts of Alvin Goldman's *Epistemology and Cognition* (1986) as an important example of the sort project I have in mind. Here is the relevant passage from Stich (1988) – the paper on which Sosa focuses.

> Goldman notes that one of the major projects of both classical and contemporary epistemology has been to develop a theory of epistemic justification. The ultimate job of such a theory is to say which cognitive states are epistemically justified and which are not. Thus, a fundamental step in constructing a theory of justification will be to articulate a system of rules evaluating the justificatory status of beliefs and other cognitive states. These rules (Goldman calls them *justificational rules* or *J-rules*) will specify permissible ways in which a cognitive agent may go about the business of forming or updating his cognitive states. They "permit or prohibit beliefs, directly or indirectly, as a function of some states, relations, or processes of the cognizer" (Goldman, 1986, p. 60).

[10] I'm grateful to Michael Bishop, Edouard Machery and Jonathan Weinberg for helpful comments on earlier drafts of this reply.

> Of course, different theorists may have different views on which beliefs are justified or which cognitive processes yield justified beliefs, and thus they may urge different and incompatible sets of J-rules. It may be that there is more than one right system of justification rules, but it is surely not the case that all systems are correct. So in order to decide whether a proposed system of J-rules is right, we must appeal to a higher criterion, which Goldman calls a "criterion of rightness." This criterion will specify a "set of conditions that are necessary and sufficient for a set of J-rules to be right." (Goldman, 1986, p. 64)
>
> But now the theoretical disputes emerge at a higher level, for different theorists have suggested very different criteria of rightness…. How are we to go about deciding among these various criteria of rightness? Or, to ask an even more basic question, just what does the correctness of a criterion of rightness come to; what makes a criterion right or wrong? On this point Goldman is not as explicit as one might wish. However, much of what he says suggests that, on his view, *conceptual analysis* or *conceptual explication* is the proper way to decide among competing criteria of rightness. The correct criterion of rightness is the one that comports with the conception of justifiedness that is "embraced by everyday thought or language." (Goldman, 1986, p. 58) To test a criterion we explore the judgments it would entail about specific cases, and we test these against our "pretheoretic intuition." "A criterion is supported to the extent that implied judgments accord with such intuitions, and weakened to the extent that they do not" (Goldman, 1986, p. 66). Goldman is careful to note that there may be a certain amount of vagueness in our commonsense notion of justifiedness, and thus there may be no unique best criterion of rightness. But despite the vagueness, "there seems to be a common core idea of justifiedness" embedded in everyday thought and language, and it is this common core idea that Goldman tells us he is trying to capture in his own epistemological theorizing (Goldman, 1986, pp. 58-59)
>
> …. I propose to use the term *analytic epistemology* to denote any epistemological project that takes the choice between competing justificational rules or competing criteria of rightness to turn on conceptual or linguistic analysis. (Stich, 1988, pp. 105-6)

I've taken a dim view of projects like this, arguing that they lead to an unwelcome xenophobia in epistemology.

> [T]he analytic epistemologist's effort is designed to determine whether our cognitive states and processes accord with our commonsense notion of justification (or some other commonsense concept of epistemic evaluation). Yet surely the evaluative epistemic concepts embedded in everyday thought and language are every bit as likely as the cognitive processes they evaluate to be culturally acquired and to vary from culture to culture. Moreover, the analytic epistemologist offers us no reason whatever to think that the notions of evaluation

> prevailing in our own language and culture are any better than the alternative evaluative notions that might or do prevail in other cultures. But in the absence of any reason to think that the locally prevailing notions of epistemic evaluation are superior to the alternatives, why should we care one whit whether the cognitive processes we use are sanctioned by those evaluative concepts? How can the fact that our cognitive processes are approved by the evaluative notions embraced in our culture alleviate the worry that our cognitive processes are no better than those of exotic folk, if we have no reason to believe that our evaluative notions are any better than alternative evaluative notions. …. [I]t's my contention that this project is of no help whatever in confronting the problem of cognitive diversity unless one is an epistemic xenophobe. (Stich, 1988, pp. 107 & 109)

It is clear that Sosa is not happy with this line of thought, though I confess that I am much less clear about what, exactly, his objection is. One might be concerned that analytic epistemology, as I have characterized it, is a straw man – that no one really does take conceptual analysis, supported by appeal to intuition, to be the final court of appeal in many epistemological disputes. But I doubt this is Sosa's view, since there is nothing in his paper that challenges my interpretation of Goldman or my contention that Nelson Goodman and Peter Strawson can also plausibly be read as practitioners of this sort of analytic epistemology.[11] Moreover, the view that many epistemologists proceed in this way is hardly idiosyncratic. A number of other authors have also commented on the pivotal role of appeals to intuition-based conceptual analysis in justifying epistemological theses.[12] Rather, my best guess is that Sosa believes that epistemologists should not try to invoke conceptual or linguistic analysis as the final arbitrator in resolving disputes about competing sets of justification rules, competing criteria of rightness and the like, because to do so they would have to endorse something like the reasoning set out in Sosa's (a) – (e), which "requires controversial claims and assumptions." (ms 5) I am far from confident of this reading, however, because Sosa does not make clear how much of (a) – (e) is offered as an explication of the reasoning required to sustain the analytic epistemologist's appeal to conceptual analysis, and how much (if any) needs to be assumed only by someone who wants to criticize analytic epistemology along the lines sketched in the previous paragraph. Also, I am far from convinced that either the analytic epistemologist or the critic would have to appeal to "some such reasoning". The argument in (a) – (e) goes far beyond anything I or the philosophers I criticize have endorsed, and Sosa offers no reason to suppose that their project or my critique can *only* be elaborated by invoking this problematic argument.

I think these concerns can safely be set to one side, however, since, while our reasons may be quite different, Sosa and I agree that the strategy of resolving debates about justificational rules, criteria or rightness and the like by appealing to conceptual analyses supported by intuition is deeply problematic. Sosa's main motive in criticizing that strategy, I think, is to set it aside so that we can focus on another way in which

[11] For some evidence in support of these interpretations, see Stich (1988), pp. 97-99 (for Goodman), and fn. 15 (for Strawson).

[12] See, for example, Cummins (1989) and Kornblith (2002), ch. 5.

intuitions are used in epistemology (and in other parts of philosophy) which makes no appeal to conceptual analysis.[13] Philosophers can and do rely on "intuition as a source of data for philosophical reflection" (ms 9) without any attempt to vindicate the practice by appealing to the analysis of meanings or concepts. "At least since Plato," Sosa notes,

> philosophical analysis has relied on thought experiments as a way to test hypotheses about the nature and conditions of human knowledge, and other rational desiderata, such as justice, happiness, and the rest.
>
> Any such practice gives prime importance to intuitions concerning not only hypothetical cases but also principles in their own right. The objective is to make coherent sense of the contents that we intuit, by adopting general accounts that will best comport with those intuitions and explain their truth. (ms 6-7)

I think Sosa is clearly right that the practice he describes is widespread in philosophy, and has been since antiquity. He's right, too, in insisting that many who pursue philosophy in this way do not take themselves to be engaged in conceptual or linguistic analysis. Where Sosa and I differ is that he thinks this is an entirely reasonable way for philosophers to proceed, while I think it is a method that philosophers should abandon.

One of the arguments that I have used against this way of doing philosophy begins with the contention that philosophical intuitions certainly could and probably do differ in different cultural groups. The "certainly could" part of this – the claim that cross-cultural differences in philosophically important intuitions is logically possible – is often conceded and then quickly dismissed as a way of discrediting intuition-based philosophy. Granted, the critics argue, it is logically possible that people with different cultural backgrounds have quite different philosophical intuitions; but it is also logically possible that people with different cultural backgrounds have quite different perceptual experiences even in identical environments. So the possibility argument gives us no more reason to be skeptical of intuition-based philosophy than to be skeptical of beliefs based on perception. At best, the critics continue, the possibility argument is just a special case of a quite general argument for skepticism.[14] I am inclined to think that the critics are right about this. The claim that people in different cultural groups *probably do* have significantly different intuitions about philosophically important matters, and the empirical studies offered in support of this claim, have provoked much more elaborate responses from the defenders of intuition-based philosophy. Sosa's responses, in his paper in this volume and in several other recent papers, are among the best informed and most acute of these.

---

[13] In another paper on the use of intuitions in philosophy Sosa writes: "It is often claimed that analytic philosophy appeals to armchair intuitions in the service of 'conceptual analysis'. But this is deplorably misleading. The use of intuitions in philosophy should not be tied exclusively to conceptual analysis." (Sosa 2007a, p. 100) Sosa makes an almost identical comment in Sosa (forthcoming), Sec. V.

[14] For a recent argument in this vicinity, see Pust (2000) and Williamson (2007).

Though Sosa and I disagree sharply about the role that appeal to intuition should play in philosophy, there are many points on which I think we are in complete agreement. First, and most important, Sosa and I both think that *if* it is true that people in different cultural groups disagree in their intuitive judgments about philosophically important cases, this would pose a major problem for philosophers who use intuition as a source of data in the way that Sosa recounts. In the paper in this volume, Sosa hedges his bets a bit, saying only that this sort of disagreement would "allegedly pose a 'serious problem'" (ms 11). But in other papers (published earlier, though written later), he is less guarded:

> One main objection [posed by "those who reject philosophical intuition as useless"] derives from alleged disagreements in philosophical intuitions, ones due in large measure to cultural or socioeconomic or other situational differences. This sort of objection is particularly important and persuasive …. (Sosa forthcoming, ms. pp. 9-10)

> [T]here will definitely be a prima facie problem for the appeal to intuitions in philosophy if surveys show that there is extensive enough disagreement on the subject matter supposedly open to intuitive access. (Sosa 2007a, p.102)

In one of these papers, Sosa goes into some detail on why disagreement in intuition would be problematic.

> When we rely on intuitions in philosophy, then, in my view we manifest a competence that enables us to get it right on a certain subject matter, by basing our beliefs on the sheer understanding of their contents. How might survey results create a problem for us? Suppose a subgroup clashes with another on some supposed truth, and suppose they all ostensibly affirm as they do based on the sheer understanding of the content affirmed. We then have a prima facie problem. Suppose half of them affirm <p> while half deny it, with everyone basing their respective attitudes on the sheer understanding of the representational content <p>. Obviously, half of them are getting it right, and half wrong. Of those who get it right, now, how plausible can it be that their beliefs constitute or derive from rational intuition, from an attraction to assent that manifests a real competence?
>
> Not that it is logically incoherent to maintain exactly that. But how plausible can it be, absent some theory of error that will explain why so many are going wrong when we are getting it right? Unless we can cite something different in the conditions or in the constitution of the misled, doubt will surely cloud the claim to competence by those who ex hypothesi are getting it right.(Sosa 2007a, p. 102)

I think that this diagnosis of the problem posed by disagreement in intuitions is both accurate and incisive. The only observation I'd add is that if the intuitions being studied are about "the nature and conditions of human knowledge, … justice, happiness, and the rest," and if the two groups are, say, East Asians and Westerners, then producing a

plausible "theory of error" will be, to put it mildly, no easy task.[15] It is worth emphasizing the enormous importance of this point, on which Sosa and I apparently agree. For 2500 years, philosophers have been relying on appeals to intuition. *But the plausibility of this entire tradition rests on an unsubstantiated, and until recently unacknowledged, empirical hypothesis* – the hypothesis that the philosophical intuitions of people in different cultural groups do not disagree. Those philosophers who rely on intuition are betting that the hypothesis is true. If they lose their bet, and if I am right that the prospects are very dim indeed for producing a convincing theory of error which explains why a substantial part of the world's population has false intuitions about knowledge, justice, happiness and the like, then a great deal of what goes on in contemporary philosophy, and a great deal of what has gone on in the past, belongs in the rubbish bin. I think it is too early to say with any assurance who is going to win this bet – though if I were a practitioner of intuition-based philosophy I'd be getting pretty nervous. What is clear is that the stakes are very high, and this underscores the importance of cross-cultural empirical work aimed at studying philosophical intuitions and understanding the psychological mechanisms that give rise to them.

A second point on which Sosa and I are in complete agreement is that currently available evidence does not show "beyond reasonable doubt that there really are philosophically important disagreements [in intuition] rooted in cultural or socio-economic differences."(Sosa 2007a, p. 103) Nor do I have any quarrel with two of the reasons Sosa offers for denying that the experimental results he cites establish disagreement beyond a reasonable doubt. The first of these is that the subjects in the experiments might "import different background beliefs as to the trustworthiness of American corporations or zoos, or different background assumptions about how likely it is that an American who has long owned an American car will continue to own a car…." (ms 15). The second is that the results might have been quite different if subjects had been given a third choice, like "we are not told enough in the description of the example to be able to tell whether the subject knows or only believes." (ms 15-16) Though Sosa very graciously describes the experimental work that my collaborators and I have done on epistemic intuitions as "extensive" (ms 10), the truth is that to date there have been only a handful of rather unsophisticated studies. More and better studies are needed, including experiments that address the concerns that Sosa raises, and a variety of other concerns as well. It is still very early days in the empirical exploration of philosophical intuitions, and no one working in the area would claim that anything has been demonstrated *beyond reasonable doubt*. That's a very high standard to set for empirical work in the social sciences. Nonetheless, I am inclined to think that Sosa should be rather more worried than he appears to be. While new evidence certainly *might* undermine the conclusions about cross-cultural diversity in intuition that my collaborators and I have drawn from existing studies, Sosa has given us no reason to

[15] In Weinberg et al. (2001), we speculate that the differences we find between the intuitions of East Asians and Westerners are linked to systematic differences in cognitive processing that Richard Nisbett and his associates have found between these two groups. In a book reviewing some of that work, Nisbett (2003) makes a persuasive case that while East Asian and Western cognitive processing have different strengths and shortcomings, it is singularly implausible to think that one style of thought is generally superior to or more accurate than the other.

think that it *will*. Until it does, these studies stand as noteworthy straws in the wind, and most of the straws seem to be blowing in the wrong direction for those who champion intuition-based philosophy.

Part of the explanation for Sosa's nonchalance emerges when we turn to his third reason for doubting that the experimental results indicate genuine intuitive disagreement. The results would pose no challenge at all to intuition-based philosophy if the term 'knowledge' picks out somewhat different concepts for the two groups, for then "we fail to have disagreement on the very same proposition." (ms 16) Here again, Sosa is surely right when he says that this *might* be the case, but is there any reason to think that it really is? Though Sosa does not address the question directly, some of his remarks suggest that the smart money should bet on ambiguity, because covert ambiguity of the sort he is concerned about is very easy to generate. If East Asians are more sensitive to communitarian factors in deciding whether to apply the term 'knowledge' to particular cases, while Westerners are more sensitive to individualistic factors, that by itself, Sosa seems to suggests, might be enough to show that the term 'knowledge' picks out different concepts in the two groups. But if this is what Sosa thinks, it is far from clear that he is right. There is a vast literature on concepts in philosophy and in psychology (Margolis & Laurence, 1999; Murphy, 2002; Machery, forthcoming), and the question of how to individuate concepts is one of the most hotly debated issues in that literature. While it is widely agreed that for two concept tokens to be of the same type they must have the same content, there is a wide diversity of views on what is required for this condition to be met. On some theories, the sort of covert ambiguity that Sosa is betting on can be expected to be fairly common, while on others covert ambiguity is much harder to generate. For Fodor, for example, the fact that an East Asian pays more attention to communitarian factors while a Westerner emphasizes individualistic factors in applying the term 'knowledge' would be no reason at all to think that the concepts linked to their use of the term 'knowledge' have different contents. (Fodor, 1998) For theorists like Frank Jackson, by contrast, if two people have different intuitions about some Gettier cases, and if neither of them is confused about the details of the example, that's enough to show that they have different concepts. (Jackson 1998, p.32) So on Jackson's account, empirical studies like those that Sosa discusses, no matter how well designed and carefully controlled, could not possibly show that people's intuitions disagree, since prima facie disagreement is conclusive evidence of ambiguity. Since Sosa grants that cross-cultural disagreement on philosophically important intuitions is a genuine empirical possibility, he can't adopt Jackson's account of content. Though I suspect that Sosa favors an account that is more like Jackson's than like Fodor's – one on which covert ambiguity is easy to generate. But since Sosa does not tell us what theory of content he endorses, or why he thinks that the correct theory will make the sort of covert ambiguity that he envisions rather commonplace, there is not much that those of us who are skeptical about intuition-based philosophy can do to move the conversation forward. We can't do empirical studies designed to test for the sort of ambiguity Sosa is worried about, until he tells us more about what that sort of ambiguity *is*. Though I have never been very clear about the rules of burden-of-argument tennis, I am inclined to think that the ball is in his court.

In one of the papers on which Sosa focuses, my collaborators and I argue that even if Jackson is right about concept individuation, findings like ours, which suggest that culture, SES and philosophical training have an important influence on epistemic intuitions, would still pose a serious problem for intuition-based epistemology. For on Jackson's theory (and on others that make it relatively easy to establish the existence of covert ambiguity), it is very likely that the term 'knowledge' picks out lots of different concepts when uttered by members of different groups.

> East Asians, Indians and High SES Westerners all have different concepts; High and Low SES Westerners have different concepts; people who have studied lots of philosophy and people who have studied no philosophy have different concepts. And that, no doubt, is just the tip of the iceberg. Moreover, these concepts don't simply differ in *intension*, they differ in *extension* – they apply to different classes of actual and possible cases. (Nichols, Stich & Weinberg, 2003, p. 245, emphasis in the original)

If that's right, we ask, then how are we to understand traditional views like Plato's claim that "wisdom and knowledge are the highest of human things," or more recent epistemological theories which suggest that if S's belief that p is an instance of knowledge, then, *ceteris paribus*, S ought to believe that p?[16] If 'knowledge' picks out different things for different speakers, they can't all be the highest of human things. And if S's belief that p counts as in instance of knowledge, as that term is used by one speaker, but does not count as an instance of knowledge as the term is used by another speaker, ought S to believe p or not? Similar problems arise in interpreting more recent work, like Williamson's (2000) contention that knowledge is the most general factitive mental state and Hawthorne's claim that "[t]he practice of assertion is constituted by the rule/requirement that one assert something only if one knows it." (Hawthorne, 2004, p. 23). Obviously, if 'knowledge' picks out different things for different speakers, these can't *all* be the most general factitive mental state. And if Ann counts as knowing p, as 'knowing' is used by one speaker, but does not count as knowing p, as 'knowing' is used by another speaker, then what does Hawthorne's claim entail about Ann if she asserts p? Has she violated the rule or hasn't she? Of course, it would be easy enough to answer these questions by simply stipulating that 'knowledge' is to be understood as expressing the concept of some specific group – high SES white Western males with lots of philosophical training, for example. But while that would resolve the ambiguity, it is a move that cries out for some justification. Why *that* group? Why is *their* concept of knowledge better than all the others?

This is a line of argument that Sosa finds baffling. Why do we need to chose between the "commodities" picked out by these various concepts of knowledge, he asks. Why can't we value them all – just as we might value owning river banks and money banks? Sosa doubts there is any conflict between cultural groups which uses the term

[16] As noted in Weinberg et al. (2001) (p. 431 & fn. 5), theories in this vicinity have been suggested by a number of leading epistemologists, including Chisholm (1977), BonJour (1985), and Pollock & Cruz (1999).

'knowledge' to pick out different epistemic commodities. "[T]here seems no more reason to postulate such conflict than there would be when we compare someone who rates cars in respect of how economical they are with someone who rates them in respect of how fast they can go." (ms 21)

While Sosa is baffled by our argument, I am baffled by his bafflement, since the conflict whose existence he denies strikes me as clear and obvious. To make the point quite vividly, an analogy may be helpful. For theorists like Jackson, if two people have divergent intuitive judgments about whether some important cases are instances of X, and if the divergence can't be attributed to mere confusion, then they are invoking different X-concepts. So, as we've seen, if two people have divergent intuitive judgments about Gettier cases, and neither is confused, then Jackson maintains that they are invoking different concepts of knowledge. Jackson makes it clear that he would say the same about cases in the moral domain. If a Yanomamö intuitively judges that it is *morally permissible* to kill men who are not members of his tribe, take their possessions, rape their wives and enslave their children, while I intuitively judge that it is *not morally permissible* to do these things, and if the disagreement can't attributed to confusion, then the Yanomamö and I are invoking different concepts of moral permissibility. And if, as I maintain, this case is entirely parallel to the knowledge case, presumably Sosa would deny that there is any conflict here. He might even wonder why we shouldn't learn to value the "commodities" that the Yanomamö label 'morally permissible' even though they are rather different from the commodities to which we apply the label 'morally permissible'.[17]

Obviously, something has gone very wrong here, though it is no easy matter to diagnose the problem since there are a number of factors involved. One of them is that Sosa has chosen to express his bafflement by focusing on what we *value* in the epistemic domain. Norms of valuing do play a role in traditional epistemological debates, but they are not the only sorts of norms that epistemologists have considered. As we noted earlier, Goldman insists, quite correctly, that justification rules (or "J-rules") play a central role in both classical and contemporary epistemology, and J-rules specify *norms of permissibility* not norms of valuing. They "permit or prohibit beliefs, directly or indirectly, as a function of some states, relations, or processes of the cognizer" (Goldman, 1986, p. 60). When we focus on these rules, the sort of pluralism that Sosa suggests is much harder to sustain. If a rule, like the one cited a few paragraphs back, says that *ceteris paribus* we ought to hold a belief if it is an instance of knowledge, and if 'knowledge' is interpreted in different ways by members of different groups, then Sosa's pluralism leads to inconsistency. There will be some beliefs which we ought to believe on one interpretation of 'knowledge' but not on the other. Moreover, even in the case of norms of valuing Sosa's pluralism can lead to problems. Sosa is surely right to claim that someone who values owning money banks can also value owning river banks. But if

[17] Compare: "The fact that we value one commodity, called 'knowledge' or 'justification' among us, is no obstacle to our *also* valuing a *different* commodity, valued by some other community under that same label. And it is also compatible with our *learning* to value that second commodity once we are brought to understand it, even if we previously had no opinion on the matter." (ms 18-19)

there is one of each on offer and the person's resources are limited, she will have to make a choice. Which one does she value more? Similar quandaries may confront the person who values both of the "commodities" picked out by 'knowledge' by the intuitions of two different groups. There will be occasions when she can have one or the other, but not both. So she must decide which she values more.

Sosa gives us no guidance on how to go about making these choices, and I am inclined to think that this is not simply an oversight. Rather, I suspect, Sosa does not take these questions to be part of the purview of epistemology. One of the reasons Sosa and I find each other's arguments baffling is that we have very different views on what epistemology should be doing. For Sosa, epistemology is "a discipline … whose scope is the nature, conditions, and extent of knowledge" (ms 20). For me, by contrast, epistemology is a discipline that "focuses on the evaluation of methods of inquiry. It tries to say which ways of going about the quest for knowledge – which ways of building and rebuilding one's doxastic house – are the good ones, which are the bad ones and why." The quote is from the first page of *The Fragmentation of Reason*. In the paragraph that follows, I try to make the case that this conception of epistemology is widely shared.

> There is no shortage of historical figures who have pursued this sort of epistemological investigation. Much of Francis Bacon's epistemological writing is devoted to the project of evaluating and criticizing strategies of inquiry, as is a good deal of Descartes's. Among more modern epistemological writers, those like Mill, Carnap, and Popper, who are concerned with the logic and methodology of science, have tended to emphasize this aspect of epistemological theory. From Bacon's time to Popper's, it has frequently been the case that those who work in this branch of epistemology are motivated, at least in part, by very practical concerns. They are convinced that defective reasoning and bad strategies of inquiry are widespread, and that these cognitive shortcomings are the cause of much mischief and misery. By developing their accounts of good reasoning and proper strategies of inquiry, and by explaining why these are better than the alternatives, they hope others will come to see the error of their cognitive ways. And, indeed, many of these philosophers have had a noticeable impact on the thinking of their contemporaries. (*Fragmentation*, pp. 1-2.)

If I understand him correctly, the sort of epistemology Sosa favors is not in this line of work. It does not even try to offer advice on how one should go about revising one's beliefs – certainly not advice "all things considered". (Sosa, this volume, ms. p. 21) Rather, the epistemologist's aim is to characterize the phenomena picked out by terms like 'knowledge' and 'justification' as he uses the term. If it turns out that people in other groups use these terms to pick out different phenomena, Sosa's epistemologist might try to characterize those phenomena as well. But it is not the epistemologist's job to tell us which of these phenomena is better, or which we ought to pursue, or why. One could, of course, engage in an entirely parallel project in the moral domain. A philosopher pursuing that project would try to characterize the phenomena picked out by terms like 'morally permissible' and 'morally prohibited', as she uses the terms. And if people in other groups use the terms in to pick out different phenomena she might try to

characterize those as well. But this Sosa-style moral philosopher would not tell us which characterization of the morally permissible was better, or which actions we should pursue, or which actions we should avoid. This is an interesting project, to be sure, and a valuable one. But by my lights, it is closer to ethnography than to moral philosophy. Much the same, I think, is true of Sosa-style epistemology. As Nichols, Weinberg and I noted, epistemologists who rely on intuitions "have chosen to be ethnographers; what they are doing is *ethno-epistemology*." (Nichols et al., 2003, p. 235; emphasis in the original) Moreover, if these philosophers *are* doing ethnography, then as Weinberg and I have pointed out, their methodology leaves much to be desired. (Stich & Weinberg, 2001) That is a theme I'll take up in my reply to Bishop.

Reply to Bishop[18]

Bishop's rich and challenging paper begins with a detailed, largely sympathetic and exceptionally well informed overview of my work in epistemology. It ends with a section arguing that a modified version of Bishop and Trout's "Strategic Reliabilism" is pragmatically preferable to the sort of epistemic pragmatism that I defended in *Fragmentation*. Though I have a quibble here and there, I think Bishop's account of my work in epistemology is generally accurate. Indeed, on a number of topics he's understood what I was up to better than I did. In the first section of my reply, I'll elaborate on some of the points that he makes and put a rather different spin on them in a few places. In the second section, I'll focus on his suggestion that there is a tension between my attitude toward the use of intuitions in epistemology in *Fragmentation* and in the papers written with Weinberg and Nichols. In the final section, I'll explain why I'm not convinced by his pragmatic defense of Strategic Reliabilism.

1. Searching in the wrong place in the wrong way

"In a nutshell," Bishop tells us, "the problem with analytic epistemology is that it searches for answers in the wrong place and in the wrong way." (ms 12) This is a great slogan for a cluster of views that Bishop and I share. But the slogan needs to be interpreted with care. In *Fragmentation*, and earlier in Stich (1988), I used "analytic epistemology" as a technical term for epistemological projects in which intuitions are used to support linguistic or conceptual analyses, and those analyses are taken to be the ultimate court of appeal in epistemological disputes. Though I did not make this as clear as I should have, the disputes I had in mind were disputes in *normative* epistemology.[19] In the papers I wrote with Weinberg and Nichols some years later, the term "analytic epistemology" is used again. But its meaning is less clear. Though I don't recall ever

[18] My thanks to Jonathan Weinberg for his acute and enormously helpful comments on several earlier drafts of this reply.

[19] For some discussion of the distinction between normative and descriptive epistemology, see Weinberg, Nichols & Stich (2001).

discussing the issue with my co-authors, my best guess as to what we had in mind was (something like) *epistemological projects in the analytic tradition that use intuitions as an important source of data*. The difference is an important one, since there are many projects that count as analytic epistemology on the latter interpretation but not on the former.[20] Bishop's slogan is consistent with my current view on the former interpretation of "analytic epistemology" but not on the more inclusive interpretation.

To make this a bit clearer, it will help to set out a taxonomy of some of the epistemological projects in which intuitions might be used as a source of evidence.

> (i) One project aims to "capture" our epistemic intuitions (or some sub-set of them, for example our intuitions about when a person does and does not have knowledge or justified belief) by producing a theory that will entail those intuitions – perhaps, as Bishop notes, with "some light revisions in the service of power or clarity" (ms 19) Of course, if different people or different groups of people have significantly different intuitions, then this project will fragment into a cluster of projects, each aimed at capturing the intuitions of a different group.
>
> (ii) A second project assumes that there is a tacit or implicit theory underlying people's ability to produce epistemic intuitions. The goal of the project is to give an account of that implicit theory. Often this project is not clearly distinguished from (i), since it seems natural to suppose that the implicit theory underlying our intuitions just *is* the theory that captures those intuitions. However, I think it is important to keep these two distinct, since there can be *lots* of different ways of capturing a person's epistemic intuitions, just as there can be lots of ways of capturing a person's grammatical intuitions. But on one understanding of what an implicit theory is, at most one of these could be the implicit theory underlying the person's intuitions.[21]
>
> (iii) A third project aims to analyze or characterize people's epistemic concepts, like the concept of knowledge or the concept of epistemic justification. Here again, it is easy to ignore the difference between this project and the one that precedes it, since some theorists maintain that concepts just *are* the tacit theories that guide our application of the associated term. But others adamantly insist that this is not the case.[22]

---

[20] As emerged clearly in my exchange with Sosa in this volume, some epistemologists, including Sosa himself, avoid the appeal to conceptual analysis, and take intuitions to be a source of data about the nature of knowledge. Also, some have little or no interest in normative questions; they are concerned *only* to analyze epistemic concepts or to characterize the nature of knowledge.

[21] Though not directly relevant to our current concerns, the issues here are surprisingly murky and difficult, as Kelby Mason and I discovered to our dismay in a recent correspondence with Frank Jackson. (Jackson, Mason & Stich, forthcoming)

[22] For further discussion of these debates, see Margolis & Laurence (1999) and Machery (forthcoming).

(iv) A fourth project is the one that is center stage in Sosa's essay. The goal is to characterize "the nature and conditions of human knowledge and other rational desiderata" like justification. Many philosophers tend to conflate this project with the previous one, where the goal is conceptual analysis. But as Sosa rightly insists, this is a mistake. Analyzing our own (or someone else's) *concept* of knowledge is quite distinct from characterizing the nature of knowledge, just as analyzing our concept of water or of disease is distinct from characterizing the nature of water or of disease.

(v) So far, all the projects on my list are instances of what I think of as descriptive epistemology. They can all be pursued without making any explicitly normative claims. But intuitions might also be used as evidence for a variety of theories in normative epistemology. As Bishop notes, the normative epistemological project that has been center stage in my work "focuses on the evaluation of methods of inquiry. It tries to say which ways of going about the quest for knowledge – which ways of building and rebuilding one's doxastic house – are good ones, which are bad ones, and why." Other normative projects aim at evaluating beliefs or other cognitive states.

On my view, intuitions are an entirely appropriate source of evidence in the first three of these projects, provided that the cautionary note at the end of (i) is kept in mind. Other sorts of evidence may be of considerable importance in (ii) and (iii), though how that evidence is used will turn on the resolution of some vexed questions about the nature of concepts and implicit theories.[23] In (iv) and (v), the use of intuitions is much more problematic. I think the best argument against the use of intuitions in (iv) and (v) relies on an empirical premise: different people (and different groups of people) have *different* epistemic intuitions. If this is right, then it is hard to see why we should trust *our* intuitions rather than those of some group whose intuitions disagree with ours.[24] This problem does not arise for the first three projects, since in those cases, in contrast with (iv) and (v), the goal of the project is to learn something about the psychology of the individuals offering the intuitions. If different people have different intuitions, then they may well have different implicit theories or different epistemic concepts. But the fact that people's intuitions differ does not pose a prima facie problem to the use of those intuitions as data.

Another widely discussed argument against using intuitions in projects like (iv) and (v) is the "calibration" argument developed by Robert Cummins (1998). This argument draws an analogy between intuition and instruments or procedures used to make observations or gather data in science. Before they are trusted, instruments and procedures need to be calibrated, and to do this "an invariable requirement … is that there be, in at least some cases, access to the target that is independent of the instrument or

[23] Bishop agrees that intuitions are "a legitimate source of evidence" (ms 20) for theories in category (i). He does not mention categories (ii) – (iv).

[24] For further discussion, see my Reply to Sosa.

procedure to be calibrated." (Cummins, 1998, p. 117). But in most cases where intuitions are used in philosophy, Cummins maintains, there is no independent access to the target, and in cases where there is independent access, the intuitions are typically superfluous. I'm much less impressed by this argument since, as a number of authors have noted, it threatens to generalize into a much more pervasive skepticism. Here is how Sosa makes the point:

> The calibration objection, if effective against intuitions will prove a skeptical quicksand that engulfs all knowledge, not just the intuitive. No source will then survive, since none can be calibrated without eventual self-dependence. That is so at least for sources broadly enough conceived: as, say, memory, introspection, and perception. (Sosa, 2007b, p. 64)

Moreover, even if there is some way of blocking the extension of the argument to perception and memory, it is likely that renouncing reliance on uncalibrated *intuition* would undermine large parts of mathematics. And that, I think, counts as a *reductio* of the calibration argument.

When Bishop says that analytic epistemology searches for answers *in the wrong place*, he means that analytic epistemologists rely on epistemic intuitions as a principal source of data. If the epistemologist's project is (iv) or (v), and if intuitions differ in different demographic groups, then there is excellent reason to think that the epistemologist is looking in the wrong place. When Bishop says that analytic epistemologists look for answers *in the wrong way*, he means that analytic epistemologists are doing "bad science."(ms 12) Here there is no need to restrict the criticism to a subset of the epistemological projects listed above. Indeed, there is no need to restrict the criticism to analytic *epistemology*. Philosophers in the analytic tradition regularly make claims about "our" intuition and use these as data for claims about the "ordinary conception" of x or the "folk theory" of y. Moreover, these claims are often of central importance to the philosopher's project. Here is a passage in which Frank Jackson makes the motivation for such claims admirably clear.

> How … should we go about defining our subject *qua* metaphysicians when we ask about Ks for some K-kind of interest to us. It depends on what we are interested in doing. If I say that what *I* mean – never mind what others mean – by a free action is one such that the agent would have done otherwise if he or she had chosen to, then the existence of free actions so conceived will be secured, and so will the compatibility of free action with determinism. If I say what *I* mean – never mind what others mean – by 'belief' is any information-carrying state that causes subjects to utter sentences like 'I believe that snow is white', the existence of beliefs so conceived will be safe from the eliminativists' arguments. *But in neither case will I have much of an audience. I have turned interesting philosophical debates into easy exercises in deduction from stipulative definitions together with accepted facts*.

> What then are the interesting philosophical questions that we are seeking to address when we debate the existence of free action and its compatibility with determinism, or about eliminativism concerning intentional psychology? What we are seeking to address is whether free action *according to our ordinary conception*, or something suitably close to our ordinary conception, exists and is compatible with determinism, and whether intentional states *according to our ordinary conception*, or something suitably close to it, will survive what cognitive science reveals about the operations of our brains. (Jackson 1998, p. 31, emphasis in the first paragraph added; emphasis in the second paragraph is Jackson's)

Obviously, claims about "our ordinary conception" of belief or free will are empirical claims. What sort of evidence does Jackson have for these claims? Here is what Jackson says:

> I am sometimes asked – in a tone that suggests that the question is a major objection – why, if conceptual analysis is concerned to elucidate what governs our classificatory practice, don't I advocate doing serious opinion polls on people's responses to various cases? My answer is that I do – when it is necessary. Everyone who presents the Gettier cases to a class of students is doing their own bit of fieldwork, and we all know the answer they get in the vast majority of cases. But it is also true that often we know that our own case is typical and so can generalize from it to others. (Jackson, 1998, pp. 36-37)

I think the views that Jackson is expressing in this last quote are (or *were* until very recently) widely shared among philosophers in the analytic tradition. By my lights, they constitute a truly shocking indictment of that tradition. The "serious opinion polls" that Jackson and other analytic philosophers conduct to support their claims about the concepts or implicit theories underlying "our" classificatory practice violate just about every methodological standard that social scientists have adopted to avoid bias and distortion in survey research. The opinions solicited in this sort of "fieldwork" are typically those of students in elite universities, who are hardly a representative sample of "the folk". Moreover, the students who take philosophy courses – particularly advanced courses where much of this fieldwork is done – are a self-selected group who are unlikely to be representative even of students at elite universities. We know very little about the factors that lead students to take advanced philosophy courses. But one possibility that surely must be taken seriously is that students whose intuitions do not match those of their teachers do not enjoy or do well in lower level courses and do not continue on in philosophy.[25] If one were looking for a textbook example to illustrate what social scientists mean by *sample bias*, the sort of "serious opinion poll" that Jackson has in mind would be a fine candidate. It would also serve as a fine example of experimenter bias since the informal polls are conducted by an authority figure with a strong

[25] This possibility was noted, albeit briefly, in Nichols, Weinberg & Stich (2003, p. 232) where we present some preliminary data indicating that the epistemic intuitions of students who have taken little or no philosophy differ from those of students who have taken two or more courses.

antecedent belief about how the "experiment" should turn out. As if all of this were not enough, the typical classroom opinion poll conducted by philosophers requires that students indicate their judgment in a very public way, and social psychologists have long known that procedures like this have a strong tendency to suppress dissention opinions.[26]

The fact that disagreement in intuition in different demographic groups would pose a serious challenge for projects (iv) and (v) underscores the importance of doing carefully controlled methodologically sophisticated studies that *look* for disagreement across demographic divides. But even if, like Jackson, one's goal is conceptual analysis or the characterization of an implicit theory underlying classificatory practice, it is important to get out of the armchair and do well designed, carefully controlled studies that look for systematic differences in intuitions. For without this sort of work, there is no way of knowing whether one's own intuitions (or concepts or implicit theories) differ from those of others. And thus there is no way of addressing Jackson's concern that one may have turned an interesting philosophical debate into an uninteresting exercise for which there may not be much of an audience.

2. The prima facie tension

At the end of his first section, Bishop suggests that there is a prima facie tension between my view about the use of intuition in epistemology in *Fragmentation* and some of the "hints" to be found in Weinberg, Nichols & Stich (2001). I think Bishop is right; there is some tension there. But the issues are rather more delicate and complex than Bishop suggests. In *Fragmentation* I saw little use for epistemic intuitions or for the concepts of epistemic evaluation that they reflect, because I assumed they were culturally local and idiosyncratic products of a process of cultural transmission that had very likely led to different concepts and different intuitions in other cultures. Why, I asked, should we think that our intuitions and our concepts of epistemic evaluation are any better than the many alternatives that were likely to exist in other groups? But, as Bishop rightly notes, when *Fragmentation* was written, there was no serious evidence for the claim that epistemic intuitions and concepts varied cross-culturally. It was little more than a guess.

A decade later, Weinberg, Nichols and I decided to explore the issue empirically. What we found offered a bit of evidence for the speculation in *Fragmentation*. However, the picture was more complicated than I had imagined. Yes, indeed, some epistemic intuitions – including Gettier intuitions! – did appear to be different in different demographic groups. But, at least in the small number of groups we looked at, there were also *some* intuitions that did not vary significantly. All of our subjects agreed that beliefs based only on a "special feeling" do not count as instances of knowledge. So perhaps there is a universal core to folk epistemology – a set of principles on which everyone agrees. We are, to put it mildly, *very* far from knowing that this is the case. It will take a great deal more work in experimental philosophy, looking at many more cases in many more demographic groups, before we have any serious idea about the nature or even the existence of this universal core.

[26] Ross & Nisbett (1991), Ch. 2.

Suppose it turns out that there *is* such a universal core. Would there be any *normative* implications? Could the discovery of a universal core be used to argue that experimental philosophy can play more than the "normatively modest" role that Bishop describes – "confirming or disconfirming the empirical assumptions or empirical implications of genuinely normative philosophical theories" but "incapable of offering up positive normative theories or principles"? (ms 16-17) The answer is far from clear. On the one hand, the discovery of universal intuitions and core principles would undermine the demographic variation argument which, I maintain, is the best argument against the use of intuitions in epistemology – or, to be a bit more precise, it would undermine the argument when applied to *those* intuitions.

On the other hand, the mere fact that a cluster of intuitions or a set of principles are universally shared surely does not entail that the intuitions can be relied on or that the principles are good ones. In deciding whether or not to rely on universally shared intuitions and principles, one would like to know much more about the *source* of these intuitions and principles. Why does every one have them? How did they come to be universally shared? There are *some* answers to these questions that would cast serious doubt on the reliability of intuition and other answers that would encourage trust. In his critique of the use of intuitions in philosophy, Cummins (1998) argues that all the plausible hypotheses about the source of intuitions fall into the former category. I'm not sure he's right about this. What I am sure of is that before we can say with any assurance whether or not experimental philosophy can play more than a normatively modest role in epistemology we will have to know a great deal more about the etiology of epistemic intuitions and think much harder about the ways in which etiologies can sustain or undermine confidence in intuition. So I am in agreement with Bishop when he writes:

> …I don't believe there is a case to make that our intuitions are *in principle* irrelevant evidence for or against a genuinely prescriptive epistemological theory. (And as far as I know, no one, including Stich, has made such a case.) (ms 20; emphasis in the original)

If there is a case to be made for or against an evidentiary role for intuitions in prescriptive epistemology, it will not be an "in principle" argument. Rather, it will be an argument in which both an empirically supported account of the extent to which epistemic intuitions are (or are not) demographically variable and an empirically supported account of the etiology of intuitions play a central role. Moreover, even if it does turn out that some intuitions are good *prima facie* evidence for normative epistemological claims, it is entirely possible that they may be overridden by other, better sources of evidence. Here again, I am in broad agreement with Bishop, who notes that "whether it is reasonable to ignore [intuitions] in our epistemological theorizing depends on whether we have an evidential source that is better than our intuition." (ms 22)

3. Pragmatism *vs.* Strategic Reliabilism

In chapter 5 of *Fragmentation*, I offered an account of what it is for a belief to be true, and argued that if that account is on the right track, then most people won't find true belief to be either intrinsically or instrumentally valuable, once they have a clear view of what true belief *is*. I went on to argue that rather than evaluating belief-forming cognitive processes as a reliabilist would, by asking how well they do at producing true beliefs, we should instead evaluate them on pragmatic grounds, asking which system of cognitive processes would be most likely to achieve those things that we do find intrinsically valuable. In section 3.1 of his paper, Bishop does an admirable job of sketching my arguments for the conclusion that, on reflection, most of us will recognize that we have no good reason to value true beliefs over true* beliefs – where most true* beliefs are also true, and those that are not are pragmatically preferable. Though these arguments are aimed at supporting what Bishop maintains is "one of the more shocking positions in the Stich oeuvre," (ms 23), he courageously endorses them: "These are powerful arguments. And properly understood, I do not want to challenge them." (ms 26). [27]

But having endorsed the shocking bit, Bishop resists taking the next step. Even though "we have no good reason to value true beliefs (intrinsically or instrumentally)"(ms 26) he nonetheless thinks that a version of reliabilism can be defended as a way of evaluating belief-forming strategies. The sort of reliabilism that Bishop has in mind is the sophisticated and innovative "Strategic Reliabilism" that he has developed in collaboration with J. D. Trout. (Bishop & Trout, 2005) The defense he proposes is one that I could hardly oppose in principle, for what he maintains is that Strategic Reliabilism is better than direct pragmatic assessment of belief-forming strategies *on pragmatic grounds*. Despite initial appearances, there is nothing paradoxical about this approach. A nuanced reliabilism may be pragmatically superior to the direct pragmatic assessment of cognitive strategies if the latter approach turns out to be psychologically unrealistic, recommending belief-forming strategies that are either psychologically impossible or so difficult that the costs would be prohibitive. Bishop contends that we have a strong tendency to value true beliefs and a strong inclination to believe p when we are convinced that p is true, and he thinks these psychological facts spell trouble for direct pragmatism.

> The cool-eyed pragmatist will be the first to insist that a theory of cognitive evaluation should not make demands on us that we can't meet. Our judgment and decision-making capacities are deeply imperfect, and the limits on our memory, computing power, time, energy, patience, and will are legion (Stich, 1990, pp. 149-158). I'm suggesting the pragmatist add one more imperfection to the list: *We tend to value truth, even when, from a pragmatic perspective, we shouldn't*. Once we take this fact about ourselves to heart, the pragmatist is faced with a familiar challenge: What sorts of normative principles, theories or recommendations can we offer that will effectively guide our reasoning but that will clearly recognize and compensate for our built-in limitations and imperfections? Perhaps our regrettable *attraction to true belief* gives us pragmatic grounds for placing truth at the center of our epistemological theory. Not because truth is more valuable to us

[27] In endorsing these arguments, Bishop has little company. Most commentators refuse to take them seriously. The interesting critique in DePaul (forthcoming) is a noteworthy exception.

than truth* (or truth** or truth***...) but just because *we're stuck valuing true belief*.

> To compensate for our unfortunate attraction to true belief, two different strategies suggest themselves. A direct strategy, which Stich adopts in *Fragmentation*, places pragmatic virtues center stage. Normative claims about cognitive matters – generalizations about good reasoning as well as evaluations of particular cognitive states and processes – are framed directly in terms of what we intrinsically value. An indirect strategy would place truth (or some other non-pragmatic category) center stage but would find some way to license the adoption of false (or true*) beliefs when they serve our interest. What might such a theory look like? I suggest it might look something like Strategic Reliabilism. (ms 29-30, italics added)

Now I'm afraid that all of this goes by a bit too quickly for me. I'm not sure whether Bishop is claiming

> (1) that in our role as epistemologists the (putative) fact that we're stuck valuing true belief makes it impossible or very difficult to advocate belief-forming strategy A over belief-forming strategy B if we believe that B will lead to more true beliefs than A

or

> (2) that in our role as cognitive agents, we are so strongly attracted to true beliefs that we cannot adopt a belief-forming strategy that leads us to form what we take to be a false belief

or

> (3) that since (2) is the case, epistemologists ought not to advocate belief-forming strategies that urge cognitive agents to form beliefs they take to be false or resist forming beliefs they take to be true.

The first of these options strikes me as distinctly unpromising. When wearing my epistemologist's hat, I don't find it at all difficult to advocate the sort of belief-forming strategy that (1) claims it is all but impossible to advocate. Nor am I alone. In one of the more memorable sections of their book, Bishop and Trout (2005) take L. Jonathan Cohen to task for doing much the same thing.[28] The most charitable interpretation, I suspect, is that Bishop is advocating both (2) and (3). But, for two reasons, I am unconvinced. The first reason is that, while Bishop *asserts* (2) on several occasions, he offers no serious evidence. Moreover, since Bishop has an admirable record of citing well done studies to back up his empirical claims, there is good reason to think that this is not merely an oversight. I know of no studies indicating that "we are naturally drawn to true belief even when it is against our interests to do so" (ms 23), and I am quite confident that if Bishop knew of any he would have mentioned them prominently. The second reason is that, for the issue at hand, (2) is not really the relevant claim. What we need to know is not whether people are naturally drawn to true belief, but whether this putative attraction would survive if

[28] Ch. 8, Sec. 3. The strategy that Cohen recommends, I hasten to add, is very different from the one I would recommend.

people became convinced of (something like) my account of what true belief *is*. It is pretty clear that Bishop thinks that it would.

> We value the truth, despite the results of our Stich-inspired deliberations on its idiosyncrasies and practical failings. The truth is like the prodigal son. We might realize that he does not deserve our love, our care, our energy; we might realize that we would be much better off committing those feelings and resources to a more deserving child. But despite what our heads say, we can't help but embrace him. (ms 26)

He might be right, of course. But the trajectory of people's values, preferences and patterns of belief formation under counterfactual conditions – particularly counterfactual conditions that are quite different from anything we've observed, are notoriously hard to predict. And here again, Bishop offers no evidence; his prediction is simply speculation. Since his argument turns crucially on unsupported speculation, I think the verdict must be that Bishop has failed to make a convincing case for the pragmatic superiority of Strategic Reliabilism.

Reply to Goldman

In a number of publications, Shaun Nichols and I have argued that the term 'simulation', as it us used in the literature on mindreading, picks out no natural or theoretically interesting category, and we've urged that the term be retired. (Stich & Nichols 1997; Nichols & Stich 1998, 2003) The main aim of Goldman's paper is to respond to this challenge by providing a "defense of the naturalness, robustness, and theoretical interest of simulation." (ms 2) But he also has a secondary aim, which is to argue that there is "a serious lacuna" in the Nichols and Stich book, *Mindreading* (2003). We have, according to Goldman, failed to discuss the literature in cognitive neuroscience, though "this is where the best evidence for simulation, including simulation-based mindreading, resides. It is a major omission of Nichols and Stich (2003) to neglect cognitive neuroscience." (ms 2) Goldman's contention that simulation is a natural, robust and theoretically interesting category raises important issues, and most of this reply will be devoted to examining Goldman's defense of that claim. But before getting on to that, I need to say something about Goldman's accusation that Nichols and I neglected important evidence.

From the edgy tone of Goldman's remarks about our "grudging appreciation of simulationism's virtues" in our "seemingly endless stream of articles" (ms 1), the reader might perhaps infer that Nichols and I had intentionally ignored relevant neuroscience literature, or worse that, like Jerry Fodor in some of his moods, we believed that neuroscience could be of no value in the study of mindreading, or in addressing broader issues about how the mind works. Nothing could be further from the truth. Here is what we say on the issue:

> On our view … *findings about the structure and functioning of the brain can, and ultimately will, impose strong constraints on theories of mindreading of the sort we will be offering*. For the moment, however, this … is of little practical importance, since as far as we have been able to discover, there are few findings about the brain that offer much guidance in constructing the sort of cognitive account of mindreading that we'll be presenting…. *In principle, our stance regarding evidence is completely eclectic. We would have no reservations at all about using evidence from neuroscience* or from history, anthropology or any of the other social sciences to constrain and test our theory, though we have found relatively little in these domains that bears on the questions we'll be considering. (*MR*, pp. 11-12, emphasis added)

Clearly, Nichols and I have no principled objection to using evidence from neuroscience in deciding among alternative accounts of mindreading.

Were we, then, just slipshod scholars who failed to realize that the many interesting and important studies Goldman cites were there to learn from? Here again, the charge is without foundation. In Goldman's recent book on mindreading (Goldman, 2006) he draws the useful distinction between "low-level mindreading" and "high-level mindreading". Roughly speaking, high level mindreading is the formation of beliefs about propositional attitudes like beliefs, desires, intentions and decisions, while low level mindreading is the formation of beliefs about sensations, like feelings of pain, and about emotions, like fear, disgust and anger.[29] Face-based emotion recognition, which Goldman discusses in his paper in this volume, occupies almost the entire chapter on low-level mindreading in Goldman's book. But, as Goldman notes, cases of low-level mindreading are "a somewhat atypical sector of mindreading, *different from the cases usually treated in the literature, especially the philosophical literature*. The stock examples in the literature are attributions of garden-variety propositional attitudes: belief, desire, intention, and so forth." (ms 9, italics added)[30] And as Goldman argues in his book, "there seem to be different mechanisms for mindreading [emotions and feelings]… than for mindreading the attitudes, more primitive and automatic mechanisms." (Goldman, 2006, p. 20) In our book, and in our "seemingly endless stream of articles" on mindreading, Nichols and I followed the lead of the philosophical literature and focused on high-level mindreading. The theory we defend in the book is an attempt to

---

[29] Though Goldman concedes that "a strict definition of high- versus low-level mindreading processes is lacking" he offers the following somewhat more detailed characterization of high-level mindreading:

> High-level mindreading is mindreading with one or more of the following features: (a) it targets mental states of a relatively complex nature, such as the propositional attitudes; (b) some components of the mindreading process are subject to voluntary control; (c) the process has some degree of accessibility to consciousness. (Goldman 2006, p. 147)

[30] Goldman makes much the same observation at several places in his book. Here is an example: "The standard treatments of 'theory of mind' have primarily studied the locus of *high-level* mindreading…." (p. 141). Cf. also p. 20. Of course, it is no accident that philosophers have focused on high-level mindreading, since that is the sort of mindreading that is most directly relevant to debates in the philosophy of mind. For a discussion of the links between issues in the philosophy of mind and theories of high-level mindreading, see Nichols & Stich (2003), pp. 5-9.

characterize the psychological mechanisms underlying high-level mindreading. We do not even try to offer an account of the quite different mechanisms underlying low-level mindreading. To be sure, low-level mindreading is an interesting and important phenomenon. But it is not what our book was about.

> Now here are some striking facts about Goldman's paper:
>
> (1) The longest section, "Simulation and Motor Cognition," is devoted to "simulation in a certain type of *non-mindreading activity*" (ms 4, emphasis added)
>
> (2) The second longest section, "Simulation and Face-based Emotion Attribution," is devoted to *low-level* mindreading
>
> (3) There is *not a single sentence* aimed at showing that neuroscientific results are relevant to assessing theories of high-level mindreading of the sort that Nichols and I proposed.

Of course, there were limits imposed on the length of contributions to this volume. But even in the chapter devoted to high-level mindreading in his book, where presumably length constraints were not a major concern, I have been unable to find a single reference to a finding in neuroscience – published prior to mid-2002, when the Nichols and Stich volume went to press – that would be relevant to assessing the sort of theory that Nichols and I develop in our book.[31]

In light of all this, the only reasonable verdict on Goldman's accusation that our failure to discuss evidence from cognitive neuroscience is a "major omission" is that it is completely unwarranted. If the charge is that we have some principled antipathy to

[31] The chapter in question is called "High-Level Simulational Mindreading." The first three substantive sections of the chapter (7.2 – 7.4) are devoted to addressing a potential problem for Goldman's account of high-level mindreading. That account gives an important role to the process of enactment-imagining (or E-imagining) which Goldman characterizes (for the first time!) earlier in the book. The problem that Goldman seeks to address is: "Can E-imagining produce states that truly resemble their intended counterparts?" (p. 149) "How similar is E-imagined desire-that-p to genuine desire-that-p? How similar is E-imagined belief-that-p to genuine belief –that-p?" (p. 151) As Goldman notes, for his theory to be plausible the answer must be that they are often quite similar. But, Goldman laments, "detailed research on these topics, unfortunately, is sparse." (p. 151) What to do? Goldman's strategy is to look at evidence regarding "two species of E-imagination" that have little or nothing to do with mindreading, viz. motor imagery and visual imagery, "in the hope that what we learn is more widely applicable."(p. 151) In the sections devoted to these topics (7.3 & 7.4) Goldman cites *lots* of evidence from neuroscience, some of which is also discussed in his paper in this volume. But, of course, it would be absurd to suggest that Nichols and I should have taken account of *this* evidence, since (i) it has no direct bearing on mindreading, and (ii) it's indirect bearing, via the "hope" that the similarities discovered would be "more widely applicable" emerged only after Goldman spelled out the role of E-imagination in *his* theory of high-level mindreading, in a book published three years after ours. The remaining sections of Goldman's chapter (7.5 – 7.13) are devoted to issues that are more directly related to high-level mindreading. But in those sections there are only two references to neuroscientific studies that bear directly on high-level mindreading, one by Mitchell, Banaji and Macrae, the other by Samson et al., and both of these were published in 2005. So the "serious lacuna" about which Goldman complains must be that Nichols and I failed to take note of research that had not yet been done. Guilty as charged!

neuroscience, my rebuttal could not be clearer: we say quite clearly that neuroscience "can and ultimately will impose strong constraints on a theory of mindreading." If the charge is that we failed to take account of relevant research in neuroscience, my reply is that, when our book went off to the publisher, there *was* no relevant published research in neuroscience to report.

So much for Goldman's canard about there being a serious lacuna in our book. Let me turn, now, to the more interesting and substantive issue that Goldman raises. Is simulation a natural and theoretically interesting category? To start us off, a bit of history will be helpful. Prior to 1986, most philosophers of mind assumed that mindreading – the process in which we attribute mental states to people, predict future mental states, and predict behavior on the basis of these attributions – was subserved by a commonsense psychological theory, often called "folk psychology." That assumption played an important role in many philosophical debates. Functionalists maintained that folk psychology determined the meaning of commonsense mental state terms. Eliminativists agreed, though they went on to argue that folk psychology was radically false and thus that commonsense mental state terms did not denote anything. In a pair of important papers, published in 1986, Robert Gordon and Jane Heal suggested another way in which some predictions of other people's mental states might be made. Rather than use a theory to predict what someone will decide to do, we could exploit the fact that we have a decision making system that is similar to theirs. So all we need to do is pretend to be in the target's situation – having her beliefs and her desires rather than our own – and then make a decision about what to do in that situation. Of course, we don't go on to act on that decision. Rather we predict that that is what the target will decide. Since we use our own decision making mechanism to simulate the mechanism used by the target, "simulation theory" seemed a natural label for the account. The alternative account, which maintained that predictions like this were subserved by a folk psychological theory, came to be know as the "theory-theory". In 1989, Goldman published a widely discussed paper in which he endorsed and clarified simulation theory, and set out a number of arguments for it and against the theory-theory. (Goldman, 1989) Soon after, Nichols and I joined the debate with a paper that criticized many of the arguments for simulation theory offered by Gordon and Goldman. (Stich & Nichols, 1992) That paper also offered the "boxological" sketch of simulation-based decision and behavior prediction reproduced in Figure 1. The sketch apparently did a good job at capturing what advocates of simulation theory had in mind, since a number of them reproduced it in their own writings. Goldman himself has reproduced it three times. (Goldman, 1992, 1993; Gallese & Goldman, 1998)

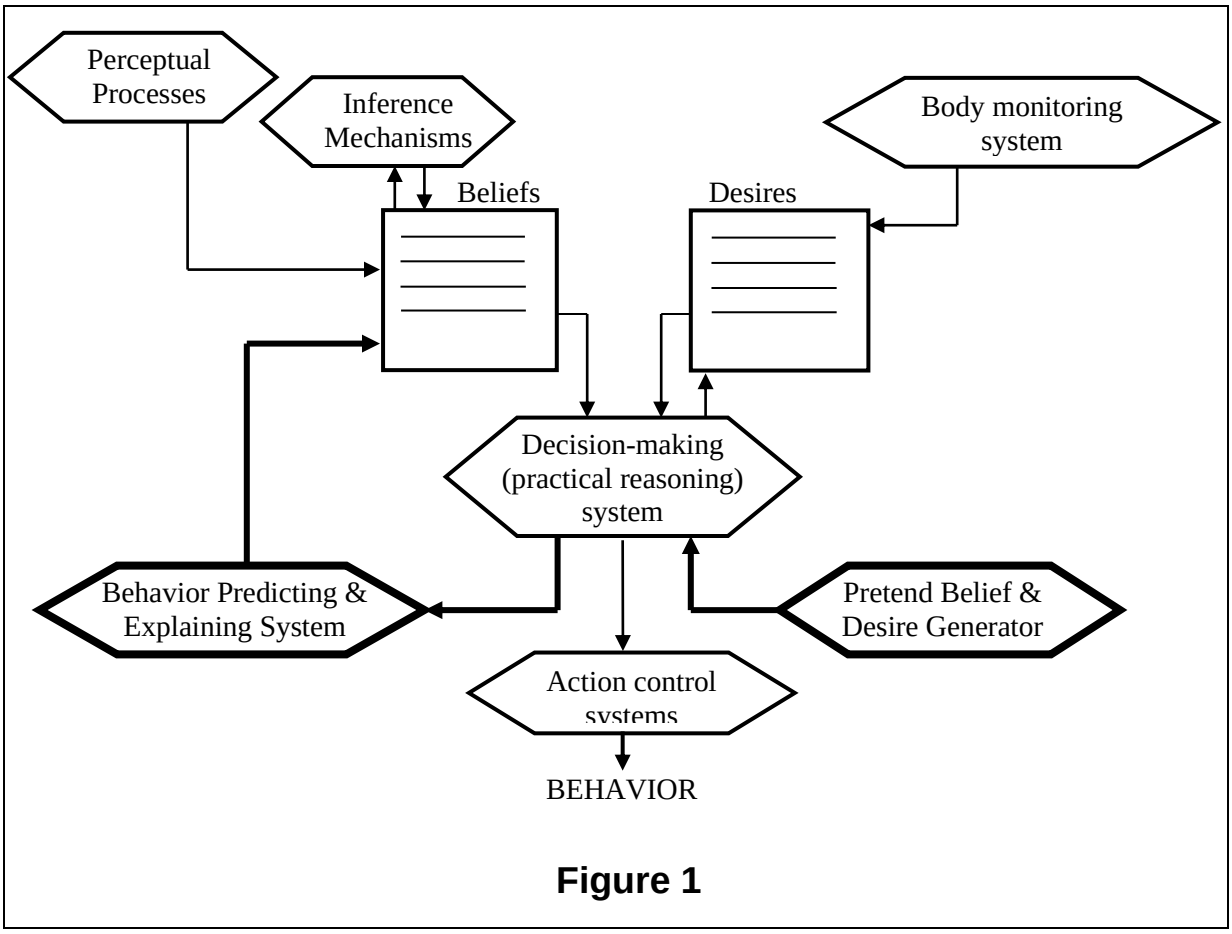
Perceptual Processes
Inference Mechanisms
Body monitoring system
Beliefs
Desires
Decision-making (practical reasoning) system
Behavior Predicting & Explaining System
Pretend Belief & Desire Generator
Action control systems
BEHAVIOR

Figure 1

In the early years of the debate over simulation theory, many writers – Nichols and myself included – viewed the debate as a two sided battle which either simulation theory or theory-theory would "win". But for two reasons this turned out to be a misleading way to think about mindreading. First, it ignored the possibility that the correct account might be a *hybrid* theory in which some aspects of mindreading are subserved by simulation and others are subserved by a folk psychological theory. As research progressed, many people, including both Goldman and Nichols and I, began to converge on the view that the right account of mindreading is indeed a hybrid account – though there is still plenty of debate about the details. Second, and more directly relevant to the issue at hand, it was far from clear what it would take for one side or the other to "win", since there was no clear and agreed on characterization of what to count as a simulation process or mechanism. For the most part, Nichols and I relied on paradigm cases of simulation mechanisms like the one depicted in Figure 1. If the correct theory for some aspect of mindreading relied on something similar to those paradigms, we counted it as a victory for simulation theory. This was at best a rough and ready strategy, since we never tried to specify *how* similar a candidate has to be, or in what respects it has to be similar. But it did enable us to say that some proposed processes were *clear* examples of simulation while others *clearly* were not.[32] But it gradually became apparent that others were not using the term "simulation" in this way. Rather, the growing group of advocates of simulation argued that just about any process or phenomenon or mechanism that could plausibly be described as a "simulation" in some sense or other counted as a victory for their side, although many of these bore no obvious similarity to mechanisms like the one depicted in Figure 1, and had little or nothing in common with each other.[33] Frustrated by a situation that made it impossible to have a well focused argument about the merits of simulation theory – since there was no consensus at all on what simulation is – Nichols and I wrote the passage that Goldman quotes toward the beginning of his paper, in which we urged that the term "simulation" be retired because "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." (Stich & Nichols, 1997, 299)

Goldman apparently has some sympathy with our frustration; he does "not want to defend every application of the term 'simulation' that anybody has ever proposed." (ms 2)[34] Nonetheless, he thinks there is a natural and theoretically interesting category in this vicinity, and if he is right, then "simulation" would surely be a reasonable name

[32] In *Mindreading* we adopt this policy quite explicitly (cf. p. 134).

[33] In *Mindreading*, pp. 132-134, we offer a list of ten examples ranging from predicting what the consequences would be if war broke out in Saudi Arabia to having a conception of people as "peepholes through which the world reappears, possibly transformed." It would have been easy to assemble a much longer list.

[34] Indeed, it appears that he no longer wants to defend every application of the term that he himself has proposed. The list mentioned in the previous footnote includes several examples of phenomena that Goldman once cited in support of simulation theory but which do *not* count as mental simulations on his current account, which we will examine below.

for that category. Giving an account of this category is "a complex and delicate matter" which, Goldman tells us, he "won't try to cover thoroughly" (ms 1) in his paper in this volume. For a more detailed account he refers us to the extended discussion in his book. Since, as is often the case, the devil is in the details, I will focus on the account in Goldman's book.

In the book, Goldman carefully sets out a set of increasingly specific and detailed definitions, starting with a first pass at defining "Generic Simulation" and ending with a definition of "Attempted Mental Simulation". The core idea, as he explains in his paper in this volume, is that "one process successfully simulates another … only if the first process copies, replicates, or resembles the target process, at least in relevant respects," (ms 3) This idea leads to his first definition:

> Generic Simulation (initial): Process P is a simulation of another process P' $=_{df.}$ P duplicates, replicates, or resembles P' in some significant respects (significant relative to the purposes of the task). (Goldman, 2006, p. 36)

In discussing this definition, Goldman notes that the target activity, P', "can be merely hypothetical rather than actual" – as when an episode in a flight simulator simulates a crash that never really happens. But this initial definition needs refinement because similarity is a symmetrical relation while simulation is not. An actual crash does not simulate what happens in the flight simulator. One way to patch the problem, Goldman notes, would be to "require that the simulating process occur out of the *purpose*, or *intention*, to replicate the simulated process…. This won't quite work, however, because it is doubtful that all simulation is purposeful. Some simulation may be automatic and nonpurposeful." (Goldman, 2006, p. 37) To remedy the problem, Goldman suggests, we can require that one phenomenon counts as a simulation of another "if it is the *function* of the former to duplicate or resemble the other." But how, exactly are we to understand this invocation of the tricky and contested notion of *function*? To his credit, Goldman admits that he has no answer. "I lack a theory of functions to provide backing for this approach, but I shall nonetheless avail myself of the notion." (Goldman, 2006, p. 37) Which he does, in the following, revised, definition of Generic Simulation.

> Generic Simulation (revised): Process P is a simulation of another process P' $=_{df.}$
> (1) P duplicates, replicates, or resembles P' is some significant respects (significant relative to the purposes of the task), and
> (2) in its (significant) duplication of P', P fulfills one of its purposes or functions.
> (Goldman, 2006, p. 37)

Generic simulation applies to both mental and non-mental processes. Goldman's next definition focuses in on simulations that are restricted to mental processes.

> Mental Simulation: Process P is a mental simulation of another process P' $=_{df.}$
> Both P and P' are mental processes (though P' might be purely hypothetical), and P and P' exemplify the relation of generic simulation previously defined.
> (Goldman, 2006, pp. 37-38)

But one more refinement is still required, since the definition of Mental Simulation requires that the simulation actually resembles the target, and "a reasonable version of ST [simulation theory] would not hold that the mental processes of mindreaders always match, or even approximately match, those of their targets. ST, like any plausible theory of mindreading, should tolerate highly inaccurate specimens of mindreading…. What ST essentially maintains is that mindreading (substantially) consists of either *successful* or *attempted* mental simulations." (Goldman 2006, p. 38) To capture the idea of attempted mental simulations, Goldman offers two more definitions.

> Attempted Generic Simulation: Process P is an attempted generic simulation of process P' $=_{df.}$
> P is executed with the *aim* of duplicating or matching P' in some significant respects.
>
> Attempted Mental Simulation: Process P is an attempted mental simulation of process P' $=_{df.}$
> Both P and P' are mental processes, and P is executed with the *aim* of duplicating or matching P' in some significant respects. (Goldman, 2006, p. 38, emphasis in the original)

Finally, Goldman tells us, "[t]he term *aim* in these definitions includes covert or implicit aims, not consciously available to the simulator." (Goldman, 2006, p. 39, emphasis in the original)

The appeal to an *aim* or *purpose* looms large in Goldman's account of high-level mindreading. However, as we noted earlier, Goldman thinks low-level mindreading is subserved by "more primitive and *automatic* mechanisms" (Goldman, 2006, p. 20) like those responsible for face-based emotion attribution, discussed in his article in this volume, and his "case for low-level mindreading … was predicated on genuine resemblances between states of the attributor and the target." (Goldman, 2006, p. 150) His "case for high-level mindreading, by contrast, rests on the ostensible purpose or function of E-imagination, not on the regular achievement of faithful reproduction." (Goldman, 2006, p. 150)

I maintain that Goldman's definitions fail to pick out a natural category either in the case of low-level mindreading or in the case of high-level mindreading. In both cases, the problems are generated by cases of *mistaken* mindreading, many of which, on Goldman's account, will not count as episodes of simulation at all. The problem is easiest to see in the case of low-level mindreading. In Goldman's article he recounts a number of fascinating studies in which brain damaged patients failed in low-level mindreading tasks at a dramatically higher rate than normal subjects. But even the normal control subjects make mistakes in these experiments, as do we all when we attribute fear, disgust and anger to other people on the basis of their facial expressions. Now consider three hypothetical cases. In the first case, the target is angry and after a

relatively brief glimpse at her face I come to believe that she is angry. The process is subserved by the sort of primitive, automatic mechanism that Goldman posits. In the second case, the target is not really angry, she is just pretending. Here too, after a relatively brief glimpse of her face I come to believe that she is angry, using the same primitive, automatic process. In the third case, the target is not angry, she is fearful. However, because of the shadow cast on her face by a nearby tree, her facial expression looks more like an anger face than like a fear face. Once again, after a brief glimpse, the same primitive, automatic process leads me to believe that she is angry. On Goldman's account, the first episode counts as a case of mental simulation. But in the second and third cases there is no "genuine resemblance between the states of the attributor and the target". So, according to Goldman's definition, it is not a case of mental simulation at all. Goldman is, of course, free to define terms as he sees fit. However, the issue at hand is whether simulation is a *natural* and *theoretically interesting* category, and it is hard to see how a category which includes episodes of *accurate* face based emotion recognition but excludes nearly identical episodes of *mistaken* face based emotion recognition is either natural or theoretically interesting.

In the case of high-level mindreading, Goldman has more wiggle room to accommodate mistaken mental state attributions, since to count as an "*attempted* mental simulation" a process underlying high-level mindreading need only be "executed with the *aim* of duplicating or matching" the mental state of the target. Successful matching is not required. Nonetheless, it seems that in many cases of high-level mindreading, there is no reason at all to think that the mindreader has any such aim or purpose. To see the point, let's once again consider some hypothetical examples. Case 1: You are sitting at a restaurant, wanting to get the check after your meal. As the waiter passes, you glance at your watch, thinking that this will be a discrete way of indicating your desire to him. This leads the waiter to form the belief that you want your check. Case 2: You are sitting at the restaurant after your meal, enjoying a pleasant conversation with your dinner companions. You have no desire at all to get the check any time soon. But as the waiter passes you happen to glance at your watch, and as before this leads the waiter to form the belief that you want your check. On the account Goldman proposes in his book (2006, pp. 183-5), this sort of mindreading is subserved by a "generate-and-test" strategy.

> The "generate" stage produces states or state combinations that might be responsible for the observed … evidence. Hypothesis generation is presumably generated by non-simulative methods. The "test" stage consists of *trying out* one or more of the hypothesized state combinations to see if it would yield the observed evidence. This stage might well employ simulation. One E-imagines being in the hypothesized combination of states, lets an appropriate mechanism operate on them, and sees whether the generated upshot matches the observed upshot. (Goldman, 2006, p. 184, emphasis in the original)

Goldman acknowledges that he "knows of no theoretical analysis or experimental evidence that bears directly on simulation's role" (Goldman, 2006, p. 184) in this sort of mental state attribution. But, for argument's sake, let's assume that he is right. If all of this generating and testing *is* going on in the waiter, it is clear that, in most cases at least,

the waiter is not consciously aware of it. That's not a problem for Goldman, since he states very clearly that "a great deal of mindreading, even high-level mindreading, is nonconscious or minimally conscious, so we should allow simulational processes to include E-imaginative states even when the latter are entirely nonconscious." (Goldman, 2006, p. 151) So, if Goldman's generate-and-test hypothesis is correct, Case 1 is a clear example of successful simulation, since the process the waiter goes through resembles the process the target went through, at least "in some significant respects". But now what about Case 2? Here, we can assume, the processes going on in the waiter, both conscious and unconscious, are identical with those in Case 1. But in this case the process is not a successful mental simulation, since the waiter gets it wrong. The target never went through a process that resembles the process the waiter went through, since the target doesn't desire to get his check and never intended to convey anything to the waiter. So is this a case of mental simulation? If it is, then it must be a case of *attempted* mental simulation, and in order for that to be the case, the process must be executed with the aim or purpose of duplicating or matching the mental state of the target. Does the waiter have any such aim? Surely he is not conscious of having it. Is it, then, one of those "covert or implicit aims, not consciously available to the simulator"? (Goldman, 2006, p. 39) Well, perhaps. But why should we think the waiter has such an aim, even unconsciously? It is hardly needed for the rest of the generate-and-test process to do its job (or fail to), since that process could be triggered *automatically*, much as low-level mindreading processes are on Goldman's account. Nor, to the best of my knowledge, is there a shred of evidence suggesting that mindreaders have unconscious aims in cases like this. Certainly, Goldman offers us none. But if there are no unconscious aims – if the process is simply triggered automatically – then the waiter's mistaken mindreading is not a case of simulation at all. Moreover, this is hardly a special case. Goldman is right that a great deal of high-level mindreading is unconscious. And in all cases of unconscious high-level mindreading, the underlying processes might well be triggered automatically, without any unconscious aim or purpose playing a role. If this is how these processes work, then for each sort of mindreading in which successful episodes are subserved by a mental simulation, unsuccessful episodes will not count as simulations at all, even though they are produced by the same mental mechanisms. And if *that* is how things turn out, then once again it is hard to see why we should regard simulation as a natural or theoretically interesting category. To make a plausible case that mental simulation – as he defines it – is a theoretically interesting category in the study of high-level mindreading, Goldman must give us a convincing reason to suppose that unconscious aims and purposes abound in the processes subserving this sort of mindreading. Perhaps he can do this. But I don't recommend holding your breath until he does.

Reply to Sterelny

Sterelny's characteristically inventive and engaging paper discusses the theory that Shaun Nichols and I develop in *Mindreading*. However, his goal is not to evaluate our theory but to "parasitize" it by arguing that it strengthens the case for an anti-nativist account of mindreading of the sort he favors.(ms 7) As Sterelny notes, the exercise would be of no interest if our theory were hopelessly implausible; it's worth pursuing only if one agrees with Sterelny's assessment that our theory is "broadly correct". Though I had never realized it before, being parasitized is flattering!

In making his anti-nativist case, Sterelny concentrates on two sorts of innateness claims. The first, *concept innateness*, maintains that intentional concepts, like the concepts of belief and desire, are innate. The second, *information (or knowledge) innateness*, maintains that "information needed for mindreading is innate."(ms 8) In our book, Nichols and I argued that a number of the mental *mechanisms* that contribute to mindreading are innate, and that some of these mechanisms require a fair amount of innate information to get the job done. But we were quite reluctant to make any claims about the innateness of the concepts invoked in mindreading. One reason for our reticence was that we knew of no evidence or argument that made a convincing case either for or against the innateness of a specific concept. A second reason was that there is a great deal of controversy about the nature of concepts and also about how innateness claims should be understood, particularly when what is claimed to be innate is a concept or some similar mental state.[35] As a result, debates about the innateness of concepts are typically more than a bit obscure. Those who claim that some concept is (or is not) innate rarely tell us what they think concepts are or what they mean by 'innate', so what they are claiming is far from clear. Because of this, Nichols and I adopted the policy of avoiding the issue whenever possible.

Sterelny is a more intrepid scholar who boldly slogs into the swamp that Nichols and I were hesitant to approach. Though I am full of admiration for his courage, and would, of course, be delighted if it turned out that our theory could make a substantive contribution to the resolution of debates about innateness in this area, I am far from convinced that Sterelny has made much progress in his defense of anti-nativism either about intentional concepts or about the information invoked in mindreading. My reasons for skepticism are quite different in the two cases. In the case of concept nativism, it is the obscurity of the issues that leaves me unconvinced. In the case of information nativism, I think that Sterelny may have taken aim at a straw man.

[35] For the debate over the nature of concepts, see Margolis & Laurence (1999), Murphy (2002) and Machery (forthcoming). For the debate about how to interpret innateness claims, see Cowie (1999), Samuels (2002) and Khalidi (2007).

It is not entirely clear what conclusion Sterelny is defending in his discussion of concept nativism. In the section headed "The Poverty of the Stimulus," he ends his remarks on concept nativism with a relatively modest claim: "… I doubt there is a special, intractable problem of learning intentional concepts, despite the unobservability of intentional states." (ms 11) Perhaps the only conclusion Sterelny wants to draw is that poverty of the stimulus arguments, which attempt to show that intentional concepts could not be learned from the available evidence, and thus must be innate, are not convincing. Indeed, it is hard to see how such arguments *could* be convincing in light of the fact that Sterelny notes at the beginning of the sentence I've just quoted: "We lack a good general theory of the nature and acquisition of concepts.…". A few pages earlier he tells us, quite correctly, that "the whole issue of concepts and their possession is deeply opaque." (ms 9) If there is no agreement on what concepts *are*, no agreement about the mechanisms or processes that account for the acquisition of concepts, and little agreement about which sorts of acquisition mechanisms or processes would *count* as nativist or as anti-nativist (or empiricist), then surely it is simply premature to debate whether intentional concepts are innate, since we have no serious idea what we are debating. So if all that Sterelny wants to claim is that poverty of the stimulus arguments about intentional concepts are not convincing, he will get no grief from me.

However, Sterelny's brief comments about "iceberg concepts" – "concepts which name a syndrome that includes both a mental state and its distinctive behavioral manifestation"(ms 10) – might be interpreted as an attempt to provide at least a sketch of an anti-nativist account of how the concept of belief and concepts for other "more cryptic mental states" can be learned. Here's how the story goes: With iceberg concepts, "the intuitive gap between observable activity and our concepts of that activity are less wide. You can point to the distinctive manifestation of an itch. Moreover, an agent who has mastered the concept of an itch has mastered the concept of an internal cause of action."(ms 10) Once iceberg concepts have been mastered, they can "facilitate the acquisition of concepts for less overt states. For they prime an agent for the possibility of internal causes of action."(ms 10) All of this may be true. But I don't think it gets us very far. Though the "intuitive gap" may be less wide in the case of iceberg concepts, it is a gap all the same. We need an account of the mechanism or process that succeeds in crossing this gap by taking the observation of activity as input and producing a concept as output. We also need a motivated way of deciding whether proposed mechanisms and processes count as empiricist (because what they accomplish counts as *learning*) or nativist (because it doesn't). We also need an account of the mechanism or process that enables someone who is "primed"[36] for the possibility of internal causes of action to acquire[37] the concept of belief and of other cryptic states. Since Sterelny proposes no mechanisms or processes and offers no way of deciding which mechanisms are nativist and which are empiricist, I don't think he's made much progress in providing an anti-nativist theory of concept acquisition.

[36] Whatever, exactly, *that* means.

[37] Whatever, exactly, *that* means. Remember that "the whole issue of concepts *and their possession* is deeply opaque."

Let's turn, now, to Sterelny discussion of information innateness, "the idea that we have innate information about intentional states and their roles." It is there, Sterelny argues, that "the Stich-Nichols model helps greatly."

> The Stich-Nichols model reinforces a non-nativist conception of the development of mindreading by decomposing the construction of doxastic and desire worlds into subcomponents each of which can be both acquired and improved more or less independently of the others. Even on their hybrid picture, interpreting others is an information-rich task. But the information needed is of a kind that can be acquired and upgraded. (ms 11)

To illustrate what he has in mind, Sterelny sketches what Nichols and I call "default attribution" – a crucial aspect of mindreading in which the mindreader attributes many of her own beliefs to the target. Early on in development, we maintain, children attribute almost all of their beliefs to other people because they have only a few strategies available for detecting "discrepant" beliefs – beliefs that they do *not* share with the target. One of the first strategies of discrepant belief detection to emerge relies on an innate Perception Detection Mechanism (PDM) which uses information about a target and her environment to produce beliefs about the target's perceptions – beliefs like (1) *target did not see the chocolate being put into the box*. As Nichols and I note, "[o]bviously, a mechanism that can pull this off must be able to draw on a fair amount of information about the links between environmental situations and perceptual states." (Nichols & Stich, 2003, p. 88) By the time normal children are about 3½, they can use beliefs like (1) to infer beliefs like (2) *target does not believe that the chocolate was in the box,* which enable them to start building models of the target's beliefs in which they do not attribute all of their own beliefs to the target. Nichols and I don't offer any details on how the inference from (1) to (2) is made, though an obvious hypothesis is that it invokes an innate principle that says (roughly) *if x didn't see that p, then x doesn't know that p,* and that this principle gets hedged or modified in various ways as the child gets older.

It might seem a bit odd that Sterelny chooses the default attribution process to illustrate how the Nichols and Stich model lends support to his anti-nativist view, since on our account the early emerging parts of the default attribution system must have access to a fair amount of innate information. However, apparently this is compatible with the "non-nativist conception of the development of mindreading" that Sterelny advocates, since that conception includes the idea that "the development of mind reading is stabilized by quasi-perceptual mechanisms (by 'shallow' modules)" which do a variety of jobs including playing "an important role in estimating the doxastic world of other agents by factoring-in the role of differences in perceptual points of view."(ms 7) It is in the later stages of the development of the default attribution system, as Nichols and I characterize it, that Sterelny finds support for his anti-nativism. As children mature, we maintain, they acquire an increasingly sophisticated bag of tricks for adding discrepant beliefs to their model of the target's beliefs. Language facilitates many of these tricks. There is evidence that by the age of 3, children attribute the belief that p to a target when the target asserts that p, even when the child believes that p is false. They also attribute discrepant beliefs as the result of third person reports about what the target believes

(Nichols & Stich, 2003, pp. 91-2). As children get older they learn that not all assertions and third person reports can be trusted and they gradually develop a body of beliefs and skills that enable them to fine tune discrepant belief attribution. When Sterelny says that much of the information utilized in mindreading is not innate and can be learned piecemeal, it is *that* sort of information he has in mind. Also, as Sterelny notes, people gradually learn to exploit social cues in discrepant belief attribution. If you are knowledgeable about rugby and others treat a target as an expert on that sport, then you will attribute most of your beliefs about rugby to the target. But if you know that the target comes from a country where rugby is not popular, you will not attribute your beliefs about rugby to him. "The crucial point," Sterelny tells us, "is that these capacities can be built one by one: they are not a package deal. Moreover, each component can be improved gradually."(ms 11) There is no need to posit innate information to explain the extensive fine tuning of the default attribution system that goes on from age three into adulthood; familiar processes of learning are all that is required.

I think it is clear that Sterelny is right about this. What is less clear is who Sterelny takes his opponent to be in this debate. His comment about the capacities that fine tune default attribution not being a "package deal" suggests that the nativist opponent he has in mind thinks that mindreading is a unitary innate capacity. This interpretation is reinforced by his characterization of the view he is criticizing as "modular nativism" in which mindreading is "assimilated to the language model, and regarded as the result of an innate module."(ms 6) Later he tells us again that the on the view he is opposing, mindreading "is … language-like in having its cognitive basis in a module."(ms 13) If this "modular nativist" is indeed the opponent that Sterelny has in mind, there can be little doubt that Sterelny has his opponent on the ropes. That's the good news. The bad news is that Sterelny's opponent may be a straw man.

Sterelny says very little about who he thinks actually advocates modular nativism. But there is some reason to think that one of the people he has in mind is my Rutgers colleague, Alan Leslie, who is widely regarded as one the main defenders of the nativist approach to mindreading, and as *the* main defender of a modular theory of mindreading. Leslie is also the only major figure in mindreading research who is quoted in Sterelny's paper. However, as Leslie makes clear in many places (including the Scholl & Leslie paper that Sterelny quotes), he does *not* think that mindreading is a package deal, nor does he think that all of mindreading is the result of an innate module. On Leslie's theory, adult mindreading is subserved by three distinct systems, only one of which is clearly modular. The modular component, which Leslie and his collaborators sometimes call "ToMM" (the Theory of Mind Mechanism) is a module that is responsible for the mindreading skills that emerge early in development.

> ToMM … is essentially a module which spontaneously and post-perceptually attends to behaviors and infers (i.e. computes) the mental states which contributed to them…. As a result, ToMM will provide the child with early intentional insight into the behaviors of others." (Scholl & Leslie, 1999, p. 147)

One of the important roles that ToMM plays is to subserve the sort of unrestricted default belief attribution that is typical of children under 3. In attributing beliefs to a target, "ToMM always makes the current situation [i.e. what the mindreader *believes* the current situation to be] available as a possible and even preferred content."(Scholl & Leslie, 1999, p. 147) So it is ToMM that is responsible for the poor performance of young children on the false belief task. Somewhat later in development, Leslie maintains, another system, the Selection Processor (SP) comes on line. The job of SP, which Scholl and Leslie tell us "may be non-modular" (Scholl & Leslie, 1999, p. 147), is to determine the correct content to attribute when a target's belief is in conflict with the mindreader's, and to override ToMM's inclination to attribute the mindreader's own beliefs. Finally, still later in development, the mindreading abilities that ToMM and SP make available "are recruited by higher cognitive processes for more complex tasks, and the resulting higher-order [mindreading] activities may well interact (in a non-modular way) with other cognitive processes, and may not be uniform across individuals or cultures." (Scholl & Leslie, 1999, p. 140)

What is striking about this account, for our purposes, is how different it is from the "modular nativism" that Sterelny is criticizing, and how similar it is to the account Sterelny defends. On Leslie's theory there is no innate module that subserves all mindreading skills. Moreover, ToMM, the one component of the mindreading system that Leslie insists is both innate and modular, could well be described in the same words that Sterelny uses to describe the innate modules posited by his account: it is a "quasi-perceptual shallow module". What is more, ToMM's job is *more modest* than the jobs that Sterelny's shallow modules perform.[38] The "higher-order" mindreading activities that emerge in the third stage of Leslie's theory presumably exploit lots of information that is not innate since, like much of what is learned, it varies across individuals and cultures. So Leslie has no trouble accommodating the ways in which beliefs about expertise in rugby, or other beliefs that are acquired by empiricist learning strategies, can influence mindreading. The upshot of all of this is that Leslie does not advocate the sort of modular nativism that Sterelny is concerned to refute. Nor, to the best of my knowledge, does anyone else who is worth refuting.

[38] Sterelny's shallow modules "play an important role in estimating the doxastic world of other agents by factoring-in the role of differences in perceptual points of view." (ms 7) On Leslie's theory this is presumably done not by ToMM, but by SP.

Reply to Prinz

Prinz's engaging and fact filled paper focuses on moral nativism, a view – or more accurately a rather tangled cluster of views – that has become increasingly prominent in recent, empirically informed discussions of moral psychology. (Dwyer, 1999, 2006; Harman, 1999; Hauser, 2006; Makhail, in press). His paper is primarily aimed at sketching and defending his own provocative and important thesis that morality is "an accident" – "a byproduct of capacities that were evolved for other purposes" (ms 2). Toward the end of his paper, Prinz suggests that, to a considerable degree, his account is compatible with the theory that Sripada and I developed in "A Framework for the Psychology of Norms"(S&S). Here is what he says:

> The general outlook defended in this discussion closely parallels ideas defended by Sripada and Stich (2006). Like those authors, I have argued that moral judgments are not universal across cultures, despite some similarities, and I have argued that emotions play an important role in acquisition, and implementation of moral norms. Like them, I have also explored these ideas with an interest in explaining how moral norms are acquired. Sripada & Stich are agnostic about how moral norms differ from other norms, and they think we are not yet in a position to determine how much innate machinery we need to explain the acquisition of moral norms. I have been less agnostic, arguing explicitly against moral nativism. Even if I am right, Sripada and Stich raise an interesting question in their discussion. Supposing there is no innate mechanism for *moralization*, might there be a more general mechanism for the acquisition of norms? Sripada and Stich suppose there is….. I think the postulation of such a mechanism is premature and methodologically risky." (ms 17-18)

Is this an accurate summary of the points on which S&S and Prinz agree and disagree? I am not convinced that it is. The problem is not that Prinz *misdescribes* the S&S view – he is far to careful a scholar for that.[39] Rather, what concerns me is that I am not at all clear about what Prinz *means* when he makes claims about morality, moral judgments, moral norms and the like. I'll devote most of my reply to elaborating on this theme, since I think it is a manifestation of a much larger and more serious problem that besets a great deal of recent discussion in empirically informed moral psychology.

"Moral norms," Prinz tells us, "are found in almost every recorded human society…. [Morality] seems to be a human universal." This leads some to conclude that "morality is an evolved capacity" and that "morality is innate." (ms 1 & 2) These are the views that Prinz is arguing against. Before one plunges into this debate, there are two clusters of questions that cry out for answers. In the first cluster are questions like: What is it for a capacity to be *innate*? and What is it for a capacity to be *an evolved capacity*? In the second cluster are questions like: What is *morality*? and Which capacities are

[39] Though there is one minor slip in the quoted passage. S&S did not endorse the view that emotions play an important role in the *acquisition* of norms, since we could find no persuasive evidence either for or against the view.

moral capacities? Most of the philosophers and psychologists involved in recent debates over moral nativism are aware that there is an extensive and sophisticated philosophical literature aimed at answering the first cluster of questions. And many of these authors offer at least a brief account of how they propose to answer them.[40]

Until quite recently, however, most of these philosophers and psychologists seemed to assume that the answers to questions in the second cluster were obvious. They indicated almost no awareness that there is a large and contentious philosophical literature aimed at answering these questions. Indeed, more than 50 years ago, Alasdair MacIntyre (1957) began an article called "What Morality Is Not" with the following sentence: "The central task to which contemporary moral philosophers have addressed themselves is that of listing the distinctive characteristics of moral utterances." MacIntyre went on to argue that two of the most widely endorsed "distinctive characteristics" – that they are "universalizable" and that they are "prescriptive" – are not, in fact, necessary features of moral utterances. In 1970, MacIntyre's article was reprinted in a valuable anthology called *The Definition of Morality* (Wallace & Walker, 1970) which also reprinted a dozen other papers by such leading figures as Elizabeth Anscombe, Kurt Baier, Philippa Foot, William Frankena and Peter Strawson, all of which, in one way or another, tackled the question of how morality is best defined. As one might expect from this distinguished array of authors, many of the arguments to be found in the volume are careful and sophisticated. And as one might expect in just about any group of 13 philosophers, no consensus was reached. Nor has there been any convergence on the issue in the decades since then.[41] In addition to debating how the notions of moral utterance, moral rule and moral norm are to be defined, some of the contributors to the *Definition of Morality* volume, as well as some more recent authors, have discussed a cluster of meta-philosophical questions including: What is a definition of morality supposed to *do*? and What counts as getting the definition *right*?[42] Though no consensus has emerged on these questions either, two sorts of answers are particularly important for our purposes.

The first is that in seeking a definition of morality, philosophers are engaged in the venerable endeavor of linguistic or conceptual analysis. They are trying to give an account of how the term 'moral' (or the expressions 'moral utterance', 'moral rule' etc.)

---

[40] For the record, here is what Sripada and I say on the matter:

> Though there is a large philosophical literature debating the best interpretation of innateness claims in psychology (Cowie, 1999; Griffiths, 2002; Samuels, 2002), for our purposes we can consider a normative rule to be innate if various genetic and developmental factors make it the case that the rule would emerge in the norm database in a wide range of environmental conditions, even if (as a result of some extraordinary set of circumstances) the child's "cultural parents" – the people she encounters during the norm acquisition process – do *not* have the norm in *their* norm data base. If there were innate norms of this sort then they would almost certainly be cultural universals. Barring extraordinary circumstances, we should expect to find them in all human groups. (S&S, p. 299, emphasis in the original)

[41] For a useful discussion of some of the literature, see Gert (2005).

[42] For a particularly useful discussion, see Taylor (1978).

are used by ordinary English speakers – or perhaps by some subset of speakers, for example those with a modicum of philosophical sophistication. A successful definition would have to comport with the intuitions of the relevant group of speakers on a wide range of actual and hypothetical cases.

The second answer is that the goal of the project is to discover what Taylor (1978) calls "the essence of morality". Philosophers who pursue this project believe that moral utterances or moral rules constitute a *natural kind*, and their goal is to discover the nature of this kind – the property or cluster of properties in virtue of which utterances or rules are members of the kind. The methodology, for those pursuing this project, is akin to the one sketched by Devitt in his contribution to this volume. We rely on intuition to pick out some clear and obvious members of the kind in question and some clear and obvious cases that are not member of the kind. We then turn to science to determine the property or properties that all (or at least almost all) of the intuitively clear members share and that all (or almost all) of the intuitively clear non-members do not. Not just any property or cluster of properties will do, however. To constitute a natural kind, there will have to be some theoretically interesting nomological generalizations in which the properties are invoked. Though it relies on intuition in its initial stages, this project is ultimately much less beholden to intuition. Once we have succeeded in determining the essential feature(s) shared by most of the intuitively obvious members of the kind, we can simply reject the urgings of intuition on lots of other cases, in much the same way that biologists rejected people's intuitive judgments that whales and dolphins are fish.

There is, of course, no guarantee that either the conceptual analysis project or the quest for the essence of morality will proceed smoothly. As Jerry Fodor has noted, "[i]t seems…to be among the most important findings of philosophical and psychological research over the last several hundred years … that attempts at conceptual analysis almost always fail." (Fodor, 1981) And while history has been kinder to the quest for natural kinds, it is sometimes the case that more than one natural kind is to be found among the intuitively clear examples. When there are just a few, as in the case of jade, there is often no motivated way of deciding which kind merits the pre-existing term. When there are many, as in the case of earth (the putative "element," not the planet), the most natural conclusion is that the intuitive category does not pick out a natural kind at all.

With this as background, let's turn to Prinz's account of morality. Here is what he tells us:

> Moral norms are a subset of norms, distinguished by their moral character.
>
> There are various theories of what moral character consists in…. According to some theories, moral norms are distinguished by their subject matter; according to other theories, they are distinguished by the procedures by which they are discovered or the reasons by which they are justified; according to a third class of theories, moral norms are distinguished by the particular way in which they are psychologically internalized and enforced. I subscribe to a theory of this last kind.

> I think a moral norm is a norm that is enforced by certain emotions (see Prinz, 2004). This view has been known historically as sentimentalism. Roughly, a person regards something as morally wrong (impermissible) if, on careful consideration, she would feel emotions of disapproval towards those who did the thing in question….
>
> There are a number of different emotions of disapproval. Core examples include guilt, shame, disappointment, resentment, anger, indignation, contempt, and disgust. There is evidence that different kinds of moral rules recruit different emotions of disapproval (Rozin et al. 1999). We feel anger toward those who harm others, contempt towards those who disrespect members of other social ranks, and disgust towards those who commit incest. If we harm another person, we feel guilty, and if we violate norms of rank or incest, we feel ashamed. I cannot defend these claims here (see Prinz, 2005). I will merely point out three relevant facts. First, emotion structures in the brain are active when people make moral judgments (Haidt and Greene, 2003). Second, people who are profoundly deficient in emotions (psychopaths), never develop a true comprehension of morality (Blair, 1999). Third, if we encountered a person who claimed to find killing (or stealing, or incest, etc.) morally wrong but took remorseless delight in killing and in hearing tales of other people killing, *we could rightfully accuse him of speaking disingenuously*. All this suggests that *emotional response is essential to moral cognition*. Norms that are not implemented by emotions of disapproval are not moral norms. (ms 12-13, italics added)

Prinz does not tell us whether he intends his account as a conceptual analysis or as a hypothesis about the essential features of a natural kind. The first of the two passages I have italicized toward the end of the long quote might be taken as a bit of evidence for the first interpretation. But I think that would be an uncharitable reading, since Prinz's version of sentimentalism has many consequences which fly in the face of intuition – or at least *my* intuition. Many violations of etiquette norms evoke disgust on the part of observers and shame on the part of the transgressor. And many violations of religious norms evoke both anger and disgust in co-religionists and guilt in the transgressor. Even when they are not closely tied to religion, transgressions of food taboos and norms for disposing of dead bodies can evoke very strong emotions of disapproval. However, for me and the four other upper middle class white males that I've consulted (adhering rigorously to the standard method of philosophical conceptual analysis) neither etiquette norms nor religious norms are intuitive examples of moral norms, nor are food taboos or burial norms.

The second passage I've italicized suggests that Prinz intends to be characterizing a natural kind whose extension overlaps significantly with the intuitive extension of 'moral norm' and that in so doing he has characterized the essential features of moral norms. However, if that's his intention, another problem looms, since even if he has succeeded in capturing the essential features of a natural kind, there may well be several other natural kinds in this vicinity. By far the best known candidate is to be found in the work of Elliott Turiel and his associates. Inspired by some of the philosophical literature

aimed at providing a definition of morality, Turiel proposed a definition according to which moral rules are those that are authority independent, universally applicable and justified by appeal to harm, justice or rights. (Turiel, 1979; Turiel, 1983; Turiel et al. 1987) With this definition in hand, he went on to design an experimental paradigm (the "moral / conventional task") which has been administered to a wide variety of subjects differing in age, religion and nationality. On one interpretation, the results of Turiel's experiments show that moral rules, as he defines them, constitute a natural kind, since the properties invoked in the definition form a robust nomological cluster.[43] I am quite skeptical about all this because I think there is abundant evidence that the putative nomological cluster is far from robust; when researchers have looked at norms and transgressions outside the narrow range of schoolyard examples that have been the focus of Turiel and his associates, the elements of the cluster come apart.[44] But it is not clear that Prinz shares my skepticism, since he endorses Blair's claim that psychopaths "never develop a true comprehension of morality," and Blair uses the moral / conventional task to assay whether his subjects have a true comprehension of morality. If I am mistaken and Turiel's work does indeed pick out a natural kind that overlaps substantially with the intuitive extension of 'moral norm', and if Prinz maintains that his sentimentalist account captures the essential features of moral norms, then Prinz owes us some additional argument to justify the claim that his account rather than Turiel's tells us what is "essential to moral cognition."

Of course, if I am right that Turiel's account does not succeed in picking out a natural kind, this problem disappears. However, there is another competitor that I am less inclined to dismiss. In "A Framework for the Psychology of Norms" Sripada and I use the term 'norm' for what we argue is "a theoretically important natural kind in the social sciences" (p. 281) We make it clear that our account of norms "is *not* intended as a *conceptual analysis* or as an account of what the term 'norm' means to ordinary speakers." (p. 281) Rather, our strategy is to give a rough and ready characterization of the kind that will enable us to pick out clear cases, and then to offer a first pass at an empirically informed theory about a psychological mechanism that can explain some of the more striking features of these cases. If the theory is on the right track, "a better account of the crucial features of norms can be expected to emerge as that theory is elaborated." (p. 281) One of the components in our theory is a norm database, and it is the job of the theory to tell us what can and cannot end up in that database. In so doing, the theory will give us an increasingly informative account of the natural kind that we call 'norms'.

A natural question to ask about the S&S theory is: what is the relation between norms, as we characterize them, and the intuitive category of moral norms. Here is what we say on the matter.

[43] For further discussion of this interpretation of Turiel, see Kelly & Stich (2007).

[44] Kelly et al. (2007), Kelly & Stich (2007).

> It … strikes us as quite likely that the *intuitive* category of *moral* norms is not co-extensive with the class of norms that can end up in the norm data base posited by our theory. Perhaps the most obvious mismatch is that the norm data base, for many people in many cultures, will include lots of rules governing what food can be eaten, how to dispose of the dead, how to show deference to high ranking people, and a host of other matters which our commonsense intuition does not count as moral. (p. 291, emphasis in the original)

But as I noted earlier, there is a roughly parallel mismatch between the intuitive category of moral norms and Prinz's sentimentalist account moral norms. If this mismatch does not prevent Prinz's account from being a proposal about the essential features of moral norms, then the S&S account can also be construed as a proposal about the essential features of moral norms.[45] But if is construed in this way, Prinz is mistaken in claiming that "[S&S] are agnostic about how moral norms differ from other norms, and they think we are not yet in a position to determine how much innate machinery we need to explain the acquisition of moral norms." Rather, the S&S account should be read as claiming that moral norms are a natural kind which is *identical* with the norms as characterized by our theory. And since our theory posits a fair amount of innate machinery, we have a clear disagreement with Prinz about how much innate machinery is required to explain the acquisition of moral norms.

Do we now have an accurate account of the relation between the S&S theory and Prinz's sentimentalist theory? I am still not confident that we do. The discussion in the last two paragraphs began with the assumption "that Prinz intends to be characterizing a natural kind whose extension overlaps significantly with the intuitive extension of 'moral norm' and that in so doing he has characterized the essential features of moral norms." But it is far from clear that this is correct. Recall that, for Prinz, morality is "an accident". Moral transgressions are just actions which, when performed by others happen to trigger one or more of a grab bag of emotions – disappointment, resentment, anger, indignation, contempt, or disgust – and when performed by oneself trigger one or more of another grab bag of emotions – primarily guilt or shame. There are, of course, lots of actions which trigger emotions in the first cluster when someone else does them, but which do not trigger guilt or shame when we do them ourselves.[46] And, though it is less clear cut, there are probably lots of actions which trigger guilt or shame when we do them, but don't trigger emotions in the first cluster when others do them. These sorts of

---

[45] Sripada and I made it clear that this is one way of understanding the relation between *moral* norms and norms as characterized by our theory.

> [One] possibility is that moral rules might turn out to constitute a natural kind that is identical with the norms characterized by our theory. On this view, our intuitions about which rules are moral are sometimes simply mistaken, in much the same way that the folk intuition that whales are kind of fish was mistaken. (S&S p. 291)

[46] For example, when a university official in charge of distributing research grants decides not to award me a grant, her act triggers disappointment in me. But when I am in charge of distributing research grants and I decide not to award one to a colleague, that action does not lead me to feel either guilt or shame. It would be easy enough to generate countless additional examples.

dissociations should not be surprising on Prinz's account, since he thinks that "moralization takes place under cultural pressure…. Our moral educators tell us that we should feel bad when we hurt each other or take things that aren't ours. They teach us by example to get angry at those who violate these norms, even when we are not directly involved. Moralization inculcates emotions of disapproval." (ms 13) If this is right, then dissociations will occur whenever an individual is taught to feel anger in response to an action, but is not taught to feel guilt, or vice versa. Moreover, the teaching itself can occur in lots of ways. "We receive a lot of moral instruction through explicit rules, sanctions, story telling, role models, and overt attitudes expressed by members of our communities." (ms 16) In light of the heterogeneity in the kinds of emotions involved, the many different ways in which they can be linked to categories of action, and the fact that a type of action can trigger emotions in one cluster without triggering emotions in the other, it is hard to see how moral norms, as Prinz characterizes them, could be natural kinds. Moral norms, on Prinz's account, don't seem to be suitable candidates for being invoked in nomological generalizations.

If that's right, then Prinz faces a menu of unpalatable alternatives. If his sentimentalist account of moral norms is intended to characterize a natural kind whose extension overlaps with the intuitive extension of 'moral norm', then there is reason to think he has failed. Accidents don't make good natural kinds. If his sentimentalist account was intended to capture the intuitive extension of 'moral norm', then again there is reason to think he has failed, since there are *lots* of cases, including food taboos, etiquette norms, fashion norms and burial norms which intuition (or at least *my* intuition) does not classify as moral norms though Prinz's sentimentalist account does. Of course, it may be that Prinz does not intend his sentimentalist theory to be either a conceptual analysis or an account of a natural kind. If that's the case, then he owes us some account of what he *is* trying to do. It is hard to see how the merits of his theory can be assessed without some guidance on what the theory is supposed to do or what counts as getting it right. And, to return to the concern with which we began, it is also hard to know whether and where his theory and the S&S theory disagree.

The problems I've posed for Prinz's view can all be traced to the fact that he has not told us enough about how he is using terms like "moral norm" and "moral rule" and that makes it all but impossible to evaluate his claims about morality. I have focused on this issue because, as I noted earlier, I think that much the same problem lies behind many debates in moral psychology. Consider, for example, the following provocative claim in the Introduction to a recent paper by Jonathan Haidt and Craig Joseph.

> [T]he psychological study of morality, like psychology itself…, has been dominated by politically liberal researchers (who includes us). The lack of moral and political diversity among researchers has led to an inappropriate narrowing of the moral domain to issues of harm/care and fairness/reciprocity/ justice. Morality in most cultures (and for social conservatives in Western cultures), is in fact much broader, including issues of ingroup/loyalty, authority/ respect, and purity/sanctity….

> This chapter is about how morality might be partially innate …. We begin by arguing for a broader conception of morality and suggesting that most of the discussion of innateness to date has not been about morality per se; it has been whether the psychology of *harm* and *fairness* is innate. (Haidt & Joseph, 2007, p. 367)

To make their case for a broader conception of morality, Haidt and Joseph offer a brief overview of norms that prevail in cultures other than our own which include "rules about clothing, gender roles, food, and forms of address" (Haidt & Joseph, 2007, p. 371) and a host of other matters as well. They emphasize that people in these cultures care deeply about whether or not others follow these rules. But this is an odd way to proceed. For surely Haidt and Joseph don't think that the "politically liberal researchers" responsible for the "inappropriate narrowing" of the moral domain are *unaware* that rules governing these matters are widespread in other cultures. The issue in dispute is not whether rules like these exist or whether people care about them. What is in dispute is whether these rules are *moral* rules. To resolve that dispute, we need an answer to the question that is center stage in the philosophical literature on the definition of morality – we need an account of what it is for a rule to be a moral rule. And if the dispute between Haidt and Joseph and those they criticize is substantive, then not just any account will do; it has to be a *correct* account. But what counts as getting such an account right? That is a question that loomed large in my critique of Prinz, and it is, I suggest, is a question that needs to be addressed by just about everyone who makes claims about moral nativism.

References

Bishop, M. and Stich, S. 1998: "The Flight to Reference, or How *Not* to Make Progress in the Philosophy of Science," *Philosophy of Science*, Vol. 65, pp. 33-49.

Bishop, M. and Trout, J. D. 2005: *Epistemology and the Psychology of Human Judgment*, New York: Oxford University Press.

BonJour, L. 1985: *The Structure of Empirical Knowledge*, Cambridge, MA: Harvard University Press.

Carruthers, P., Laurence, S. and Stich, S. (eds.): *The Innate Mind: Structure and Contents*, New York: Oxford University Press.

Carruthers, P., Laurence, S. and Stich, S. (eds.): *The Innate Mind: Culture and Cognition*, New York: Oxford University Press.

Carruthers, P., Laurence, S. and Stich, S. (eds.): *The Innate Mind: Foundations and the Future*, New York: Oxford University Press.

Chisholm, R. 1977: *Theory of Knowledge*, Englewood Cliffs, NJ: Prentice Hall.

Clement, J. 1983: "A Conceptual Model Discussed by Galileo and Used Intuitively by Physics Students," in D. Gentner and A. Stevens (eds.), *Mental Models*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 325-339.

Cowie, F. 1999: *What's Within? Nativism Reconsidered*, New York: Oxford University Press.

Cummins, R. 1989: *Meaning and Mental Representation*, Cambridge, MA: MIT Press.

Cummins, R. 1998: "Reflection on Reflective Equilibrium," in M. DePaul, and W. Ramsey (eds.), *Rethinking Intuition*, Lanham, Maryland: Rowman and Littlefield, pp. 113-127.

DePaul, M. (forthcoming): "Ugly Analyses and Value," to appear in D. Pritchard, A. Millar and A. Haddock (eds.), *Epistemic Value*, Oxford: Oxford University Press.

Devitt, M. 1981: *Designation*, New York: Columbia University Press.

Devitt, M. 1996: *Coming to Our Senses: A Naturalistic Program for Semantic Localism*, New York: Cambridge University Press.

Devitt, M. 2006: *Ignorance of Language*, New York: Oxford University Press.

Devitt, M. and Sterelny, K. 1999: *Language and Reality: An Introduction to the Philosophy of Language*, 2nd edition (1st edition 1987), Oxford: Blackwell.

Dretske, F. 1981: *Knowledge and the Flow of Information*, Cambridge, MA: MIT Press.

Dwyer, S. 1999: "Moral Competence," in K. Murasugi and R. Stainton (eds.), *Philosophy and Lingusitics*, Boulder, Colorado: Westview Press, pp. 196-190.

Dwyer, S. 2006: "How Good Is the Linguistic Analogy?" in P. Carruthers, S. Laurence and S. Stich (eds.), *The Innate Mind, Volume 2, Culture and Cognition*, Oxford: Oxford University Press, pp. 237-256.

Egan, F. 1992: "Individualism, Computation, and Perceptual Content," *Mind*, Vol. 101, pp. 443-459.

Egan, F. 1995: "Content and Computation," *Philosophical Review*, Vol. 104, pp. 181-203.

Egan, F. 1999: "In Defense of Narrow Mindedness," *Mind and Language*, Vol. 14, pp. 177-194.

Egan, F. 2003: "Naturalistic Inquiry: Where Does Mental Representation Fit In?" in L. Antony and N. Hornstein (eds.), *Chomsky and His Critics*, Oxford: Blackwell, pp. 89-104.

Field, H. 1986: "The Deflationary Concept of Truth," in G. MacDonald and C. Wright (eds.), *Fact, Science and Value*, Oxford: Blackwell.

Field, H. 1994: "Deflationist Views of Meaning and Content," *Mind*, Vol. 103, pp. 249-285.

Fodor, J. 1981: "The Present Status of the Innateness Controversy," in *Representations*, Cambridge, MA: MIT Press, pp. 257-316.

Fodor, J. 1987: *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*, Cambridge, MA: MIT Press.

Fodor, J. 1990: "A Theory of Content, I and II" in *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press, pp. 51-136.

Fodor, J. 1991: "Reply to Stich," in B. Loewer and G. Rey (eds.), *Meaning in Mind: Fodor and His Critics*, Cambridge, MA: Blackwell, pp. 310-312.

Fodor, J. 1998: *Concepts*, Oxford: Oxford University Press.

Gallese, V. and Goldman, A. 1998: "Motor Neurons and the Simulation Theory of Mind-Reading," *Trends in Cognitive Science*, Vol. 2, 12, pp. 493-501.

Gert, B. 2005: "The Definition of Morality," *The Stanford Encyclopedia of Philosophy (Fall 2005 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2005/entries/morality-definition/>.

Gilovich, T., Griffin, D. and Kahneman, D. (eds.) 2002: *Heuristics and Biases: The Psychology of intuitive Judgment*, Cambridge: Cambridge University Press.

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

Goldman, A. 1989: "Interpretation Psychologized," *Mind and Language*, Vol. 4, pp. 161-185.

Goldman, A. 1992: "In Defense of the Simulation Theory," *Mind and Language* Vol. 7, pp. 104-119.

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

Goldman, A. 2006: *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*, New York: Oxford University Press.

Gordon, R. 1986: "Folk Psychology as Simulation," *Mind and Language*, Vol. 1, pp. 158-170.

Griffiths, P. 2002: "What is Innateness?" *Monist*, Vol. 85, pp. 70-85.

Haidt, J. and Joseph, C. 2007: "The Moral Mind: How 5 Sets of Innate Moral Intuitions Guide the Development of Many Culture-Specific Virtues, and Perhaps Even Modules," in P. Carruthers, S. Laurence and S. Stich (eds.), *Innateness and the Structure of the Mind, Volume III, Foundations and the Future*, New York: Oxford University Press, pp. 367-391.

Harman, G. 1999: "Moral Philosophy and Linguistics," in K. Brinkmann (ed.), *Proceedings of the 20th World Congress of Philosophy*, Vol. I: *Ethics*, Bowling Green, Ohio: Philosophy Documentation Center, pp. 107-115.

Hauser, M. 2006: *Moral Minds*, New York: Echo Press.

Hawthorne, J. 2004: *Knowledge and Lotteries*, Oxford: Oxford University Press.

Heal, J. 1986: "Replication and Functionalism," J. Butterfield (ed.), *Language, Mind and Logic*, Cambridge: Cambridge University Press.

Horwich, P. 1990: *Truth*, Oxford: Blackwell.

Jackson, F. 1998: *From Metaphysics to Ethics: A Defence of Conceptual Analysis*, Oxford: Oxford University Press.

Jackson, F., Mason, K. and Stich, S. (forthcoming): "Implicit Knowledge and Folk Psychology," to appear in D. Braddon-Mitchell and R. Nola (eds.), *Conceptual Analysis and Philosophical Naturalism*, Cambridge, MA: MIT Press.

Kahneman, D., Slovic, P. and Tversky, A. (eds.), 1982: *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press.

Kelly, D., Stich, S., Haley, K., Eng, S. and Fessler, D. 2007: "Harm, Affect and the Moral / Conventional Distinction," *Mind and Language*, Vol. 22, April 2007, pp. 117-131.

Kelly, D. and Stich, S. 2007: "Two Theories About the Cognitive Architecture Underlying Morality," in P. Carruthers, S. Laurence and S. Stich (eds.), *Innateness and the Structure of the Mind, Volume III, Foundations and the Future*, New York: Oxford University Press, pp. 348-366.

Khalidi, M. 2007: "Innate Cognitive Capacities," *Mind and Language*, Vol. 22, pp. 92-115.

Kitcher, P. 1993: *The Advancement of Science*, Oxford: Oxford University Press.

Kornblith, H. 2002: *Knowledge and its Place in Nature*, Oxford: Oxford University Press.

Kripke, S. 1972: "Naming and Necessity," in D. Davidson and G. Harman (eds.), *Semantics of Natural Language*, Dordrecht, The Netherlands: Reidel, pp. 253-355.

Lycan, W. 1988: *Judgement and Justification*, Cambridge: Cambridge University Press.

MacIntyre, A. 1957: "What Morality Is Not," *Philosophy*, Vol. 32, pp. 325-335.

Machery, E., Mallon, R., Nichols, S. and Stich, S. 2004: "Semantics, Cross-Cultural Style," *Cognition*, Vol. 92, pp. B1- B12.

Machery, E. (forthcoming): *Doing Without Concepts*, New York: Oxford University Press.

Makhail, J. (in press): *Rawls' Linguistic Analogy*, Cambridge: Cambridge University Press.

Margolis, E. and Laurence, S. (eds.), 1999: *Concepts: Core Readings*, Cambridge, MA: MIT Press.

McCloskey, M. 1983: "Naive Theories of Motion," in D. Gentner and A. Stevens (eds.), *Mental Models*, Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 299-324.

Millikan, R. 1984: *Language, Thought, and Other Biological Categories: New Foundations for Realism*, Cambridge, MA: MIT Press.

Mitchell, J., Banaji, M. and Macrae, C. 2005: "The Link Between Social Cognition and Self-Referential Thought in the Medial Prefrontal Cortex," *Journal of Cognitive Neuroscience*, Vol. 17, pp. 1306-1315.

Murphy, G. 2002: *The Big Book of Concepts*, Cambridge, MA: MIT Press.

Nichols, S. and Stich, S. 1998: "Rethinking Co-cognition," *Mind and Language*, Vol. 13, pp. 499-512.

Nichols, S. and Stich, S. 2003: *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*, Oxford: Oxford University Press.

Nichols, S., Stich, S. and Weinberg, J. 2003: "Meta-Skepticism: Meditations on Ethno-Epistemology," in S. Luper (ed.), *The Skeptics*: *Contemporary Essays*, Aldershot, U.K.: Ashgate Publishing, pp. 227-247.

Nisbett, R. 2003: *The Geography of Though: How Asians and Westerners Think Differently … and Why*, New York: The Free Press.

Papineau, D. 1987: *Reality and Representation*, Oxford: Blackwell.

Pessin, A. and Goldberg, S. (eds.) 1996: *The Twin Earth Chronicles: Twenty Years of Reflection on Hilary Putnam's "The Meaning of Meaning,"* Armonk, N.Y.: M. E. Sharpe, Inc.

Pollock, J. and Cruz, J. 1999: *Contemporary Theories of Knowledge*, Lanham, Maryland: Rowman and Littlefield.

Putnam, H. 1975: "The Meaning of 'Meaning'," in K. Gunderson (ed.), *Language, Mind and Knowledge: Minnesota Studies in the Philosophy of Science*, Vol. 7, Minneapolis: University of Minnesota Press.

Quine, W. 1960: *Word and Object*, Cambridge, MA: MIT Press.

Ramsey, W. 2007: *Representation Reconsidered*, Cambridge: Cambridge University Press.

Ross, L. and Nisbett, R. 1991: *The Person and the Situation*, New York: McGraw-Hill.

Scholl, B. and Leslie, B. 1999: "Modularity, Development and 'Theory of Mind'," *Mind and Language*, Vol. 14, pp. 131-153.

Sosa, E. 2007a: "Experimental Philosophy and Philosophical Intuition," *Philosophical Studies*, Vol. 132, pp. 99-107.

Sosa, E. 2007b: *A Virtue Epistemology*, New York: Oxford University Press.

Sosa, E. (forthcoming): "Intuitions: Their Nature and Epistemic Efficacy" to appear in *Grazer Philosophische Studien*, special issue *Philosophical Knowledge - Its Possibility and Scope*, ed. by C. Beyer and A. Burri, Amsterdam: Rodopi.

Sosa, E. (this volume). "A Defense of the Use of Intuitions in Philosophy. "

Samson, D., Apperly, I., Kathirgamanathan, U. and Humphreys, G. 2005: "Seeing It My Way: A Case of Selective Deficit in Inhibiting Self-Perspective," *Brain*, Vol. 128, pp. 1102-1111.

Samuels, R. 2002: "Nativism in Cognitive Science," *Mind and Language*, Vol. 17, pp. 233-265.

Sripada, C. and Stich, S. 2006: "A Framework for the Psychology of Norms," in P. Carruthers, S. Laurence and S. Stich (eds.), *The Innate Mind, Volume 2, Culture and Cognition*, Oxford: Oxford University Press, pp. 280-301.

Stich, S. 1983: *From Folk Psychology to Cognitive Science: The Case Against Belief*, Cambridge, MA: MIT Press.

Stich, S. 1990: *The Fragmentation of Reason*, Cambridge, MA: MIT Press.

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

Stich, S. 1988: "Reflective Equilibrium, Analytic Epistemology and the Problem of Cognitive Diversity," *Synthese*, Vol. 74, pp. 391-413. Reprinted in M. DePaul and W. Ramsey (eds.), *Rethinking Intuition*, Lanham, Maryland: Rowman and Littlefield) 1998, pp. 95-112. Page references are to the DePaul and Ramsey volume.

Stich, S. and Nichols, S. 1992: "Folk Psychology: Simulation or Tacit Theory," *Mind and Language*, Vol. 7, pp. 35-71.

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

Stich, S. and Weinberg, J. 2001: "Jackson's Empirical Assumptions," *Philosophy and Phenomenological Research*, Vol. 62, pp. 637-643.

Swain, S., Alexander, J., and Weinberg J. (forthcoming): "The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp," to appear in *Philosophy and Phenomenological Research*.

Taylor, P. 1978: "On Taking the Moral Point of View," in P. French, T. Uehling and H. Wettstein (eds.), *Midwest Studies in Philosophy*, Vol. 3, *Studies in Ethical Theory*, Morris, Minnesota: University of Minnesota.

Turiel, E. 1979: "Distinct Conceptual and Developmental Domains: Social Convention and Morality," in H. Howe, and C. Keasey (eds.), *Nebraska Symposium on Motivation, 1977: Social Cognitive Development*, Lincoln: University of Nebraska Press, Vol. 25, pp. 77-116.

Turiel, E. 1983: *The Development of Social Knowledge*, Cambridge: Cambridge University Press.

Turiel, E. and Nucci, L. 1978: "Social Interactions and the Development of Social Concepts in Preschool Children," *Child Development*, Vol. 49, pp. 400-407.

Turiel, E., M. Killen, and Helwig, C. 1987: "Morality: It's Structure, Functions, and Vagaries," in J. Kagan and S. Lamb (eds.), *The Emergence of Morality in Young Children*, Chicago: The University of Chicago Press.

Wallace, G. and Walker, A. (eds.) 1970: *The Definition of Morality*, London: Methuen.

Weinberg, J., Nichols, S. and Stich, S. 2001: "Normativity and Epistemic Intuitions," *Philosophical Topics*, Vol. 29, pp. 429-460.

Williamson, T. 2000: *Knowledge and Its Limits*, Oxford: Oxford University Press.

Williamson, T. 2007: *The Philosophy of Philosophy*, Oxford: Blackwell.