Folk Psychology Revisited:The Methodological Problem and the Autonomy of Psychology

Daniel F. Hartner

Department of Humanities and Social Sciences, Rose-Hulman Institute of Technology

'Folk psychology' is a term that refers to the way that ordinary people think and talk about minds. But over roughly the last four decades the term has come to be used in rather different ways by philosophers and psychologists engaged in technical projects in analytic philosophy of mind and empirical psychology, many of which are only indirectly related to the question of how ordinary people actually think about minds. The result is a sometimes puzzling body of academic literature, cobbled together loosely under that single heading, that contains a number of terminological inconsistencies, the clarification of which seems to reveal conceptual problems. This paper is an attempt to approach folk psychology more directly, to clarify the phenomenon of interest, and to examine the methods used to investigate it. Having identified some conceptual problems in the literature, I argue that those problems have occluded a particular methodological confound involved in the study of folk psychology, one associated with psychological language, that may well be intractable. Rather than attempt to solve that methodological problem, then, I suggest that we use the opportunity to rethink the relationship between folk psychology and its scientific counterpart. A careful look at the study of folk psychology may prove surprisingly helpful for clarifying the nature of psychological science and addressing the contentious question of its status as a potentially autonomous special science.

Keywords: folk psychology, theory of mind, mindreading, autonomy of psychology, theory-theory, philosophical methodology, psychological language

1. What is folk psychology?

Without any specialized training in psychology, people appear to think about —and certainly talk about—mental states. They talk about such states as

Corresponding author's address: Daniel F. Hartner, Department of Humanities and Social Sciences, Rose-Hulman Institute of Technology, 5500 Wabash Avenue, Terre Haute, Indiana 47803, USA. Email: hartner@rose-hulman.edu.

Studia Philosophica Estonica (2016) **9.1**, 22–54
Published online: December 2016
Online ISSN: 1736–5899
www.spe.ut.ee
http://dx.doi.org/10.12697/spe.2016.9.1.02

though they were real things in the head, both in their own heads and in the heads of others. That is, they seem to have mental state concepts like beliefs, desires, hopes, fears, wishes, dreams, intentions, and so forth. And, further, they seem to use their understanding of those concepts to predict and explain what others will do, and even to explain what they have done themselves, almost as though they were applying a kind of general theory about the relationship between mental states and behavior to specific cases. Why does Bill Bob tear the pages out of his pocket bible? Because he *believes* it is a checkbook and he *wants* to buy some candy.[1]

'Folk psychology' is the term coined by philosophers for this commonsensical understanding of the mind or mental content.[2] Folk psychology is a kind of psychology, but the 'folk' prefix indicates that the subject matter needs to be differentiated from the subject matter of other forms of psychology, perhaps most obviously, from academic psychology. Academic psychology is the scientific study of the mind. Folk psychology, by contrast, is the unscientific understanding of the mind as possessed by, for want of better terms, laypeople or folk—people with no special training in the formal academic or scientific study of the mind.

It may help to observe that the term 'folk psychology' mirrors the term 'folk physics,' which similarly distinguishes the putatively commonsensical theory of physics ubiquitous among laypeople from the kind of physics taught in formal academic settings.[3] Just as laypeople possess basic ideas or a basic understanding (which some might call 'a theory') about the movements of bodies in space and the effects of gravity and so forth without any formal training in physics, which they use to predict what physical bodies will do or to explain why such bodies behave as they do, so too, presumably, the layperson possesses a set of ideas about the minds of other people or organisms without any formal training in psychology. The term 'folk psychology' designates, in a very general way, this lay capacity. In sum, 'folk psychology' is just a general term used rather loosely by philosophers to cover a number of phenomena connected to the way that ordinary people seem to think about minds.

But that looseness can become problematic. Consider, for instance, that in the preceding overview we can identify not one but two distinct, albeit related, phenomena. First is the act of thinking about other minds: laypeople actually seem to make use of mental state concepts in order to understand what others will do, that is, to read minds. I will use the term 'mindreading'

[1] The example comes from David James Duncan's novel, *The River Why* (1983).

[2] According to Stich (1983), Adam Morton (1980) coined the term.

[3] For more on the analogy with folk physics, see (Baron-Cohen et al. 2000) and (Greene and Cohen 2004).

(not 'theory of mind') when I want to refer, narrowly, to just this activity. The second is the set of linguistic practices that ordinary people tend to engage in when they read minds and when they talk about their mindreading. They talk about beliefs, desires, wishes, fears, intentions, and so forth. I will frequently call this folk psychological linguistic practice 'description,' since it amounts to describing the activity of mindreading. Thus the lay capacity designated by the term 'folk psychology' actually consists of two related activities: the primary act of mindreading, and the closely associated linguistic practice of description that may well reveal something about that primary behavior (henceforth where I use 'folk psychology' I mean it in just this general way).

Of course, this basic distinction between mindreading and linguistic description has already been recognized in the literature (e.g., Nichols and Stich 2003; Strijbos and de Bruin 2013) and even identified as a potential source of confusion for disputes about theory of mind (e.g., Slors 2012). But its implications continue to be underappreciated. So too are the problems that arise from inconsistent and imprecise employment of the term 'folk psychology' and its associated philosophical lexicon, which includes widely used terms like 'theory of mind' and 'theory-theory.' Concerning this last point, it now seems to be conventional to use the term 'theory of mind' for what is properly called 'mindreading.' But the term 'theory of mind' implies that folk mindreading is facilitated by a *theory*. The view (or theory) that folk mindreading is facilitated by a theory is called the 'theory-theory.' Thus the terms 'theory of mind' and 'theory-theory' are properly interchangeable, and that means that the term 'theory of mind' is now regularly used in places where 'mindreading' is actually intended (see also Note 10 below). Oddly enough, in a move characteristic of the literature on folk psychology, Marraffa (2011) offers just this observation about the misuse of the term 'theory of mind' in the *Internet Encyclopedia of Philosophy* (IEP) before proceeding to preserve this terminological convention in an entry on mindreading that is actually entitled "Theory of Mind." Terminological inconsistency abounds in this literature.

In the next section I outline four of what seem to me to be worrisome cases of terminological—and perhaps also conceptual—inconsistency and confusion that have appeared in arguments concerned with folk psychology among philosophers and psychologists. Even where these problems have already been identified, I think that too little has been done to reevaluate the existing literature in light of them. In fact, in two of the most prominent electronic encyclopedias in academic philosophy, the aforementioned IEP and the *Stanford Encyclopedia of Philosophy* (SEP), there exists no general entry for the term 'folk psychology' at all, despite widespread use of the phrase

in academic work. No doubt there are other closely related entries in those resources—good ones—that summarize and address a great deal of the academic literature that falls somewhere under the heading of folk psychology, including Marraffa's entry, "Theory of Mind" in the IEP, and Ravenscroft's entry, "Folk Psychology as a Theory" in the SEP. But, as I have just showed, they come prepackaged with the literature's tangled terminology, and generally work around rather than reassess terminological difficulties. To be clear, I do not mean to blame the problem on these authors, who no doubt rightly intend their entries to capture and explain existing conventions, however problematic they may be. I only want to point out that terminological inconsistency is readily apparent in this literature, and that it may be more problematic than it first seems.

Even so, I will not attempt an exhaustive encyclopedic overview of the literature on folk psychology. The goal is to offer a fresh look at folk psychology by getting back to basics, as it were. This involves reexamining some of the existing terminological and conceptual conventions and drawing attention to some very basic questions about folk psychology that some specialists writing highly technical papers directed at other specialists now regularly skip. In doing so I think we will find that some of the terminological problems actually reflect more significant conceptual problems. These include the conflation of folk and philosophical perspectives on the mind and mental states, the conflation of mindreading with theory of mind, a tendency to treat folk verbalizations as neutral evidence for theoretical postulates, and a tendency to lump disputes about philosophy of science and mind together with questions about the mechanisms that facilitate folk mindreading. Because these difficulties have clouded the study of folk psychology, I suspect that they have had the tendency to occlude more pressing concerns about the methodology we use to study folk psychology. I will thus sketch what I take to be a methodological problem for folk psychology, and finally offer some reasons to think the problem is insurmountable. Ironically, though, in examining that insurmountable methodological problem, I think we stand to gain much more useful insight into the way people—folk and specialists alike—think about minds than we ever got from ignoring it.

2. Some terminological and conceptual problems

In this section I explain four cases of terminological, and potentially also conceptual, problems identifiable in the literature on folk psychology. Then in the following section I will try to draw some broader conclusions about their impact on the study of folk psychology.

2.1 Mindreading vs. the language of folk psychology

In the previous section I suggested that the distinction between mindreading and folk psychological description is sometimes ignored. That problem is visible in Stich's (1983) influential treatment of folk psychology. He writes:

> In our everyday dealings with one another we invoke a variety of commonsense psychological terms including 'believe', 'remember', 'feel', 'thinking', 'desire', 'prefer', 'imagine', 'fear', and many others. The use of these terms is governed by a loose knit network of largely tacit principles, platitudes, and paradigms which constitute a folk theory. Following recent practice, I will call this network *folk psychology*. (Stich 1983, 1)

As introduced, the target phenomenon here—the thing to be explained—is folk psychology. It is taken as a given that the use of the language of folk psychology is governed by principles that form a theory of concept usage. That is a problem because those principles, if they existed as part of an implicit theory, would directly govern the theoretical mental state postulates that the terms track, not the terms that in turn track those concepts. Though it is not yet clear that the conflation is philosophically significant, it is a mistake to lump the phenomena together from the start. It is not self-evident that the principles governing terminological conventions are identical to the principles governing the deployment of concepts. Indeed adult speakers regularly deploy terms and phrases in ways that indicate that do not fully grasp the corresponding concepts and are simply parroting more proficient speakers.[4] The line between language and concept deployment needs to be scrutinized here rather than shaded over.

2.2 Phenomenon vs. mechanism

Stich's description of this target phenomenon immediately steers us toward the idea that ordinary people (the folk) are governed by tacit principles and theoretical postulates, that is, by a theory. This view that mindreading is facilitated by the layperson's grasp of an actual theory of mind is a view held by specialists called the 'theory-theory' or sometimes 'theory of mind.' Proponents of the theory-theory argue that what explains the ability to predict and explain behavior is that the layperson grasps, even if only implicitly, a set of theoretical mental state postulates, like beliefs, desires, hopes, fears, wishes, and so forth, that they rely on in understanding behavior.

[4] Some common examples of the problem have come to be called eggcorns, e.g., using the term 'Old-timer's Disease' where the goal is to pick out what more competent speakers mean by 'Alzheimer's Disease' or using the expression 'one in the same' where 'one and the same' is intended. There are many such (rather entertaining) examples.

It should be evident that in tying the explanation of the target phenomenon directly to the theory-theory we conflate two different things: what people do, and how specialists think they might do it. Stich is not alone in presenting the target phenomenon this way. Paul Churchland (1989) opens his book with the following outline of mindreading:

> We understand others, as well as we do, because we share a tacit command of an integrated body of lore concerning the lawlike relations holding among external circumstances, internal states, and overt behavior. (Churchland 1989, 2)

Similarly, Horgan and Woodward (1985) write in the very first sentence of their paper in *The Philosophical Review*:

> Folk psychology is a network of principles which constitutes a sort of common-sense theory about how to explain human behavior. (Horgan and Woodward 1985, 1)

These opening descriptions—which seem intended to be relatively neutral descriptions of the target phenomenon—actually come much closer to sketching what we might call the mechanisms that facilitate mindreading than either the behavior of mindreading itself or the linguistic practices associated with mindreading. What they describe is one possible theoretical explanation—a theory according to which the folk rely on a tacit theory of mind (the theory-theory, or the having of a theory of mind)—rather than just the target phenomenon itself.

It is worth noting that the conflation of phenomenon with mechanism is probably reducible to the first terminological problem. Because our only direct access to the prevalence of the phenomenon we intend to explain, i.e., mindreading, comes to us by way of contact with ubiquitous folk psychological description, it is easy enough to slip into seeing that description as reflecting access to a set of concepts that constitute the mechanisms of mindreading. Yet, where we should be especially cautious about navigating this complex relationship between language and mechanism, some authors actually explicitly take folk linguistic practices for clear evidence of what is actually in the heads of ordinary mindreaders. And, interestingly, not all of them are even proponents of the theory-theory.

For example, Maibom (2003, 301), who actually rejects the theory-theory, does so by insisting that ordinary people "seem to master the concepts" involved in mindreading attributions. That is, she rejects the traditional theory-theory account in favor of a model-based account of mindreading, but in the course of the argument nevertheless commits explicitly to the view taken for granted by some proponents of the theory-theory, like Stich, that folk linguistic descriptions are genuine attributions involving the deployment of concepts:

> I am not first a competent attributer of folk psychological states and *then* learn to use psychological concepts in my various projects… Children acquire psychological concepts at roughly the same time they learn to make psychological attributions. (Maibom 2003, 301).

The worry is that she would seem to have us just accept that linguistic practices, like the use of the terms 'belief' and 'desire,' amount to genuine *attributions* of mental state concepts. But, as I have just said, we have to recognize that phenomenon and mechanism can be readily teased apart. In fairness, she does follow up the claim by citing the work of Wellman. However, as I will show in §4, Wellman too seems to mistake linguistic practices (the use of the terms 'belief' and 'desire') for mechanisms of mindreading (deployment of concepts) in his own empirical work.

What is particularly interesting about this example is that Maibom actually sets out to help "flesh out" (Maibom 2003, 314) the theory-theory by developing a model-based account of the mechanisms of mindreading. This is necessary, she says, because proponents of the theory-theory, like Stich, cannot make good on the claim that the folk theory is implicit, and the model-based view furnishes a way to give up the claim without having to insist that the folk have explicit knowledge of universal generalizations about mental states. But in the course of fleshing out the view, she ends up committing the very same error that I have just said proponents of the theory-theory sometimes do: committing, on too little evidence, to the view that linguistic utterances (i.e., folk descriptions of behavior) amount to genuine mental state attributions that require a grasp on concepts (i.e., committing to a particular mechanism that would explain mindreading).[5] The only difference is that Maibom at least seems to commit explicitly to the idea, gesturing toward some empirical evidence to support it, rather than merely neglecting the crucial distinction. But, as I have said, I will try to cast doubt on the quality of that evidence in §4.

[5] It would be interesting to investigate whether it is possible to develop model-based accounts of mindreading, like Maibom's, without importing the still unsubstantiated assumption that ordinary linguistic practices must reflect a folk grasp on mental state concepts. If so, perhaps model-based accounts could help us get traction on understanding the actual mechanisms of mindreading. Nevertheless, I remain somewhat pessimistic for the reasons I detail in the final section. Separating linguistic practice from the mechanisms of mindreading is a difficult task, and the kind of work that psychologists are currently doing—which I also discuss in §4—does not seem sufficient to show that people really have a grasp on the relevant concepts. So, as it stands, the model-based accounts seem to struggle with the same problem that all accounts of mindreading seem to struggle with: the problem of distinguishing between the observable phenomenon (linguistic practices) and mechanisms. I discuss that problem more formally in the next section.

2.3 Mechanism vs. philosophical theory of mind

A third common conflation is evident in the opening sentences of Stich's next chapter, which is tellingly titled "The Theory-Theory" despite being the very first chapter in a section titled "Folk Psychology." There we are told that:

> What may well be the most widely accepted theory about the nature of commonsense mental states is the view Morton has labeled *the theory-theory*. 'Functionalism' and the 'causal' theory are more common labels for the doctrine. (Stich 1983, 13)

This too neglects a crucial conceptual distinction. The theory-theory is a specialist's explanation, the truth of which is heavily disputed, of the mechanisms that facilitate folk mindreading. That is not the same thing as, or a synonym for, a philosophical account of the nature of commonsense mental states. It could not be, because we do not know whether commonsense mental states map onto anything at all in reality (i.e., to *real* mental states), and the view that they do is a contentious theory in philosophy of mind called functionalism.

Functionalism is a philosopher's theory of mind, not a theory of the mechanisms of folk mindreading. It holds, very roughly, that mental states are to be defined and delineated by what they do, i.e., by the function they serve in the organism. It is thus a view about the actual nature of mental states. Of course, this philosopher's theory of mind may well have implications for philosophical views about how ordinary people think about and talk about what is in each other's heads (i.e., 'folk psychology' understood generally). Nevertheless, that is not the same thing as a theory about the mechanisms that facilitate folk predictions and explanations of behavior, which in this case is held to be an implicit theory that includes mental state postulates (which themselves could be either roughly accurate or completely defective). At best, the theory-theory and functionalism will turn out to be rather closely related if it turns out that the folk really do use a theory (i.e., the theory-theory is true) *and* that the analytic functionalist was right all along that something in reality *must* fill the roles postulated by the folk theory (see Jackson and Pettit 1990). But even if the theory-theory turns out to be true and the analytic functionalist is right, the former is still a theory of folk mindreading mechanisms and the latter a philosophical theory of mind. They are theories that purport to explain different things, and so they cannot be interchangeable.

2.4 Theory of mindreading vs. theory of mind: Whose propositional attitudes?

Since at least the time that Churchland (e.g., 1989) and Stich (e.g., 1983) were offering widely discussed arguments about folk psychology, a fair amount of

work that we now tend to associate with folk psychology has focused on the status of eliminativism and the associated question whether beliefs and desires can be analyzed as propositional attitudes (see Gordon 2009). This is perhaps a rather peculiar feature of the literature, however, because in retrospect some of these arguments may have had—and were perhaps even intended to have—relatively little to do with how ordinary people conceive of mental states, and much more to do with the development of philosophical theories of mind. Much like in the previous section, there are two different questions here: 'How does folk mindreading really work?' and 'What theory of mind can philosophers construct from folksy resources, regardless of whether the folk actually use such a theory?'. In the absence of some argument demonstrating the contrary, these questions must be distinguished because the status of the philosophical theory may not necessarily tell us anything about the actual mechanisms of mindreading in ordinary people, and vice versa. Strangely, though, there are cases in the literature in which it is difficult to tell—even within one author's single chapter—which of these two distinct questions is under consideration.[6] I'll demonstrate using arguments from both Churchland and Stich.

The basic connection between the two questions might seem clear enough. The dominant philosophical explanation for mindreading is the theory-theory account, and that account holds that folksy mental state terms like 'belief' and 'desire' designate genuine theoretical constructs. Prominent defenders of this theory-theory view (though in very different forms and with very different motivations) include Paul Churchland, Alison Gopnik, and the Canberra School analytic functionalists Frank Jackson, Philip Pettit, and Stephen Stich. Proponents of the theory-theory sometimes claim that these constructs can be analyzed as propositional attitudes. This propositional-attitudes analysis, as I'll call it, essentially says that the states the folk postulate in their folksy theory take the form of propositional attitudes, i.e., attitudes we hold toward propositions, which are themselves statements about the world that bear truth-values. If I believe that the Earth is round, one might say that I take an attitude of assent toward the proposition, 'the earth is round.'[7] Analyzing folk mental state postulates as propositional at-

[6] I exclude here work on the so-called platitudes analysis associated with David Lewis (e.g., 1970) that is explicitly engaged in constructing a *psychological* theory from folksy resources.

[7] Philosophers sometimes explain such propositional attitudes in the form of that-clauses: I believe (i.e., assent to the proposition) *that* the Earth is round. Some, e.g., Schroeder (2004), even conceive of desires in this way as well. When I want or desire to drink a beer, I have the attitude of desiring toward the state of affairs: I desire *that* I should drink a beer. That is, I take an attitude of desire toward the state of affairs expressed in the proposition 'I drink a beer.' This is obviously clumsy, and so there is dispute about whether it is appropriate to conceive of desires as propositional attitudes.

titudes has been a real focal point of the literature on folk psychology. Yet it seems to me that the crucial question for determining whether this issue really has much to do with folk psychology is a rather basic one: Whose propositional attitudes—the philosopher's or the folk's? The former would seem to make the propositional-attitudes analysis part of a philosophical theory of mind, which is, once again, not the same thing as an account of the mechanisms that enable folk mindreading.

Presumably there are two ways of addressing this question that would succeed in preserving the relevance of the philosopher's propositional-attitudes analysis to the actual facts about folk mindreading: either argue that there is no real distinction between these two projects (i.e., between developing a philosophical analysis of mental states as propositional attitudes and explaining the actual mechanisms of mindreading), or argue explicitly for the claim that the folk really do use a theory of mental states as propositional attitudes and then proceed to explain the precise sense in which they actually use it, i.e., is it explicit or only implicit? In either case, one would expect the chosen solution to be readily detectable in the project. It is strange then that at least in two prominent projects that extensively integrate folk psychology and the propositional-attitudes analysis—from Churchland (1989) and Stich (1983)—this is never really made clear. Neither explicitly attempts to show that the two projects are identical, or even explicitly states that assumption from the start. Perhaps this suggests that their projects intend to take the second route, holding that the propositional-attitudes analysis is relevant to folk psychology because the folk *actually* use such a theory.

But, in turn, if the second approach were intended, one would expect it to be clear *how* the folk use such a theory. Is the claim that the folk explicitly understand mental states as attitudes toward propositions, or is the claim rather that they somehow use a propositional-attitudes analysis only implicitly, without any explicit grasp on how they really are conceiving of mental states like beliefs and desires?

The case for the claim that the folk actually use the propositional-attitudes analysis—if that is what is really intended—would seem to depend on assessing such crucial details as whether the theory is explicit or implicit. Yet both Churchland and Stich at times seem to move back and forth between these views, making it rather difficult to tell whether the project is really supposed to reveal anything about *folk* psychology at all. For example, Stich (1983, 1, emphasis added) introduces folk psychology as the use of mental state terms "governed by a loose knit network of largely *tacit* principles, platitudes, and paradigms." Yet throughout the argument he seems to rely on the assumption that the theory is explicit. Here he describes the

development of philosophical and psychological behaviorism—theories of the nature of the mind—as being at odds with folk psychology:

> As behaviorism flourished, a chasm began to open in our culture. The picture of the mind shared by the historian, the poet, the political theorist, and the man in the street was being rejected by the vanguard of scientific psychology. (Stich 1983, 1)

This passage, like many others in the chapter, clearly presupposes that the ordinary "man in the street" really has a picture of the mind that could be upended by a technical theory of mind like behaviorism. Indeed if the ordinary person's theory were really tacit, it is hard to see how it could be upset by the rise of behaviorism in academic philosophy and psychology. This is peculiar. The easiest, and I think most charitable, solution here is to simply read Stich as having little interest in what ordinary folk are *actually* doing when they talk about beliefs and desires, except insofar as their folksy ways of talking about minds can be—and perhaps have been—spun into a fuller, more technical philosophical or psychological theory of the mind. After all, that seems to be the way that others have interpreted the project (see Gordon 2009) and here Stich even seems to say so himself:

> From antiquity to the beginning of the twentieth century, such *systematic* psychology as there was employed the vocabulary of folk psychology. Those who theorized about the mind shared the bulk of their terminology and their conceptual apparatus with poets, critics, historians, economists, and indeed their own grandmothers. (Stich 1983, 1, emphasis added)

The real target, then, is the *systematized* version of folk psychology built up from the familiar folk terminology. This suggests that Stich's project was always about specialized philosophical or psychological theories of mind. But, in that case, that it is so closely tied to the theory-theory and folk psychology seems really misleading in retrospect. Perhaps Stich intended this to be rather obvious from the start. Even so, the point remains that our thinking about folk psychology stands to benefit from drawing a much clearer line than we find here between what the folk actually do and what philosophers can do with folksy resources in trying to determine the proper foundations for cognitive science.

That same line is blurred in a nearly identical way in Churchland's (1989) work on folk psychology and the propositional attitudes. Like Stich, Churchland repeatedly describes folk psychology as a "tacit command of an integrated body of lore" (Churchland 1989, 2). But he is also prone to implying—even in the very same chapter—that the propositional understanding of mental states is explicit. In describing the effects of folk psychology's eventual

elimination on the future human beings who will witness its demise, he writes:

> How will such people understand and conceive of other individuals? To this question I can only answer, "in roughly the same fashion that your right hemisphere 'understands' and 'conceives of' your left hemisphere: intimately and efficiently, but not propositionally!" (Churchland 1989, 21)

Now this particular passage might be mere hyperbole or rhetorical flourish, but it is nevertheless part of an entire section devoted to the eventual effects of elimination, some of which are supposed to pertain to ordinary people and human social and legal institutions. It is hard to understand why the (explicit) philosophical realization that the propositional-attitudes analysis is radically defective would have any effects on people who are using such an analysis only implicitly, which is to say, without any real awareness of it. If the theory is currently both implicitly used and actually false, then whatever effects its falsity has should *already* be detectable. Ordinary people would presumably only feel the effects of the realization that the theory is false if they were using the theory explicitly. Maybe Churchland really does think that ordinary people use the theory in some explicit way. To be clear, my goal is not to try to upend Churchland's view. I only mean to point out that it is rather peculiar that this issue about whether the theory is implicit or explicit—as central as it is for understanding the actual mechanisms of folk mindreading—would be unclear in a project so closely connected to folk psychology. That is why I think we get a clearer idea of Churchland's project by simply detaching his defense of eliminativism from any real connection to folk psychology or the theory-theory. It is more sensible to read eliminativism as a theory concerned with assessing a philosophical theory of the mind built up from folk vocabulary rather than as a theory of how the folk actually read minds. And since the theory-theory is concerned with how the folk actually read minds and not with whether we specialists can use folk vocabulary to develop an accurate, technical theory of mind, it seems Churchland's view does not really require much recourse to the theory-theory after all.

In fact Churchland's commitment to the theory-theory just seems to cause unnecessary trouble for him, since he then has to square the claim that the folk really use a theory with the eliminativist claim that that putative theory of mind is "a radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by completed neuroscience" (Churchland 1989, 1). But, as critics have rightly asked, if the theory is *that* defective, how could it really be an implicit guide to anything for ordinary

people? What can it really mean to claim that the folk implicitly or tacitly use an analysis of mental states *as* propositional attitudes—an analysis they do not explicitly grasp—that turns out to be so radically defective that the principles of the theory have to be displaced rather than reduced by neuroscience? Presumably what it would mean for ordinary folk to implicitly grasp the theory without an explicit understanding is that in some way their approach is latching onto the true account of mental states as propositional attitudes. But then it is hard to make sense of the claim that the theory is so radically defective. Horgan and Woodward (1985) and Horgan and Graham (1991) have used exactly this observation against the forms of eliminativism offered by both Churchland and Stich. But if eliminativism is really just concerned with a philosophical theory of mind built from folk vocabulary rather than with an actual folk theory, this objection loses its force. It makes more sense to think of eliminativism as being aimed at a defective philosophical theory constructed from folk vocabulary rather than at a genuine folk theory, and in that case the theory-theory seems to have little relevance to the issue.

It is hard to say whether the initial plausibility of the connection between the theory-theory and eliminativism was the cause of confusion over what the theory-theory is really about or simply facilitated by it. In either case that connection is much weaker than it looks. These views are readily detachable. For example, a theory-theorist could insist that the folk do use a theory to read minds, adequate or defective, and nevertheless still maintain that specialists can use the folk vocabulary to construct a different, presumably better, theory of mind that serves as an adequate foundation for cognitive science. Conversely, an eliminativist could argue that the specialist's theory of mind that is built from folk vocabulary is radically defective and therefore an inadequate foundation for cognitive science, and nevertheless not claim that (or even care whether) the folk ever actually used *that* constructed theory. Perhaps they use some other simpler one. At times this seems to be exactly what Churchland and Stich are claiming. So unless we stipulate that for some reason the theory-theorist has to mean that the folk use the same theory as the one that philosophers can reconstruct from folk resources, the theory-theory and eliminativism seem like they can be quite easily separated. In other words, insofar as the term 'theory-theory' designates the view that the *folk* really do use a theory, it is not clear that either Churchland or Stich need or want to talk about the theory-theory at all.

And so it seems the eliminativist could just as well argue that the propositional-attitudes analysis of mental states is not an actual folk *theory*, not a layperson's tacit command of an integrated body of lore that includes the propositional-attitudes analysis. It is rather a defective philosophical theory

that fails to fit the psychological facts. As Churchland himself has argued so persuasively, the time frame for our everyday, folk cognitive processes does not fit very well with the hypothesis that ordinary people rely in some way on propositional deductions to reason about the world or about what other people will do, nor do people understand propositional logic well enough to reason with propositions, as by deducing one proposition from others (see Churchland 1989, 199). As it stands, these considerations make it rather difficult to see the point of his commitment to a propositional-attitudes analysis of the theory-theory, of *folk* psychology, in the first place. That the dispute about the propositional-attitudes analysis has been so closely connected to folk psychology and mindreading strikes me as deeply misleading.

These issues should have been more clearly disentangled before they had a chance to generate further confusion. For example, if eliminativism is not actually aimed at eliminating a commonsensical view of the mental, but rather claims that the *philosophical* theory of mind built from the resources of commonsense vocabulary is radically defective, then one could be an eliminativist of Churchland's sort without committing to the claim that folk psychology constitutes a genuine theory, i.e., without committing to the theory-theory, or indeed without saying much at all about how the folk actually read minds. This is somewhat ironic given that, as Gordon (2009) points out, the development of the simulation theory—which offered new hope of upending the theory-theory—was met with particular interest precisely because it seemed to open up the possibility that the dispute between eliminativists and psychological realists over the proper foundation of cognitive science was baseless. Given the preceding, these two disputes may have had much less to do with each other than it seemed. Whether a theory of mind externally constructed from the resources of folk vocabulary is the proper foundation for cognitive science seems like a perfectly viable question—the question of interest to Churchland and Stich—even if the simulation theory had managed to show that ordinary people do not actually use a theory.

2.5 How did we get here?

In retrospect, some of the foregoing terminological and conceptual problems were perhaps inevitable given at least two features of the development of this literature. First, folk psychology as an area of serious philosophical inquiry has had an unfortunate developmental trajectory. As I just showed in §2.4, over the last thirty years or so philosophical explorations of folk psychology have been tied to other disputes or theoretical commitments in analytic philosophy. Even as disputes about folk psychology were having their heyday in the philosophical literature in the 1980s and '90s, owing largely to the contributions of the Churchlands and Stich, the emphasis was

less on understanding and unpacking the phenomenon than using it as a springboard to the defense of complicated methodological and metaphysical commitments in philosophy of mind and science. Intertheoretic reduction, eliminativism, neurocomputationalism, the representational theory of mind, and so on, were the real focus, not understanding folk psychology per se. The result has been a literature that often approaches folk psychology from the periphery, as a means to some other philosophical end. This in itself has probably abetted a fair deal of the terminological inconsistency and yielded a body of literature that is not particularly accessible to those working outside some of the central disputes in analytic philosophy of mind and science, exacerbating the problem.

Secondly, and perhaps in part as a result of the diversions caused by the preceding terminological and conceptual confusions, far too little attention has been paid to the role that folk psychological description plays in the study of folk psychology. Folk psychology is a research area that deals extensively with how ordinary people talk about minds. Talking about minds can be tricky enough. Talking about how *other people* talk about minds requires keeping track of quite a few layers of analysis. Furthermore, this second-order mind-talk tends to be carried out by philosophers and psychologists who make observations and formulate theories about how such mind-talk works using much of the very same vocabulary—the language of beliefs and desires and so forth—that ordinary folk use in their (first-order) mind-talk. They do that, of course, because they have no alternative. Psychology's specialized theories are built up from a familiar folksy vocabulary.

Folk psychology is an area of research that is quite unlike any other in this particular way. Technical professions like medicine, economics, law, and so on, typically adopt distinctive technical terms to help avoid precisely the kinds of confusions generated by employing ordinary lay terminology to more rigorous forms of inquiry. In the case of folk psychology that approach appears to be unavailable. Though the problem has been recognized in various ways (e.g., Hutto 2008; Slors 2012), its impact on research in this area continues to be underappreciated. I think it poses a significant methodological problem for research on folk psychology that warrants more attention.

3. Folk psychology's methodological problem

To outline the problem more formally, it will help to establish some terminological conventions consistent with the preceding discussion. I will continue to treat *folk psychology* as a blanket term covering both the practice of *mind-reading*—the apparent ability of ordinary people to predict what others will do—and the ways in which they talk about minds. I will call the latter folk psychological *description* or *linguistic practice*. These terms refer strictly to

the ordinary, and seemingly unavoidable, talk about mental states (e.g., 'Bill Bob believes his pocket bible is a checkbook'). I will use the term *mechanism* to refer to the actual mechanism—the psycho-neural process(es), whatever they may be—that actually make it possible for ordinary people to successfully read minds (assuming that they do).

Finally, the distinction between description and mechanism is in turn further complicated by the presence of two distinct perspectives on folk psychology: that of the lay mindreader and that of the specialist (the philosopher, psychologist, etc.). I propose to call the lay folk perspective the *internal* perspective, and the specialist's perspective the *external* perspective.[8] We might think of internal and external here as perspectives on a shared framework, a shared language of folk psychological description. When I mindread, my perspective is internal; when I study the nature of mindreading, it is external. When I describe behavior using the language of beliefs and desires, my perspective is internal; when I attempt to study the way in which ordinary people employ the language of beliefs and desires, my perspective is external. For example, there is a difference between how a layperson might explain her ability to read minds when pressed for information and the way in which a researcher would describe what facilitates that person's ability. When a layperson is pressed for information about how they understand Bill Bob's real mental states, they are engaged in hypothesizing about mechanisms from the internal perspective. By contrast, when a specialist investigates the neural and/or psychological processes that facilitate successful reading of Bill Bob's mind, they are engaged in the mechanism question from the external perspective.[9]

The methodological problem is the problem of determining how it is possible to separate, from an external perspective, folk psychological description from the mechanisms of mindreading well enough to study the latter directly. It is thus the question: How can we objectively investigate the mechanisms of a phenomenon (mindreading) observable only through the very same linguistic framework (folk psychological description) that ap-

[8] Unfortunately, Stich and Ravenscroft (1994) have already introduced an internal/external distinction in some of their joint work. Theirs distinguishes between mindreading (internal sense of folk psychology) and Lewis' platitudes account of folk psychology (external). This distinction is rather different from what I'm proposing here. Theirs appears to ignore the distinction between differences in perspective in description or folk psychological language. Also, as Ravenscroft (2010) notes, they have dropped their internal/external distinction in part because the terminology has not caught on.

[9] It is possible that the internal perspective on the mechanism question never really arises at all. Here I'm not interested in whether every product of intersecting distinctions is a tenable research program in itself. I'm only marking conceptual distinctions and showing that they intersect.

pears to answer the mechanism question? Everything we know, or think we know, about the way ordinary people read minds comes to us through the lens of folk psychological description. Some of that information comes from internal self-report (I know how I think about minds), some from others' reports of their internal experiences (we can ask how others go about thinking about minds), and some from external observation (I can observe in the field, so to speak, that others talk in ways that seem to reflect their thinking about minds), but in all cases the data comes by way of folk psychological description. In this way our linguistic practices are a permanent confound to investigating the mechanisms of mindreading. Delineating mindreading as a phenomenon requires employing a descriptive vocabulary that invariably appears from the external perspective to track a set of theoretical mental state constructs. It is a problem that grows out of the shared language—folk psychological description—of the internal and external perspectives on folk psychology.

Given that researchers, like laypeople, are stuck with the vocabulary of folk psychology, we might expect them to read too much off of folk vocabulary and to find abundant evidence for the view that ordinary people really do have a theory of mind, i.e., for the theory-theory.[10] And in fact this seems to account for some of the conceptual and terminological errors outlined earlier, for example, the tendency explain folk vocabulary as a set of theoretical postulates, and the tendency to treat mindreading as a synonym for having a theory of mind. Indeed some philosophers have even argued *explicitly* that the inevitability of talk about beliefs and desires seems to support its claim to empirical adequacy (see Horgan and Woodward 1985 and Horgan and Graham 1991). But as major milestones in the development of psychology and philosophy of mind show, that line of reasoning has to be defective. The fact that my introspective examination reveals my own dependence on beliefs and desires does not tell me which came first: the language or the concepts. It is just as plausible that I think I rely on beliefs and desires pre-

[10] In this section I continue to resist convention and use the terms 'theory of mind' and 'theory-theory' interchangeably when grammar permits. The only difference is that we should tend to use 'theory of mind' in first-order discussions of mindreading (e.g., 'Churchland believes that people have a theory of mind') and 'theory-theory' in second-order discussions of specialists who are arguing about mindreading (e.g., 'Churchland subscribes to the theory-theory'). But the target referent is the same. The theory that ordinary people have a theory of mind is called the 'theory-theory' for short. Either you think ordinary people have a theory of mind or you deny it. The philosophical/empirical claim that ordinary people do *not* use a theory of mind is typically defended by way of the simulation theory, but there may be other non-theory-postulating accounts of mindreading to oppose the view that mindreading is facilitated by a theory of mind. Not so for the theory of mind. Either people have a genuine theory of some kind (the theory-theory is true) or they do not (simulation theory or some other non-theory-theory alternative is true).

cisely because that is the only way I know how to talk about them as it is that I talk about them because I actually understand and use them. It is, as the old expression goes, a kind of chicken-and-egg question, and an old one with a rich history.

Exactly this mystery about the right causal ordering of language and concepts has figured centrally in such psychological developments as the death of introspection as a serious method in psychological science (am I really reporting my mental states or just filling them in as I talk?), the development of the James-Lange theory of emotion, which reversed our ordinary way of thinking about emotions as causal forces (perhaps I'm scared because my parasympathetic nervous system is activated, not vice versa), and of course in the rise of philosophical behaviorism and logical positivism (so-called 'mental states' are just ways of talking about behavioral processes). It was also the central question that led to the development of folk psychology as an area of research. That question, going back Sellars' (1956) "Myth of Jones," is essentially whether our folk psychological language tracks our concepts or vice versa. One possibility is that my linguistic descriptions of others' behavior come to be applied, internally, to myself. In this way, I find in my own head, somewhat literally, the mental state concepts that begin as linguistic characterizations of hard-to-observe behaviors, like subtle bodily expressions (fears, hopes, and so on). Once I learn to wield these linguistic tools, I apply them to myself, creating mental state postulates. An alternative account might say that the mental states are real and language simply tracks them—language has developed out of a need for a way to talk about mental goings-on.

So the history of this old problem of the relationship between language and concepts really is in some sense just the history of psychology's development and its tangled relationship with philosophy of mind, especially, the rise of behaviorism, which had hoped to address the problem of unreliable introspective access by doing away with the focus on mental states and their familiar attendant language altogether. But for such a familiar problem, the significance of its implications for research on folk psychology has been consistently underappreciated. How is that possible?

Perhaps it is because this issue has remained for so long unsettled that it gradually became common practice to ignore it and return to other interests in philosophy of mind, like the mechanisms of mindreading. Or perhaps, as I argued in §2.4, it is because some of the putative research on folk psychology was never really all that concerned with actual folk mindreading at all. Whatever the explanation, it makes little sense to carry on talking about folk psychology without returning to that problem directly. Any investigation of the mechanisms of mindreading is bound to face the question of whether

we are studying genuine mental processes or merely the way that people talk about behavior and dispositions in terms of psychological states. I have two unreliable options for understanding what is in someone else's head: inferring from my own experience to hers on the basis of her behavior (when it parallels mine) or simply asking her. Both rely on introspective examination and report. With any such report, I do not know whether the reports track the mental states or vice versa. As Freud showed, *pace* Descartes we lack direct access or awareness to our own mental processes. Furthermore, as we might say in more contemporary parlance, people are prone to rationalization, confabulation, and self-deception.

But after Sellars, research on mindreading pressed on, and eventually two predominant theories of the mechanisms of mindreading emerged, namely the theory-theory and the simulation theory. Proponents of the simulation theory seem to deny that mindreading involves the use of a set of theoretical postulates. They hold instead that reading other minds is an act of first-person simulation. To understand what others do, and what they will do, we (the folk) simulate their mental experience within ourselves. No set of theoretical postulates is required, then, since we simply employ our own cognitive facilities, shifting to the perspective of another agent. The dispute wages on and so far the results have been, in light of the preceding, rather predictable. The theory-theory is, it seems fair to say, the dominant account. This is predictable because the dispute is obviously confounded by the fact that there is only one linguistic framework for reporting mental processes: the way we all naturally talk about mental processes. The research question is predisposed to generate an answer that supports the view that ordinary people have a theory of mind (i.e., the theory-theory). Because the mechanism problem has to be settled by research that has no possible platform other than the platform that generated the theory-theory, it is to be expected that its main competitor, the simulation theory, emerged much later and has always seemed difficult to define or explain without emphasizing the parts of the theory-theory it denies (see the accounts in Marraffa 2011 and Gordon 2009 for examples).

Oddly enough, critics of eliminativism very nearly brought the methodological problem to the fore in insisting that eliminativism was self-defeating since it could not even be formulated without clear linguistic recourse to the very folk states it proposed to eliminate (e.g., Rudder-Baker 1987). The eliminativist believes there are no beliefs! The objection has been more than adequately addressed by eliminativists. But where it should have succeeded was in returning our attention to the methodological problem: researchers interested in the mechanisms that facilitate mindreading are always stuck behind description, i.e., a vocabulary that invariably seems to posit a set of concepts

that serve as a mechanism for reading minds. It is just that the right conclusion to draw from this is not that folk psychological language must track real theoretical mental state postulates that actually facilitate mindreading because everyone has to use such language including eliminativists, but precisely the opposite: that we cannot even address the mechanism question properly without a more neutral linguistic framework, which seems incredibly unlikely given the unavoidable tendency for self-professed eliminativists to regularly use commonsense mental state terms just like the rest of us. That more important point was missed, I suspect, because the real focus was not so much understanding folk mindreading as it was defending philosophical theories of mind. Having formally outlined the problem, it is worth one more demonstration of its impact on research programs, but this time from the empirical side. I think the illustration can help to shed light on the nature of psychological science.

4. The methodological problem in empirical psychology

Gopnik and Wellmann (1992) have argued that the evidence from studies of child development overwhelmingly supports the theory-theory over the simulation theory. Since the argument has garnered considerable philosophical attention over the years, it makes a poignant example. Children, they claim, really do rely on an implicit theory containing postulates like beliefs and desires in order to predict and explain what others do. The theory-theory account of mindreading, they maintain, fits the data in ways that its primary competitor, the simulation theory, does not. Their argument depends on the demonstration of some key empirical results from studies that aim to explain how children predict, explain, and interpret behavior. But failure to address the methodological problem is plainly visible here.

First, they claim, children's explanations, in open-ended explanation tasks, show a "theory-like pattern":

> In open-ended explanation tasks…children are simply presented an action or reaction ('Jane is looking for her kitty under the piano') and asked to explain it ('Why is she doing that?'). There are many mental states that might be associated with such situations. Yet 3- and 4-year-old children's answers to such open-ended questions are organized around beliefs and desires just as adults' are ('she wants the kitty'; 'she thinks it's under the piano'). Moreover, there is a shift in explanatory type between two and five. Two-year-olds' explanations almost always mention desires, but not beliefs. Asked why the girl looks for her doll under the bed they will talk about the fact that she wants the doll, but not the fact that she believes the doll is there. Three-year-olds invoke beliefs and desires, and some threes and most 4- and 5-year-olds consistently refer to the representational character of these

> states, explaining failure in terms of falsity. (Gopnik and Wellmann 1992, 153–154)

There are clear potential confounds here associated with the role of linguistic development in general and the development of folk psychological language (i.e., the way that researchers prompt children to talk about mindreading) in particular. The task is an open-ended question—'Why is the girl looking for the doll under the bed?'—which, as Gopnik and Wellman are using it here, is intended *not* as a test of linguistic comprehension but as a neutral test for the mechanisms of mindreading.

But because the goal is to use the task to distinguish between competing accounts of how mindreading works, it is only a viable task to the extent that the child succeeds in mindreading. Thus, the task will succeed in its intended purpose only if two conditions are met: (i) the child's response clearly indicates adequate comprehension of task (otherwise it cannot serve as a genuine test of the mindreading mechanism but rather of linguistic development), and (ii) it is a genuinely neutral test for the mechanisms of mindreading, i.e., there are conceivable answers capable of satisfying (i) that would not immediately serve as evidence in favor of the theory-theory. It is impossible for any response to the task to satisfy both of these conditions. Any answer which succeeds in satisfying (ii)—by avoiding the language that would favor the theory-theory account—will succeed only in virtue of violating (i), i.e., by being so linguistically incoherent that it would simultaneously show that the task has not been adequately comprehended. If this is right, such a task could show nothing about the mechanisms of mindreading.

In fact, the only conceivable response that even an adult could give to this question that would not reveal a misunderstanding of the task is something resembling the claim that the girl is looking under the bed because she *believes* the doll is under the bed, and because the girl *wants* to find the doll. Nothing else will do. Suppose, for example, the child utters complete nonsense in response (e.g., 'because kitty,' or 'she eats waffles,' or even something more grammatically sophisticated but plainly irrelevant like 'I stayed at a Holiday Inn Express last night'). These answers would immediately suggest that the task has not been comprehended and so cannot reveal anything about the mechanisms of successful mindreading. The same goes for a result in which the child offers no response at all, or responds to the question with bewilderment.

Now consider a more coherent response that might otherwise be taken to undermine the theory-theory. Suppose the child responds with an answer couched in first-person language. The child says that the girl is looking for the doll under the bed because, 'I saw you put it there.' Here we would still be forced to conclude that the task has been inadequately comprehended, since

the question is about a third party (a girl looking under the bed) rather than about the experimental setup. There is simply no conceivable response that will succeed in confirming that the task has been adequately comprehended without thereby vindicating, or at least providing evidence in favor of, the theory-theory. The very task—at least as it is being used in this context—is built on the methodological problem, that is, the problem of determining how it is even possible to extricate the study of mindreading from the study of the language of mindreading. At most what a task of this sort is capable of showing about the mechanisms of mindreading is the near-tautology that *when* children respond coherently to questions couched in the language of folk psychology, they do so with responses couched in the language of folk psychology.

Second, Gopnik and Wellman argue that something resembling the theoretical constructs of belief and desire facilitate the child's ability to predict what others will do in complicated cases. Some of the best evidence for the child's use of the theory-theory, as they see it, comes the well-known false-belief task, which was itself designed as a test for the presence of theory of mind. In a common version of the task, a child is presented with a story of two characters (typically portrayed using dolls) in which the first character, Sally, leads the second character, Anne, to have a false belief by moving an item (a box of crayons, a marble, etc.) from its original location to a new location while Anne is not looking. A child who understands the events and wields a theory of mind should, in theory, be capable of predicting that Anne will look in the original location, because she falsely believes the item remains in the same location. Thus when children are asked where Anne will look for the item, they pass the test if they report the original location and fail if they report the new location. Gopnik and Wellman maintain that young children, with their rudimentary non-representational concept of perception and desire, cannot account for complex representational relations between the mind and the world. Thus we should expect that, if children really do have an early non-representational form of the theory-theory at work, they should fail the false-belief task:

> Both desires and perceptions, on the 2-year-old view, involve simple non-representational causal links between the world and the mind… This theory cannot handle cases of misrepresentation…The theory also cannot handle other problems that require an understanding of the complexity of the representational relations between mind and world…The most well-known instance of such an incorrect prediction is, of course, the false-belief error in 3-year-olds.… (Gopnik and Wellmann 1992, 155)

The thrust of the argument is that because children's false-belief predictions fail in the ways that the theory-theory mechanism would predict were that

mechanism under development, the theory-theory must be under development in children. And since it is under development in children, it, rather than the simulation theory, is more likely to be the mechanism explains mindreading.

The initial assumption here—the idea that if the theory-theory account is right, a rudimentary form of it will be detectable in experiments with children—is plainly problematic. Given the methodological problem with disentangling the deployment of folk psychological description from the mechanisms, it is quite plausible that we would find evidence of basic folk psychological vocabulary in conversations with children about others' behavior even if no such mechanism were at work. This is because, as we have just seen, the questions posed to the children themselves employ folk psychological language and, as a corollary, any answer to an experimental task couched in the language of folk psychology will be read as coherent only if it too responds with that language (the only language anyone, including the researcher, has at his or her disposal). Any answer that does not will succeed not in undermining the theory-theory mechanism but the child's very ability to adequately participate in the study.

Moreover, the argument depends on the following conditional: if children are using a rudimentary, non-representational form of the theoretical constructs 'belief' and 'desire', then their predictions will fail in cases that require representational states. The methodological problem looms large here. The setup of the conditional is problematic. There are lots of explanations beyond the limitations of a rudimentary desire perception theory that would equally well explain the child's failure in tasks that require explaining mental representation (like false-belief tasks). The most obvious alternative, which is inexplicably never considered in the course of the argument about child development, is that the child simply lacks the linguistic capacity to fully comprehend and/or adequately respond to the question. What evidence do we really have that the child even understands what is being asked? Prima facie, it seems problematic to assume that a child who purportedly lacks the conceptual capacity to deal with others' representational states well enough to verbally attribute false beliefs to others nevertheless adequately understands questions about the behavior of other agents when those questions could only be satisfactorily answered by wielding mental state terminology in conjunction with a complex grammar.

To be sure, some controls are used in such studies to rule out the possibility that the child comprehends nothing at all about the task. Typically the child's eligibility can be determined by simply testing whether they're capable of engaging in a very basic discussion about objects like puppets. But such controls are irrelevant. In principle, none could ever rule out the

conflation of language with mechanism in the way required to genuinely vindicate the theory-theory. For, once again, the only possible control that could rule out the confound without already begging the very question the study is intended (in this context) to test is one in which the child (i) demonstrates proficiency in using the kind of complex language about others that would be needed to attribute representational mental states in the course of an explanation but (ii) without using the kind of language that a researcher will take to vindicate the theory-theory. There is no conceivable response that could satisfy both of these requirements. A coherent response to the question requires recourse to terms resembling beliefs and desires. Thus any child that understands the questions will have achieved that milestone only by progressing along far enough in their linguistic development to begin using the language of the theory-theory.

There are just two final points worth considering. First, we have to be careful to separate the validity of the false-belief task, which is really designed only to test *whether* children can read minds, from its use as a test for *how* children read minds, which is how Gopnik and Wellman are using it. The fact that linguistic competence cannot be readily extricated from responses that would seem to favor the theory-theory does not seem to present any special problem for the false-belief task in itself, so long as in using it we are content to study facts about a child's linguistic development in conjunction with the child's development of mindreading. All competent speakers use the language of the theory-theory, and in that sense it seems we have to accept that part of the very question the false-belief task is rightly after is whether kids are far enough along in their development to begin using such language. I am concerned here only with evaluating its use as a test of the mechanisms of mindreading. This is a kind of off-label use. When one's aim is to use a false-belief task as a neutral test of the mechanism of mindreading, the methodological problem becomes a barrier. Any child who can legitimately participate in the study by wielding the language of mental state representations will, in virtue of using the only language available for that purpose, appear to be employing a theory of mental states, thereby supporting the theory-theory.

Second, the preceding critique has focused exclusively on so-called *explicit* false-belief tasks in which children are required to verbalize their reactions to false-belief scenarios. But *implicit* (or nonverbal) versions of the task, in which researchers rely strictly on behavioral cues to determine whether the child is mindreading (or, as they wrongly tend to say, whether the child has a theory of mind), require no such verbalizations from the children. They rely instead on researchers' observations of looking times (e.g., Onishi and Baillargeon 2005) and even on more sophisticated data from

eye-tracking software (Southgate et al. 2007). In both cases, however, the researchers must interpret visual cues—how long a child looks at the false-belief scenario—as a sign that the child is surprised by the kinds of false-belief experiments I described above. Is it possible that these implicit versions of the task offer a way out of the methodological problem?

The proposal may be worth exploring, though I remain skeptical. In fact I suspect that these tasks actually manage to show even less about the mechanisms of mindreading than the explicit cases. As I just said of the explicit versions, the implicit tasks are built to test *whether* the child is reading minds, not *how*. Even if the researchers mistakenly claim the test is looking for a 'theory of mind' (a common terminological error), it is actually testing for mindreading. At bottom the test is merely a test of whether the child's expectations about what another person will do have been defied. But showing that they have is not thereby evidence for the theory-theory, since presumably those defied expectations could also be explained by simulation theory or some other non-theory alternative.

In fact, the moment we begin to think that the implicit tasks really do serve as evidence for the theory-theory, we should start to wonder whether there is any data that we could ever accept as evidence for any account of mindreading *other* than the theory-theory. For if the evidence for the very existence of the phenomenon that the theory-theory is intended to explain (i.e., the child looks longer at false-belief scenarios) is *also* evidence for the theory-theory, then it has become impossible to extract the phenomenon from the proposed mechanism and there is no genuine test for mechanisms here at all. Instead, it looks a lot like the test for the mechanism comes loaded with a virtually unfalsifiable research hypothesis. It is only falsified when the entire test is rejected. To make this point even more obvious, consider it from the other direction. If we assume that the implicit test might be able to tell us something about the mechanisms of mindreading, and then run the test and find that the child fails it, that is, she does not in fact look longer at the false-belief scenario, does this show that the child is not using the theory-theory, or only that she is not yet able to read minds (does not have a 'theory of mind' as the researchers say)? The point here is that, just as in the explicit case, if this is to be taken as a genuine test of the mechanisms of mindreading, we have to be able to specify the conditions under which the child demonstrates mindreading competence but does not thereby automatically provide evidence for the theory-theory. The only difference between using the implicit test for evidence of the theory-theory and using the explicit test is that in the implicit case we simply skip the step that makes the methodological problem more obvious. In the explicit case the researcher poses a question the only coherent response for which requires that the child use language that

would automatically vindicate the theory-theory. In the implicit case the researchers need not even bother to lead the subject to vindicate the theory-theory. They simply interpret the child's behavior on their behalf using the only language any of us have for talking about mental states: the language of the theory-theory. For these reasons it seems to me that no false-belief task of any kind has much to offer the question whether mindreading requires a genuine theory of mind.

5. Folk psychology and the autonomy of psychological science

I want to briefly sketch one of the larger conclusions that I am inclined to draw from the preceding explication of folk psychology's methodological problem. As space is limited, it is more a sketch than a complete argument. But it may offer a useful perspective on what the challenge of studying folk psychology reveals about the nature of psychological science and the contentious question of its status as an autonomous science.

In total I have outlined five conflations in the literature on folk psychology: mindreading with description, phenomenon with mechanism, mechanism of mindreading with philosophical theory of mind, folk vs. philosophical use of propositional attitudes, and, in empirical practice, the conflation of evidence for description with evidence for mechanism. What appears to drive most of these conflations is a vocabulary that ranges over both internal and external perspectives on the way people appear to deal with mental states. The internal perspective is the first-order perspective of the mindreader as they (apparently) manipulate mental state terms. The external perspective is the second-order perspective from which we assess how mindreading works. It seems to be remarkably difficult to avoid running these together in the course of an argument about folk psychology, particularly because specialists always occupy both perspectives simultaneously. Thus a philosophical account of how mindreading might actually work comes to be blended with an account of the way ordinary people talk about minds, thereby treating an internal linguistic practice as a window into a mechanism; an external philosophical analysis of mental states as propositional attitudes gets projected onto folk minds as a theory visible in their own account of their internal processes; intuitive and presumably learned ways of talking come to be treated as empirical evidence for an actual psychological mechanism of mindreading from an external perspective rather than as mere evidence for the development of the same language we all use from the internal perspective. In such cases, perhaps we export something internal, taking it as genuine evidence of mechanisms visible from an external perspective. If this diagnosis is right, it suggests a simple solution: researchers can do better about disentangling the perspectives, keeping them distinct in

the course of research on folk psychology. I am inclined to think that that suggestion is not so simple after all.

The most obvious reason that the internal and external perspectives cannot be so easily disentangled is, as we have already seen, precisely because there exists no alternative descriptive or linguistic framework for articulating observations about mindreading from the external perspective. Even for specialists the language of folk psychology is unavoidable. Philosophers, psychologists, and neuroscientists of various stripes deal in beliefs and desires and the like. No compelling solution to this problem has ever been provided. The Churchlands, for example, worked hard to piece together descriptions of observable behavior couched entirely in the language of low-level neuroscience for the purpose of suggesting that, with the proper training, people might over time learn to speak the complex language of neuroscience and leave off any talk about beliefs and desires. But it always seemed a stretch at best. And it ought to, because even really good neuroscientists, for all their concern with careful scientific methodology, tend to trade in beliefs and desires inside their own research projects—indeed much of neuroscience is, or is related to, *cognitive* neuroscience, which relies heavily on familiar psychological concepts like belief. Serious scientists employ these concepts regularly even in technical work because they simply have no alternative. The prospects for suddenly developing an alternative to folk psychological description seem really poor.

There is a helpful parallel here with Nagel's famous (1974) paper addressing the likelihood of our developing a reductive physicalist explanation of consciousness. Nagel showed that there is a plain conceptual problem with reducing a fundamentally subjective first-person phenomenon to an objective perspective: objective perspectives work by taking you away from subjectivity, which in the case of consciousness means taking you away from the very phenomenon you had hoped to explain. He concluded that we might have to hold out for a new set of concepts and a new method of "objective phenomenology not dependent on empathy or the imagination" (Nagel 1974, 449). Forty years later, we are no further along on that project either.

The problem that Nagel identified is probably a species of the same problem that plagues folk psychology. It is a problem of language, and the odds of solving it in the case of folk psychology are equally grim. For in the case of folk psychology, the task is to develop a new linguistic framework for dealing with mindreading that would allow us to take an objective, external perspective on how mindreading really works and still somehow account for the familiar language of beliefs, desires, and so on as an integral part of the phenomenon of interest. But, to adapt Nagel's point to folk psychology, any linguistic framework that could succeed in objectively describing how mind-

reading works without begging the question in favor of the theory-theory will succeed in virtue of moving away from the target phenomenon we had hoped to describe: it will take us *away* from the familiar language of beliefs and desires, leaving us no way to say anything useful about that language, which is integral to understanding the phenomenon that piqued our interest in the first place.

As a result, I'm inclined to draw a different conclusion. Rather than seeing the problems with studying folk psychology as a challenge to develop new concepts and methods, as Nagel did for consciousness, I think we ought to see it instead as evidence for a version of the autonomy of psychology thesis, though not of the sort that traditional proponents of psychological autonomy would want to endorse.

The autonomy of psychology is a phrase that grew in popularity with the publication of Fodor's famous (1974) paper "Special Sciences," that advanced a form of functionalism at the expense of reductive physicalism. Proponents of the autonomy of psychology generally hold that psychology enjoys (a common way of putting it) a kind of independence from neuroscience. It has long been a fancy way of insisting that psychology has nothing to fear from the progress of neuroscience, since it is inconceivable that the kind of high-level cognitive work conducted by psychologists could be done more accurately by scientists engaged directly with real physiological processes in the brain, as reductive physicalists were betting. Special sciences like psychology thus cannot be reduced to more basic physical sciences by way of neuroscience. The implication has always been that psychology is and will remain a viable, empirically respectable science in its own right, regardless of what neuroscience might turn up about how the brain works.

In suggesting that folk psychology's methodological problem exposes the autonomy of psychology, I do not mean to endorse this view. Though I am drawn to the view that psychology is an autonomous science for a variety of reasons, I am much less convinced that it remains the kind of independent science that genuinely aims to reveal bare facts about the world as it is independently of human experience. Psychology probably is in many respects an autonomous science, but my suspicion is that it does not exactly enjoy that autonomy. I think, rather, the main line of argument in this paper shows that it suffers from it.

What plagues psychology, unlike lower-level (e.g., cellular and molecular) neurosciences, for example, is a particularly insidious kind of theory-ladenness in the course of picking out the phenomena to be investigated. The unavoidable vocabulary of psychology common to the internal and external perspectives compromises the ability of psychology to reach the world as it is in itself by generating research goals that focus on the world as we shape

it with our linguistic conventions. For example, a psychologist might want to understand the nature of our beliefs, or the effects of having particular beliefs on some other hypothesized state like motivation. Unlike trying to understand how a particular gene contributes to the development of a particular phenotype, or how complexes of proteins facilitate the transport of ions across cell membranes, such questions depend much more heavily on the terminology we use to formulate the question than they do on the bare empirical world. That is not to deny that 'genes' or 'cells' or 'proteins' are terms that shape our empirical investigations in some way. But their impact is significantly less than in the case of psychology. We can empirically confirm that something like ions cross something like cell membranes, however we describe the process. But we still do not even know what it *really* means to say that Bill Bob believes his pocket bible is a checkbook. As the logical behaviorists and ordinary language philosophers taught us, it is possible that the entire question rests on a confused way of talking, a mere artifact of our peculiar language. However wrong we might be about ion channels, it is not likely that we could be wrong in the same way we could be wrong about mental states. In the case of psychology, the linguistic framework wreaks havoc not merely on understanding the mechanisms of the phenomena of interest, but on picking out the phenomena themselves that are to serve as proper targets of empirical investigation of the world.

There is a slightly different way of putting this point, which is due originally to Carnap (1950) and which has been helpfully revived by Bickle (2003). Carnap distinguishes between internal and external questions about linguistic frameworks, where internal questions are those that work from within, or presuppose the entities specified by, the linguistic framework, and external questions are questions about the framework itself, such as whether the entities it postulates are real. For Carnap, as for Bickle, a linguistic framework agreed upon by scientists does not entail ontological commitments. Scientists agree for practical and linguistic reasons to use certain terms, but they are not attempting to establish ontological facts with such terms. As a result, the terms are meaningless outside the framework, and so external questions about whether the terms used in the framework map onto reality are meaningless. Carnap used the distinction to deal with pesky metaphysical questions about science, such as whether numbers are real, by showing that such questions wrongly take scientific terms outside their agreed upon framework, rendering those questions meaningless.

The distinction between internal and external perspectives in folk psychology maps neatly onto Carnap's internal/external distinction. We may either be situated internally with respect to the language of folk psychology, employing beliefs and desires to predict and explain behavior as ordinary

people do, or we may step outside that framework, as the specialist (say a scientific psychologist) wants to do, to make sense of that folk language and its relationship to mindreading. The peculiar problem in the case of psychology and its folk counterpart is that psychological terms like belief and desire figure centrally in both the language of the layperson and the language of the psychological specialist. Academic psychology, insofar as it studies concepts like belief and desire as part of the mind, is forever taking internal language outside its internal framework.

Thus we might say that the autonomy of psychology is a product of its unavoidably moving between internal and external perspectives given the inevitable vocabulary of beliefs, desires, and the like. To want to understand what beliefs really are, in the real world, for example, is to take a term internal to a ubiquitous and widely accepted linguistic framework and use it to pose external, objective questions about what such terms refer to out there in the world. But that question has already been answered: they just refer to verbal utterances the folk use in the course of trying to explain why Bill Bob is tearing pages out of his pocket bible. That *is* what 'belief' means because the term comes from established folk convention, not from psychology. Of course, the psychologist could insist that her project is rescuing the term from problematic usage, but in that case the term would need to be assigned a new operational definition built from the resources of scientific psychology rather than folk vocabulary, and the subsequent research would have nothing to do with the concept as ordinarily understood. The psychologist could just bite that bullet, except that in the peculiar case of psychology it is impossible, because there simply is no alternative conception of belief disconnected from the way that folk use the term. The only observable phenomenon on which to build the operational definition for 'belief' is the way that ordinary people talk about beliefs, for that is where the term comes from in the observable world.

Thus any empirical test for the validity of the operational definition depends on an analysis of folk linguistic conventions rather than any real empirical work having to do with minds as they really are in the world. Specialists can still argue about how to 'properly' conceive of the nature of beliefs, but that is just good old-fashioned conceptual analysis in philosophy of mind, not the empirical study of the psychological world.

This suggests that psychology actually suffers from its autonomy in the sense that its inability to objectively investigate its own terminological decisions empirically results in its detachment from the empirical world. Psychology, unlike neuroscience, formulates the distinctive language it uses to build theories about the mind from facts about folk linguistic practices, not from facts about how minds really work in some objective sense. By simply

assuming that its terminological decisions track facts about minds rather than folk language, it can proceed as any good science should, with careful methodology designed to test theories. In that sense it remains a kind of science. It may well be a very valuable one for a variety of purposes. But perhaps as with some forms of economics, its theories, to the extent that they actually purport to address minds rather than behavior, are to be understood not as accurate depictions of the external world as it is independently of human conventions (the "science of the mind") but rather as accounts of how the world is through the lens of our unavoidable terminological conventions. Its ability to say anything about the bare facts about our world is always constrained by its inability to empirically vet its own operational definitions. Thus it is a kind of science tethered more to human language than the bare empirical world. And since the study of folk psychology relies on all of the same terminological conventions as academic psychology, it is no surprise that our attempts to study it objectively suffers from the same methodological problem.

Folk psychology is just a term for the way that ordinary people dabble in the subject matter of "real" psychology. But unlike the relationship between physics and folk physics, in the case of psychology the academic version may well be nearly as indifferent to the bald empirical facts about minds (if there are any) as its folk counterpart. It is the autonomy of psychology in this sense that probably best accounts for all the trouble we've had making sense of folk psychology.

Acknowledgments

Thanks to Bruno Mölder and an anonymous reviewer for very helpful comments, criticisms, and suggestions.

Bibliography

Baron-Cohen, S., Tager-Flusberg, H. and Cohen, D. (eds) (2000). *Understanding Other Minds*, 2nd edn, Oxford University Press, Oxford.

Bickle, J. (2003). *Philosophy and Neuroscience: A Ruthlessly Reductive Account*, Kluwer, Dordrecht.

Carnap, R. (1950). Empiricism, semantics, and ontology, *Revue Internationale de Philosophie* **4**: 20–40.

Churchland, P. (1989). *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science.*, Bradford, Cambridge.

Duncan, D. (1983). *The River Why*, Sierra Club Books, San Francisco.

Fodor, J. (1974). Special sciences (or: The disunity of science as a working hypothesis), *Synthese* **28**: 97–115.

Gopnik, A. and Wellmann, H. M. (1992). Why the child's theory of mind really *is* a theory, *Mind & Language* **7**: 145–171.

Gordon, R. (2009). Folk psychology as mental simulation, *in* E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, Fall 2009 Edition.
URL:*http://plato.stanford.edu/archives/fall2009/entries/folkpsych-simulation/*

Greene, J. and Cohen, J. (2004). For the law, neuroscience changes nothing and everything, *Philosophical Transactions of the Royal Society of London B (Special Issue on Law and the Brain)* **359**: 1775–1785.

Horgan, T. and Graham, G. (1991). In defense of southern fundamentalism, *Philosophical Studies* **62**: 107–134.

Horgan, T. and Woodward, J. (1985). Folk psychology is here to stay, *Philosophical Review* **94**: 197–225.

Hutto, D. (2008). Lessons from Wittgenstein: Elucidating folk psychology, *New Ideas in Psychology* **27**: 197–212.

Jackson, F. and Pettit, P. (1990). In defence of folk psychology, *Philosophical Studies* **5**: 7–30.

Lewis, D. (1970). How to define theoretical terms, *Journal of Philosophy* **67**: 427–466.

Maibom, H. (2003). The mindreader and the scientist, *Mind & Language* **18**: 296–315.

Marraffa, M. (2011). Theory of mind, *Internet Encyclopedia of Philosophy*. Accessed on 1 July 2015.
URL: *http://www.iep.utm.edu/theomind*

Morton, A. (1980). *Frames of Mind*, Oxford University Press, Oxford.

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

Nichols, S. and Stich, S. (2003). *Mindreading: An Integrated Account of Pretence, Self-awareness, and Understanding of Other Minds*, Clarendon Press, Oxford.

Onishi, K. and Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs?, *Science* **308**: 255–258.

Ravenscroft, I. (2010). Folk psychology as a theory, *in* E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*, Fall 2010 Edition.
URL: *http://plato.stanford.edu/archives/fall2010/entries/ folkpsych-theory*

Rudder-Baker, L. (1987). *Saving Belief: A Critique of Physicalism*, Princeton University Press, Princeton.

Schroeder, T. (2004). *Three Faces of Desire*, Oxford University Press, Oxford.

Sellars, W. (1956). Empiricism and the philosophy of mind, *Minnesota Studies in Philosophy of Science* **1**: 253–329.

Slors, M. (2012). The model-model of the theory-theory: Why 'theory of mind' seems ubiquitous, even though it isn't, *Inquiry* **55**: 521–542.

Southgate, V., Senju, A. and Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year olds, *Psychological Science* **18**: 587–592.

Stich, S. (1983). *From Folk Psychology to Cognitive Science*, MIT Press, Cambridge, MA.

Stich, S. and Ravenscroft, I. (1994). What *is* folk psychology?, *Cognition* **50**: 447–468.

Strijbos, D. and de Bruin, L. (2013). Universal belief-desire psychology: A dilemma for theory theory and simulation theory, *Philosophical Psychology* **26**: 744–764.