

Willem deVries

FOLK PSYCHOLOGY, THEORIES, AND THE SELLARSIAN ROOTS

ABSTRACT. Almost fifty years ago, Wilfrid Sellars first proposed that psychological concepts are like theoretical concepts. Since then, several different research programs have been based on this conjecture. This essay examines what his original claim really amounted to and what it was supposed to accomplish, and then uses that understanding of the original project to investigate the extent to which the later research projects expand on it or depart from it.

1. Introduction: Folk Psychology as Theory

We use mentalistic or psychological terms constantly in describing, explaining, understanding, and generally coping with the behavior of others and ourselves. For several hundred years in the modern era (until, say, 50 years ago), the bad old explanation of how we do that and what exactly is going on when we do seemed, well, not unproblematic, but at least uncontroversial in its fundamental outlines: We enjoy direct and privileged access to our own mental states. This enables us, first, to abstract a set of psychological concepts from actual and self-intimating instances of psychological states. Furthermore, since the psychological concepts we have are abstracted directly from self-intimating instances, first-person application of those concepts is unproblematic. It is incorrigible and provides each of us with certainty about his or her own psychological states. This knowledge is primordial. On its basis we build up knowledge of everything external to our minds. From the materials we have direct access to we (somehow) construct and apply concepts of physical objects and events. More problematically, once we have the core set of concepts and some knowledge of our own psychological states, we can begin to apply psychological concepts to describe, explain, understand, interpret, and generally cope with others as well. This requires the assumption that the observable behavior of others is tied to their internal psychological states in the same ways and in the same kinds of patterns that we find connect our own psychological states and behavior. The argument for that assumption is a pretty weak induction by analogy, and this was a stumbling block for the

In: M.P. Wolf and M.N. Lance (eds.), *The Self-Correcting Enterprise: Essays on Wilfrid Sellars (Poznań Studies in the Philosophy of the Sciences and the Humanities, vol. 92)*, pp. 53-84. Amsterdam/NewYork: Rodopi, 2006.

picture as a whole, but the rest of the picture seemed so solid, so obvious, that most preferred trying to find some way to solve the associated “problem of other minds” rather than rethinking the whole picture.

About fifty years ago some brave souls started rethinking the whole picture.¹ Principal among the attempts to revise the whole picture explicitly was Wilfrid Sellars’s trail-blazing effort in “Empiricism and the Philosophy of Mind.” Sellars proposes that our psychological concepts and terms are like *theoretical* concepts and terms. In Sellars’s picture our most primordial knowledge is not our first-person acquaintance with our own mental states, but a knowledge of “medium-sized dry goods,” the public, physical objects and events that form the arena within which we conduct our lives. Sellars argues that we should reverse the hitherto standard picture that our knowledge grows “from the inside out”: if we begin with a knowledge of the public objects and events that shape our lives and at least a rudimentary set of metalinguistic, particularly semantic, concepts, we would have all the material necessary to developing the psychological concepts that Cartesians think are simply given elements of our mental lives.

Sellars’s proposal, radical at the time, struck a chord and has since become enough of a commonplace that many no longer know its provenance. Sellars’s original idea has served as inspiration for subsequent developments that have picked up the notion that psychological concepts are like theoretical concepts. If psychological concepts are like theoretical concepts, then our use of those concepts, even in our everyday attributions of psychological states to ourselves and others, is like using or applying a theory. Commonsense psychology is itself, then, really like a theory.

While Sellars’s proposal has influenced almost everything in subsequent philosophy of mind, there are two different projects that have drawn particular inspiration from this aspect of Sellars’s proposal that I want to focus on.² The first, whose principal spokesmen have been Richard Rorty (1965) and Paul Churchland (1979), both of whom worked with Sellars, is *eliminative materialism*. They point out that a salient property of theories is susceptibility to replacement by a better theory. If commonsense psychology is a theory, then it, too, is susceptible to replacement by a more adequate theory. Given its

¹ Hegel had already rethought it from the ground up, but left little mark on English-speaking philosophy, the domain I am most concerned with here. And Wittgenstein started rethinking it in the 1930s, but nothing appeared in print until the 1950s, though there were many reports of his thought circulating.

² Other projects, such as the functionalist analysis of intentional states, have drawn inspiration from other aspects of Sellars’s proposal.

hoary past and extended run, they argue that commonsense psychology is not only susceptible to replacement, we should *expect* it to be replaced. It is not only a theory, it is a bad theory. It took several thousand years for folk physics to be superseded, even after Aristotle codified it well enough to enable some more or less clear-cut tests. Folk psychology has endured through the first few centuries of serious efforts at scientific psychology, but as neuroscience advances, it, too will eventually succumb. We will find ourselves in the seemingly paradoxical position of denying that the psychological concepts in terms of which we (currently) understand our own behavior really designate any robust features of the world. In our current lingo: we'll end up believing that there are no beliefs.

The second major project that takes off from Sellars's proposal is not concerned with the large-scale picture of the conceptual evolution of psychological concepts in human society, but the small-scale picture of the individual's development and application of psychological concepts. If psychological concepts are like theoretical concepts, then one should investigate whether the kinds of processes by which psychological concepts are acquired and applied by the individual are like the kinds of processes by which theoretical concepts are acquired and applied. Indeed, in (probably unbeknownst) tribute to Sellars's suggestion, the investigation of the human ability to attribute psychological states to oneself and others and utilize such attributions to understand, explain, and sometimes predict human behavior is often called inquiry into (or the theory of) Theory of Mind.³ A number of psychologists are vigorously pursuing this project, e.g. Janet Astington, Alison Gopnik, Josef Perner, Henry Wellman, Alan Leslie, and Paul Harris. Since 'theory of mind' is now a general name for a complex field of investigation, however, it covers a wide range of specific proposals that take the phrase and attendant analogy more or less seriously. Gopnick and Wellman (1992) take the phrase 'theory of mind' most literally; she argues that the "theory of mind" that children acquire and the process by which children acquire and learn to use it is very much like a scientific theory and the corresponding process of

³ Consider, for example, the titles of representative contributions to the field:

"What Is Theoretical About the Child's Theory of Mind? A Vygotskian View of Its Development" (Astington 1996); "The Theory of Mind Deficit in Autism: How Specific Is It?" (Baron-Cohen 1991); "Does the Autistic Child Have a Theory of Mind?" (Baron-Cohen, et al, 1985); *Theories of Theories of Mind*. (Carruthers and Smith 1996); *Introduction to Theory of Mind: Children, Autism, and Apes*. (Mitchell 1997); "Theory of Mind Is Contagious: You Catch It from Your Sibs." (Perner, et al, 1994); "Insights into Theory of Mind from Deafness and Autism" (Peterson and Siegal 2000); "Does the Chimpanzee Have a Theory of Mind?" (Premack and Woodruff 1978); *The Child's Theory of Mind* (Wellman 1990).

theory acquisition and revision in the sciences. Other contributors to the field discount the analogy to varying degrees, claiming, e.g. that our “theory of mind” is a highly implicit set of representational capacities. Some claim that our “theory of mind” is contained in a set of informationally encapsulated, innate modules that need only appropriate triggers at the proper point in one’s developmental history to blossom. By this point, ‘theory of mind’ functions arguably only as a label and has lost any descriptive functionality. In this paper we will focus on those who take the “theory of mind” metaphor quite seriously.

Both of these Sellars-inspired projects have drawn substantial critical response. Eliminative materialism has been vilified as well as criticized, and the literature examining the claims of eliminative materialism is very extensive. Sellars has himself been branded an eliminativist, though I will argue that this is unjustified guilt by association. Against those who take ‘theory of mind’ to be a description and not a mere label, a group of “simulationists,” e.g. Robert Gordon, Alvin Goldman, and Jane Heal, have argued that we cannot appropriately understand our own use of psychological concepts as anything like an instance of theoretical practice. Though there may be some abstract analogy between theoretical concepts and psychological concepts, the mechanisms by which we apply psychological concepts are not at all like the kinds of mechanisms theorists employ in their work.

I am now finally in a position to state my intent in this paper. I want to go back to the original source in Sellars of the idea that psychological concepts are like theoretical concepts, see what his original claim really amounted to and what it was supposed to accomplish, and then use that understanding of the original project to investigate the extent to which the later research projects expand on it or depart from it. There is some purely historical interest in my effort here, but I think we will also find that it helps us get a better understanding of the terrain in which all these programs operate. Eliminative materialism is a broad, general challenge to a pervasive and fundamental aspect of our conceptual framework; the debate about the theoreticity of our “theory of mind” is a fairly narrow and ultimately empirical debate in psychology. Understanding the connections and disconnections among them is important to understanding the current state of the philosophy of mind.

2. Sellars’s Original Proposal

Sellars’s stated goal in “Empiricism and the Philosophy of Mind” (1997) is to attack the “Myth of the Given” (§1/p. 14). Sellars believes this myth is not a purely *epistemological* fiction, however. It needs and finds support in a set of

metaphysical assumptions that must also be overthrown if the attack is to succeed. In one sense, Sellars's attack on the Myth of the Given is complete halfway through the paper, in Part VIII, where he pulls together the threads of his critique of traditional notions of the given and sketches his own nonfoundationalist theory of observational knowledge as an alternative. Why does he not stop there?

Through the halfway mark in "Empiricism and the Philosophy of Mind," Sellars has worked hard to displace first-personal knowledge of one's mental states from the foundation (or the core, if you prefer that metaphor) of our knowledge. Having achieved that, however, he must then develop an alternative, non-Cartesian explanation of the privacy and authority of first-person knowledge as well as of the nature and structure of our knowledge of others' minds. For Sellars does not want to *deny* that our self-ascriptions of mental states usually have a special, first-person authority, nor that we have in some respectable sense privileged access to them. And it is crucial to his overall view that our intersubjective knowledge of others' mental states can be as firm as our purely subjective knowledge of our own states.

So the situation at this point in the essay is that Sellars believes he has given an adequate, non-Cartesian reconstruction of the human capacity for observational knowledge of public, physical objects and events, a reconstruction that notably does not itself *require* that humans have prior concepts or knowledge of their own (or others') mental states as *inner, private, or privileged*. He then needs to show that he can account for concepts and knowledge of one's own (and others') mental states in ways that explain privacy, first-person authority, and intersubjectivity without falling back into Cartesian conceptions of givenness.

Sellars's strategy is fairly direct, given the situation. He tells a little myth himself about some people who exemplify the dialectical situation reached at this point in the essay: the "Ryleans" have rudimentary observational knowledge. Indeed, they exhibit fairly sophisticated abilities to describe and explain the behavior of public physical objects, utilizing the subjunctive conditional and other sophisticated syntactic forms. But they have no conception (and therefore no knowledge) of their own mental states as inner. Could these Ryleans suitably acquire concepts and knowledge of mental states (their own and others') as inner, private states to which they have privileged access?

Sellars's answer is *yes, if* they also have (1) semantical concepts they can and do apply to their language, and (2) the ability to theorize, i.e., to formulate the conception of a new domain, not otherwise perceptible, modeled on one's conception of an already familiar and more or less well comprehended domain. Sellars's mythical character Jones provides the theory, formulating a new

conception of an inner domain consisting of *two* kinds of internal states. One kind of state (thought) is attributed properties that are modeled on the semantical properties of speech acts. Such states have meanings, entailments, truth, etc. The other kind of state (sense impression) is attributed properties that are modeled on the perceptible properties of physical objects. Such states can be blue or red, sharp or flat, bitter or sweet.

Jones develops his theory (or is it theories?) in order to explain certain kinds of behavior that appear anomalous in the Rylean framework.⁴ Once he has developed the theory and taught others how to use it, though, “It is but a short step to the use of this language in self-description . . . [I]t now turns out – need it have? – that Dick can be trained to give reasonably reliable self-descriptions, using the language of the theory, without having to observe his overt behavior” (1997, §59/p. 106). Thus are born the notions of privileged access and privacy.

If one accepts the possibility (not the actuality) of Sellars’s myth, then the Cartesian story is not the only one in town. We have available an understanding of our concepts of the mental that is both methodologically and ontologically compatible with empirically and naturalistically respectable science and that does not require any given. Sellars’s attack on the given has been brought to completion: he has not only shown us the problems with the given, he has shown us how to do without it.

3. Eliminativism

This is the basis for attributing to Sellars the view that folk psychology is a theory, and it is a strong basis for that attribution. But I want now to argue that things are not as straightforward as they might seem, for we also have to be careful to ensure that “folk psychology is a theory” is construed the same way by all the parties to this conversation.

3.1. *Theories and Models*

There is one thing to be clear on from the outset, something that some commentators on Sellars have not recognized: whereas Jones’s theory is about our behavior and mental states, Sellars’s theory is about the nature and status

⁴ For more detailed discussion of what Jones’s theories allow him to explain, see deVries and Triplett (2000), pp. 142-144 and 159-164.

of our *concepts* of mental states; Jones was the first psychologist, Sellars is a philosopher.

Sellars does offer us a theory, and it is part of his overall view that theories are often constructed using *models*. A model is a domain of which one presumes to have an understanding, and which one proposes to be in some way analogous to the domain one is currently trying to understand. Interestingly, in his metaconceptual inquiry, Sellars uses as his model for the domain of our mentalistic concepts (our psychological language), which we are trying to understand, theoretical concepts (theoretical language), which he presumes 20 or 30 years of philosophy of science has helped us get a good grasp of.

Now, it is also part of Sellars's understanding of the use of models in theory formation that (1) the analogy is always limited, and (2) the terms of a theory can acquire "surplus value" from their use in their new context, which entails that they become independent of the model and can proceed to develop in their own way in confrontation with experience. To say that the analogy is limited is to say that the analogy is always accompanied, more or less explicitly, by a commentary that specifies which aspects of the model domain are denied to the domain theorized and which aspects of the model domain are attributed to the domain theorized. These conditions entail that it is a mistake to *identify* the domain of the theory with the domain of the model unless there is independent reason to do so. The presumption should be against such an identification.

This already affords us a quick and easy argument – though not a deeply convincing one – that it is a mistake to say that Sellars thinks of folk psychology as a theory.⁵ He employs theoretical concepts and theoretical language as his *model* in order to understand psychological concepts and language, but we should assume that, as in all theoretical models, the analogy is limited and that psychological concepts have some "surplus value" above and beyond the analogous theoretical concepts that disrupts the analogy still further.

⁵ The reason this is not a deeply convincing argument is that the theory Sellars is proposing here is not a scientific theory that postulates a new domain of theoretical entities – the kind of theory to which his speculations on the nature of theories and models directly applies. Philosophical or interpretive theories have to work in a different way, and the role of a model in a philosophical theory will correspondingly be different. But I see no reason why the fundamental principle defended in the argument – that it is a mistake to simply identify the domain of the explanandum with the model in a theory – would itself cease to apply to philosophical theories and the models used in them.

This argument invites us then to investigate more deeply the analogies and disanalogies between theoretical concepts and theoretical language *sensu stricto* and folk psychological concepts and language. If there are systematic and strong disanalogies, then we should indeed abandon the claim that Sellars believed that folk psychology is itself a theory.

3.2. Theory and Observation

The first item in the “commentary” on the theory-model that Sellars is using to help understand folk psychology I want to investigate is the theory-observation contrast. In Sellars’s treatment this is a very important theme. It is in the context of a discussion of the theory-observation contrast that Sellars most definitively rejects identifying mentalistic concepts as theoretical concepts:

. . . I am going to argue that the distinction between theoretical and observational discourse is involved in the logic of concepts pertaining to inner episodes. I say ‘involved in’ for it would be paradoxical and, indeed, incorrect, to say that these concepts are theoretical concepts. (§51/p. 97)

This is a complex topic, but I will try to be brief. Sellars does not hold that *all* theories postulate unobservables, but he does hold that this is true of (scientific) theories in their “most developed or sophisticated form.” (§51/p. 94). There is a continuum of theoretical activity from very low-level empirical generalizations entirely in observation vocabulary to highly mathematized theories that postulate complex domains of unobservable entities. It is important for Sellars that “the ways in which the scientist seeks to explain empirical phenomena are refinements of the ways in which plain men, however crudely and schematically, have attempted to understand their environment and their fellow men since the dawn of intelligence” (§51/p. 97). For instance, commonsense often postulates an *unobserved*, though *observable*, entity or event to help explain some occurrence. How big a step is it to postulate a kind of systematically unobserved event to help explain other occurrences? How large is the gap between the *unobserved* and the *unobservable*? Is there not a difference in *kind* between the unobservable and the observable, whether in fact observed or not?

It is important to see that for Sellars, any such difference is not ontological. The line between the observable and the unobservable is not fixed by ontology, but by our current methods, and it is correspondingly movable as our

methods change (and hopefully improve).⁶ For Sellars, “theoretical entity” and “observable entity” denote *methodological statuses* not *ontological genera*. Given the theory of observational knowledge that Sellars sketched in Part VIII of “Empiricism and the Philosophy of Mind,” anything that one can make a noninferential report of, knowing that that report is a reliable indicator of the matter reported, is thereby observable. Vocabulary that gets its start in a theory (e.g. “There is an alpha-particle track”) can come to be used in noninferential reports. If such usage becomes widespread, that vocabulary thereby enters the observation language.

We can now see why, in the context of a discussion of the theory-observation distinction, Sellars says “it would be paradoxical and, indeed, incorrect, to say that these concepts *are* theoretical concepts” (§51/p. 97). To call a term or concept theoretical is to say something methodological about it; it characterizes the conditions of its use. Among other things, it characterizes it in contrast to observation terms as *not* having a use in noninferential reports.⁷ But when a concept that perhaps originated in some theory becomes regularly used in direct and immediate noninferential reports, it has lost its status as theoretical. And this is precisely the situation in which we find mentalistic vocabulary in Sellars’s myth. Jones’s disciples learn to apply mentalistic language to themselves directly, immediately, and noninferentially. “*What began as a language with a purely theoretical use has gained a reporting role*” (§59/p. 107). Sellars’s phrasing seems to hold out the possibility that some terms can retain a theoretical use while adding to it a reporting role. But in gaining a reporting role, mentalistic language also loses a principal characteristic of the theoretical.

⁶ Many have been tempted to believe that, though the line between the observable and the unobservable is mutable, there are determinate and significant limits in principle to how far it can be moved. I am not sure anyone has clearly maintained that there are entities or features of the world that are necessarily unobservable (and can therefore be known only theoretically). (Classical early modern theorists would not have thought so, because there is nothing in the world that can be known only inferentially, since God’s knowledge is not inferential.) But the idea that there are certain features of the world that are necessarily observable – in particular, the proper sensibles – has been a leitmotif in modern epistemology. It is, however, tied to the Myth of the Given, the idea that there are certain features of the world that are such that when they present themselves to us (or when we encounter them) we thereby know them.

⁷ Notice that having a use in sentences of the theory is not sufficient to make a term a theoretical term. Presumably, there are some points at which the theory makes contact with observation, and therefore there will be sentences in which both theoretical and observation terms occur. This does not make the observation terms theoretical. We have to conclude that calling something a theoretical term involves denying to it an observation or reporting role.

Thus, when Sellars says that “the distinction between theoretical and observational discourse is involved in the logic of concepts pertaining to inner episodes,” we cannot take this to mean that concepts pertaining to inner episodes are to be straightforwardly construed as theoretical and therefore as opposed to observational. At most, he is pointing again to the idea that the theory-observation contrast offers us a useful *model* for illuminating the inner episode-outer behavior relationship.

3.3. Theory and Self-Reports

As I noted above, Sellars seems committed to the idea that anything we can immediately and knowingly report is observable, but we do not normally consider our self-reports as expressions of *observations*. (Imagine the fit Ryle would throw, and properly so, at the idea that I *observe* my mental states!) In normal usage ‘observation’ seems tied to the use of our five senses, and our self-reports about our mental states do not employ the five senses as intermediaries (though sensory states may be the *object* of such reports).⁸ Undoubtedly, this is one reason Sellars employs the more neutral vocabulary of “reports.” Sellars’s story does help us understand the constant temptation (which some philosophers have succumbed to) to assimilate our mentalistic self-reports to sensory observation reports: the justificational structures are very similar, even though the causal mechanisms involved differ in important respects.

Sellars does not bring out clearly enough that there are two different kinds of cases in which the originally theoretical vocabulary proposed by Jones could gain a reporting role. It could be the case that Jones’s disciples learn to apply his new mentalistic vocabulary *to others* directly and noninferentially. In this case, they would learn to *report* (in Sellars’s special sense) the mental states of others. To the Cartesian, of course, this is a blasphemous idea: we have and can have no immediate access to the mental states of others. And it is certainly true that very often we must figure out, laboriously, what other people are thinking or feeling. But it has been pointed out that, at least phenomenologically, we sometimes see another’s anger or happiness. There does not seem to be a Sellarsian reason against saying that, in the proper conditions, I can be both a reliable and a knowing meter of a[n] (admittedly

⁸ Proprioception is a difficult case here. It is not counted among the basic 5 senses, yet does not seem reducible to their output. Self-reports of some mental states, e.g. (some?) sensations such as having a pain in one’s foot, seem like proprioceptive observations. In any case, self-reports of intentional states are not proprioceptive nor are they comfortably treated as observations.

limited) range of the mental states of others.⁹ Perhaps the fact that there are a great many *caveats* on my ability to report immediately and noninferentially on the mental states of others (e.g. I may be able to do so only for a very limited range of people – my family and good friends or maybe the members of my cultural community – and/or for only a very limited range of mental states – principally stronger emotions – and under fairly stringent conditions) discouraged Sellars from mentioning this possibility. But it is worth taking seriously, for it is an indication of how far Sellars’s position moves us from the traditional Problem of Other Minds, and it emphasizes the important point that in Sellars’s theory, psychological vocabulary is univocal across first- and third-person application.

Sellars does focus on the other case in which the Jonesian vocabulary becomes a reporting or observation vocabulary: “And it now turns out – need it have? – that Dick can be trained to give reasonably reliable self-descriptions, using the vocabulary of the theory, without having to observe his overt behavior” (§59/p. 106). The parenthetical ‘need it have?’ is important, for it raises the question of the *necessity* of privileged access.¹⁰ In one important sense, that question is already settled by Sellars’s myth. Some within the Cartesian tradition think that privileged accessibility is a defining characteristic of the mental. There are, as has been noted, varieties of privileged access, however. The most hard-core Cartesians might then believe that it is a defining characteristic of a mental state that it is self-intimating: mental states are such that if one is in mental state *F*, then, necessarily, one knows one is in *F*. If Sellars’s Rylean myth is at all coherent, however, then it is not a conceptual truth that mental states are self-intimating. But weaker versions of privileged access are also threatened by Sellars’s myth. Since the

⁹ As Jay Garfield points out, (personal communication) there is good reason to believe that our expression of and abilities to perceive emotions and other mental states evolved as mechanisms to facilitate social interaction. It is very natural to treat emotion detection mechanisms as perceptual, because they are fast, mandatory, domain specific etc.

¹⁰ But Sellars’s ‘need it have?’ is also frustrating, for it raises important questions without even hinting at an answer. Could there be a linguistic community in which there is a robust psychological vocabulary in use, but in which first-person reports using that vocabulary have no special status? If I read him correctly, Davidson argues that this is not really a coherent possibility: objective knowledge is possible only when all three legs of the triangle self-others-world are, in some sense, immediately available to the knower. (See the essays in Donald Davidson 2001.) Sellars seems to be contemplating the possibility of a linguistic community with a psychological vocabulary but without first-person authority. But he certainly does not commit himself to such a possibility, nor does he explore the extent to which citizens of such a community would differ from us.

Ryleans cannot be thought of as operating with *exactly* our current concepts of the mental, one lesson of the myth is that there could be respectable concepts of mentality that do not include privileged access as an indispensable feature. This is important in deciding what to make of animals, for instance.

3.4. *The Malleability of Observation*

Churchland claims that the malleability of the theory/observation distinction plays in his favor, because it shows that we could learn to report (more adequately) in the vocabulary of his projected ideal neuroscience the presence of those states we now normally report on (less adequately) in psychological vocabulary. There are two problems with this view, however. First, it proves too much for Churchland, and second, it does not fully appreciate the complexity of the relationship between the observational and the theoretical.

It proves too much for Churchland because it ultimately allows us no good reason to stop at the level of neuroscience in our replacement of psychological language. If we can be trained to report our inner states and episodes in neuroscientific terms and thereby do a better job of describing and explaining, why not dispense with the neuroscience and report those states and episodes directly in the vocabulary of physics and do the best possible job of describing and explaining? Churchland is unimpressed by the pragmatic difficulties of using the complex language of neuroscience in our self-reports, so he could not use similar pragmatic difficulties to block the call for discarding neuroscience in favor of flat-out fundamental physics.

More importantly, this view oversimplifies the relationship between observation language and theoretical language. Churchland seems to think of our ordinary observation language as consisting of a variety of modules, each dealing with some subject area, and each liable to replacement by some scientific theory that would improve it. Slowly, bit by bit, then, we will upgrade our language until we find ourselves at home in the world as it really is, as described and explained by science. But this is *not* the way Sellars sees things, even though it does retain traces of Sellars's influence. The problem is that Churchland's view as I have sketched it does not make provision for any overarching structure that would organize the various subject-matter modules in the language, what Sellars would call "framework principles." The issue runs deep for Sellars, for it concerns the nature of the clash between the "manifest" and the "scientific" images of the world – another piece of Sellars's philosophical apparatus that has broken off and acquired a life of its own in philosophy.

3.5. *Categorical Structure and the Manifest and Scientific Images*

In Sellars's view, though science may begin as a sophistication of commonsense methods of explanation, and may thus at first propose revisions and replacements of pieces of our manifest image of the world, it ultimately comes to pose a challenge to that manifest image *as a whole*. This means that the fundamental structures by which things (in the broadest possible sense of the term) are identified and individuated in the manifest image are challenged in the developing scientific image, and the entire ontology of the manifest image is revealed as ill grounded (ultimately) compared to the ontology developed in the sciences. The sciences perform an ontological coup, but not simply by replacing talk of pains with, e.g. talk of brain states. The sciences will reorganize our conceptions of objects, events, properties, processes, particles, space, time, etc., the very categories that organize the world for us. The entire world of the manifest image will have been revealed as phenomenal in a more-or-less Kantian sense: a representation of reality that is well-founded in relation to the native abilities we (as evolved and socially developed creatures) bring to the world, but that cannot claim to be an adequate representation of the world as it is in itself. Sellars believes that the framework under development in the sciences does have a reasonable claim to be an adequate representation of the world as it is in itself. (In this, Sellars departs radically from Kant, however many other Kantian inspirations he retains.) This is the source of Sellars's *Scientia Mensura*: "in the dimension of describing and explaining the world, science is the measure of all things, of what is that it is, and of what is not that it is not" (§41/p. 83).

Sellars's scientific realism, summed up in the *Scientia Mensura*, seems to support the eliminativists. The sciences, presumably including neuroscience, will tell us what exists. The categories of the manifest image will be replaced. Does that not mean that mental states as we now understand them will be revealed as "merely phenomenal," and thus not real? But, of course, things are more complex than that.

First, in order to infer the truth of eliminativism from Sellars's scientific realism, one needs the premise that folk psychology is not only *false*, but *radically false*. And it turns out this does not follow automatically from the claim that the manifest image is merely phenomenal. Eliminativism would not follow from Sellars's position if the following is possible: Though the manifest-image concepts of mental states, both intentional states and sensory states, are false in the sense that they are part of an overall framework that cannot withstand the rigors of repeated, exhaustive, empirical test, there is nonetheless in the successor framework that replaces the manifest image a distinctive set of concepts that retains enough of the general structure of folk

psychology that we cannot say that folk psychology has been so falsified that it has been abandoned. Rather, we have to say that folk psychology is the predecessor of scientific psychology, a scientific discipline that retains some independence from the neuroscientific investigation of the embodying substratum of psychology. Churchland needs to claim that folk psychology is so empirically inadequate that there will be no successor science that will retain even its most general structural features, and Churchland's arguments that that will prove to be the case have been widely criticized.

Here's a revealing parallel: The concept of a brain is clearly present in the manifest image framework. If the phenomenality of the manifest image established the falsehood or inapplicability of its component concepts, then we would have to conclude that there could be no brain science in the projected scientific millennium, since brains do not really exist. Churchland's beloved baby, neuroscience, would be thrown out with the bath water of folk psychology. The inference from the supercession of the manifest image to the elimination of the psychological proves too much, for it would do away with brain science (and physics) as well. The further premise that there is no scientific-image successor to psychology is also necessary, and that could be established only by the failure to develop any such successor, which Churchland is not in a position to ascertain.

Second, Sellars does not think that we are destined to end up with the scientific image, pure and simple. He argues instead that, though science is king "in the dimension of describing and explaining the world," (§41/p. 83) describing and explaining the world are not the only activities of interest to us. We are not only cognizers of the world, we are *agents* in the world, and Sellars does not believe that the concepts essential to our being agents in the world are going to be reconstructed within the scientific image. Therefore, the scientific image will not (and cannot) totally displace the manifest image; we must marry the agent-constituting concepts of the manifest image to the scientific image in order to obtain a "synoptic vision" of humanity's place in the world. The concepts essential to us as agents include precisely the concepts of folk psychology: intentions, desires, beliefs, etc. Psychological concepts are clearly ineliminable, according to Sellars.

This enables us also to see another dimension in which it would be wrong by Sellars's lights to treat folk psychology as a theory in Churchland's sense. Psychological concepts do not, in Sellars's view, constitute a self-contained module within the manifest image that could be unplugged and replaced with something else. They rather infect the whole manifest image, influencing its structure and "framework principles." To the extent that this is the case, to the extent that psychology is present not only in our vocabulary, but in the syntactic structures available to us, in the very structure of the language

through which our experience of the world is filtered, we cannot treat psychology as merely one theory among others, to be adopted or discarded as dictated by the success or failure of a certain research program. Psychological concepts are provided for by the very structure of the language/framework by which the world is constituted for us.¹¹

With these considerations we have emerged beyond the theory/observation distinction, in the following sense. Is ‘physical object’ an observation concept? In one sense, *of course* it is. We can directly, immediately, and noninferentially apply the concept “physical object” to items of experience. But that does not entail that every physical object is observable, nor that the concept “unobservable physical object” is paradoxical or contradictory. Thinking of “physical object” as one concept alongside many others, in fact, gets things wrong. “Physical object” is better understood as a *category*, a fundamental classification that designates a most basic type in the framework or language. Calling something a physical object determines what kinds of things can (and must) be said about it, what kinds of properties it can be said to have, what forms of explanation can be applied to it, etc. It determines the most generic forms of language that apply to it. The concept of a mental event or state is no less categorial for Sellars, determining the fundamental linguistic/conceptual forms that apply to such things. Some mental states may well be observable or reportable, some may well not be. The notion of a mental state is so deeply woven into our language that it is not obvious how it might be related to the notion of a physical object. Sellars’s Rylean myth gave us a deep and fruitful insight into that relationship. But it does not have the consequence that we should expect the progress of science to eliminate the concept of the psychological from our repertoire.

There is, thus, no sense of ‘theory’ in which Sellars would have conceded that folk psychology is an eliminable theory, not even as a consequence of his claim that the manifest image is ultimately to be superceded by the scientific in matters ontological. The eliminativists, as far as Sellars is concerned, took his original insight down a garden path.

¹¹ The biggest impediment to taking Sellars’s Rylean myth seriously, my experience indicates, is the immense difficulty of trying to imagine a language or point of view on the world that is really bereft of the concepts and consequent structures of the psychological.

4. “Theory of Mind,” Psychological Concepts, Theoretical Concepts, and Empirical Hypotheses

We can now turn to consider whether those who take Sellars’s suggestion at face value and make it the inspiration for a research program in developmental psychology are just unpacking elements already prefigured in Sellars, or creatively extending Sellars’s original suggestion, or have somehow misconstrued the meaning of Sellars’s insight. As noted, ‘theory of mind’ covers a broad range of proposals explaining the development and structure of our abilities as folk psychologists. I will focus here on those who take most seriously the idea that our “theory of mind” is really theory-like, such people as Alison Gopnick. I will call this position the “theory-theory.”

4.1. *The Prima Facie Argument against the “Theory-Theory”*

We have already seen that it is a mistake, from Sellars’s point of view, to identify without further ado the domain about which Sellars is theorizing (psychological concepts) and the domain of the model Sellars uses in his theoretical activity (theoretical concepts). Just as it did with the eliminativists, this affords a *prima facie* argument against a Gopnick-style “theory-theory” approach to the theory of mind.

But the situation is, in fact, different in the theory-theory case from the eliminativism case. The eliminativists assume the theory-likeness of psychology and infer the presence in folk psychology of a particular property of theories strictly so-called, namely their eliminability in favor of some better theory. It is an argument from analogy and suffers the weaknesses of such arguments. That is not the overall form of the theory-theorists’ project. Rather, they take the original Sellarsian suggestion, not as an assumption, but as a hypothesis, and investigate the extent to which it can inform an empirically adequate theory, not of our psychological concepts, but of our actual psychological competence, our ability to describe, explain, understand, and generally cope with the behavior of others and ourselves. The theory-theorists are not assuming that the analogy is good and arguing from there, they are empirically investigating how far the first-order form of the analogy can be pushed.

One way to see the theory-theorists, then, is as neo-Jonesians, who, now puzzled by our psychological competence in particular, follow their Master (now fading into mythological history) by explaining those behavioral abilities by postulating inner episodes that are like a specific *kind* of linguistic activity, namely the activity of employing a theory to inform or control one’s behavior. This is a further specification and extension of Jones’s strategy, it seems, and

thus in keeping with the overall thrust of Sellars's story, especially if it makes sense out of the empirical data. The theory-theorists want to see whether Sellars's analogy affords not just a *philosophical* (meta-)theory about the nature of psychological concepts, but also a *psychological* theory about a certain set of psychological abilities.

Thus, the theory-theory seems at least *consistent* with Sellars's original suggestion. But we can also ask whether it might be what Sellars originally had in mind. In order to do this, we need a more detailed examination both of Sellars's text and of the characteristics of theories. Did Sellars believe that the analogy between psychological and theoretical concepts would provide a good empirical explanation of our folk psychological skills? I will start this next phase of the investigation by stating what seem to be the most important relevant characteristics of theories, then examine Sellars's text for clues to what he thought might be the empirical import of his analogy.

4.2. Theories and the Attractions of the Theory-Theory

What are the characteristics of scientific theories that "theories of mind" might share? The details differ slightly depending on which theory-theorist one looks at, but the core relevant properties of theories are the following. Theories are defeasible structures of (generally) propositional knowledge used to provide explanations and (falsifiable) predictions (even in new or counterfactual cases) by postulating a set of unobservable states or entities that are related in certain specifiable, lawlike ways, which relations implicitly define the unobservables in question.¹² To the extent that a conception of the human "theory of mind" deviates from this picture of the "theory" involved, it is not trying to exploit Sellars's analogy explicitly, and is not a "theory-theory" as I am using the term.

Botterill (1996) points out four different kinds of reasons for finding the theory-theory attractive:

There are *epistemological attractions*, associated with the special epistemic status of theories – whether one goes on to add "It is a theory, so it may be radically wrong" à la Churchland, or "It is a theory that works very well, so probably it is broadly correct" à la Fodor. There are *semantic attractions*, associated with the idea that theories provide implicit definitions: on a functionalist view this promises to solve the problem of how we understand the meaning of mental state terms. There are *developmental attractions*, associated

¹² This characterization picks up themes found in, for instance, Gopnik and Wellman (1992) and Botterill (1996). Whether this is an adequate characterization of theories is not a topic for today.

with the way in which theories are discovered and elaborated, which may be supported either by identification of specific proto-theoretical stages, or merely by the claim that a theoretical structure would facilitate an otherwise formidable learning task. Finally, there are *cognitive processing attractions*, associated with the application of theory, which hold out the hope of answering the “How do we do it?” question by employing a frequently successful strategy – i.e., positing a body of tacit theoretical knowledge. (Botterill 1996, p. 106)

How does Sellars’s treatment of the psychological stand with respect to these; which attractions are the ones that drew Sellars? First, notice that these attractions fall into two groups. The first and second, the epistemological and semantic attractions, are *philosophical* in the sense that construing the psychological as theory-like offers interesting answers to (non-empirical) questions usually posed by philosophers, questions about the nature of knowledge and understanding or meaning. The other two, the developmental and the cognitive processing attractions are *empirical*. In these cases the psychological as the theory-like promises interesting answers to essentially *empirical* questions about how we acquire and apply our psychological competence. We expect some kind of causal theory to respond to both developmental and cognitive processing questions, but there is no presumption that the answers to epistemological or semantic questions would be contained in a causal theory.

It is clear that the first, epistemological, attraction, was very significant for Sellars. Remember the context of Sellars’s Rylean myth: a critique of the given. The significance of this for Sellars, its importance in overcoming a faulty Cartesian picture of our relation to the world, has already been discussed above. I want to reserve discussion of the second, semantic attraction, for later, for it turns out also to be very significant for Sellars. So let us turn now to the empirical attractions of taking Sellars’s analogy seriously. And let us start by returning to Sellars’s text to see whether it contains clues about whether Sellars thought his approach would bear empirical as well as philosophical fruit.

Before we move on, however, a remark on the relations among these four attractions. Though they divide into philosophical and empirical pairs, it would be a major advantage to have a unified treatment of the psychological that responds effectively in all four of these dimensions. Notably, a strictly Cartesian answer in the first two, epistemological and semantic dimensions, an answer which would treat the mind as a separable realm from the bodily, makes it difficult to understand how to even begin answering questions about development or cognitive processing. We should expect that Sellars’s anti-Cartesian epistemology and semantics of the mental would put some

constraints on reasonable answers to the developmental and cognitive processing questions.

4.3. *Learning Psychology: The Reception of Jones's Theory*

Looking for further detail on how other humans come to appropriate and apply Jones's theory, perhaps unsurprisingly, we find the text frustratingly brief:

[O]nce our fictitious ancestor, Jones, has developed the theory that overt verbal behavior is the expression of thoughts, and taught his compatriots to make use of the theory in interpreting each other's behavior, it is but a short step to the use of this language in self-description. Thus, when Tom, watching Dick, has behavioral evidence which warrants the use of the sentence (in the language of the theory) "Dick is thinking 'p'" (or "Dick is thinking that p"), Dick, using the same behavioral evidence, can say, in the language of the theory, "I am thinking 'p'" (or "I am thinking that p'.") And it now turns out – need it have? – that Dick can be trained to give reasonably reliable self-descriptions, using the language of the theory, without having to observe his overt behavior. Jones brings this about, roughly by applauding utterances by Dick of "I am thinking that p" when the behavioral evidence strongly supports the theoretical statement "Dick is thinking that p"; and by frowning on utterances of "I am thinking that p," when the evidence does not support this theoretical statement. Our ancestors begin to speak of the privileged access each of us has to his own thoughts. What began as a language with a purely theoretical use has gained a reporting role (1997, §59/pp. 106-107).

And again, about sensations:

Let us suppose that as his final service to mankind before he vanishes without a trace, Jones teaches his theory of perception to his fellows. As before in the case of *thoughts*, they begin by using the language of impressions to draw theoretical conclusions from appropriate premises. . . . Finally he succeeds in training them to make a *reporting* use of this language. He trains them, that is, to say "I have the impression of a red triangle" when, and only when, according to the theory, they are indeed having the impression of a red triangle (1997, §62/p. 115).

In both cases, Jones's protégées first apply psychological concepts to others in a strictly theoretical manner. That is, they note the presence in the other of certain kinds of nonpsychologically characterized observable behavior. They then infer, on the basis of Jones's theory, the presence of certain psychological states, and, if called upon to make a further inference to other psychological states or a prediction of behavior, also do so via the mediation of Jones's theory. They first apply Jones's theory to themselves on exactly the same basis. Later, Sellars tells us, they come to be able to apply

Jones's new vocabulary to themselves directly and make reports of their own states. But notice that Sellars does *not* tell us anything further about the second- or third-person cases. Does Sellars think that Jones's protégées can also be trained to report immediately on the psychological states of others? I argued in the previous section of this paper that being able to *report* on a fact or entity, in Sellars's view, removes it from the realm of the theoretical *sensu stricto*. And I also argued that at least phenomenologically speaking, we can and do often report immediately and directly on the mental states of others. (These points are most directly relevant to the theory-theory as an account of our use of our mature psychological competences. We might still acquire that competence in theory-like ways, even if we eventually learn to apply it in non-theory-like ways.) If these claims are correct, then in neither the first-, second-, nor third-person cases need we be literally *applying a theory* when we exercise our psychological abilities.

4.4. *Explicit and Implicit Theories*

So here is one point on which the theory analogy cannot be pushed. But this point need not faze a theory-theorist, for there is a common move available (and noted by Botterill) that allows the theory-theorist to hold on to the analogy in a fairly strong form while denying that in the exercise of our psychological abilities we are always *applying a theory* in an explicit fashion. That is, the theory-theorist can attribute to us a *tacit* rather than an explicit theory. Although phenomenologically we often respond directly and immediately to our own and other's psychological states, those responses are, the theory-theorist claims, mediated by unconscious internal processes that are like those involved in the application of a theory. So, a theory-theorist might say, when Jones's protégées come to be able to report on psychological states, they have become so fluent in the application and manipulation of the theoretical vocabulary that the cognitive operations involved have become automatic and habitual, so automatic and habitual that they recede behind the line of consciousness. Just as the musician practices until he no longer has to think about what fingers to move where, until it is all so automatic that he need only think about the music itself, we become expert at on-the-fly psychological theory. For the musician, it can be the death of music if she begins to think about her movements, and she could not possibly construct an explicit account of those movements; similarly, we may be incapable of any explicit account of our psychologizing and may get significantly worse at it when we think about it.

Possessing and acting on a *tacit* theory differs in some important respects from possessing and acting on an *explicit* theory. Phenomenologically the two

are very different. But that does not entail that it is wrong to think of them as two different ways of possessing and acting on a theory. The notion of tacit knowledge has been utilized by philosophers since Plato, but it has been at the very heart of the cognitive revolution in psychology: explain complex behavioral capacities by attributing to the organism some internal representational structure(s) that mediate(s) its responses to its environment. Very often, the representational structures that enable a particular capacity or competence are thought to be (like) theories of the relevant domain. That is, the representational structures are taken to consist of items with propositional form that are interrelated by law-like generalizations that are also elements in the system, knit together by some inference engine. Such tacit theories share many characteristics with explicit theories; differing with regard to our explicit, conscious command and manipulation of the theory is not itself sufficient to destroy the otherwise fruitful analogy.

Is there any indication that Sellars thought that our actual, empirical competence at psychology would be best accounted for by literally moving Jones's theory inside our heads and retaining it as implicit? No, there really is not. As I noted earlier, Sellars is (and knows himself to be) a philosopher and not a psychologist. As such, he is not interested in either the developmental story behind our psychological competence (note, there is no mention in the myth of Jones of how children can come to acquire psychological expertise by four or five, well before they would be capable of mastering an explicit Jonesian theory), nor is he interested in the cognitive processing story beyond the commitments contained in Jones's theory that *something, somehow* like overt language is going on in us when we think and that *something, somehow* like a replica of a physical object is present within us in perception. In this light, we have to conclude that the latter-day theory-theory is a creative extension of Sellars's proposal that, though consistent and coherent with his original proposal, is not already contained or intended therein. As such, the theory-theory is subject to empirical confirmation and refutation, not philosophical support or refutation. Should the theory-theory prove not to be the best available empirical explanation, Sellars's original suggestion remains untouched, for the theory-theory goes beyond anything Sellars himself claimed.

Indeed, the hard-core theory-theory, in the light of current research, faces an uphill battle as an empirical theory:

It requires an implausible degree of inductive and hypothesis-testing competence in three- and four-year-olds, an implausible universal, uniform and rather selective fixation on the mind as a domain of study, and fails to explain why [Theory of Mind] would be selectively impaired in those like high-functioning autistics who are nonetheless theoretically competent in other

domains (that is, it does not account for the false picture/false belief performance dissociation). (Garfield et. al. 2001, p. 526)

5. Beyond the Theory-Theory: Sellarsian Constraints on the Theory of Mind

5.1. *Semantics and the Mental*

The astute reader will undoubtedly have noticed that I have so far ignored the second attraction of the theory-theory mentioned by Botterill, the *semantic attraction*. That is because the story here gets both more complex and more interesting. As Botterill notes, since theories constitute implicit definitions of the primitive terms of the theory, the theory-theory offers us an approach to what is involved in our comprehension of psychological terms. But notice that, for Sellars, this is not a particular advantage of the theory-theory approach, for Sellars possesses a general theory of meaning that is thoroughly functionalistic. For him, the fact that theoretical terms possess their meaning in virtue of functional relationships to other terms, to observations, and to action does not distinguish them from any other kinds of terms. Theoretical terms are distinguished by the fact that they have no reporting use in observation statements, but their meanings are otherwise determined in much the same way any other terms' meanings are determined. The story of Jones offers a distinctive and illuminating account of the semantics of psychological terms insofar as the Cartesian-empiricist tradition presumes that psychological terms, if any, get their meanings from direct connection to or acquaintance with the very phenomena they mean. That is, the myth of Jones can help us leave behind the old Cartesian-empiricist theory that there must be some terms that get their meaning by direct connection to the object, event, state, or property meant (i.e., by ostensive definition), and that the meanings of all other terms are constructions out of or derived from this foundational level of meaning. But once we have left the old foundationalist conception of meaning behind, we realize that there is nothing special in this regard about psychological terms.

But, I believe, there is a deeper level at which the semantic attraction is a powerful draw for Sellars. The story of Jones gives us not just a way to understand how psychological terms can be meaningful, but also a way to understand *what* meanings they have. Combined with Sellars's theory of meaning, it gives us a (theory of) psychological nomenclature, a first but very important step towards a theory of the development of and the cognitive processes underlying our psyches. But, as we will see, it does *not* give us

anything like a set of distinctive, lawlike psychological generalizations, and thus falls short itself of being the kind of psychological theory empirical psychologists are looking for.

It is part of Jones's theory that intentional states are inner states that are tied (on the action side) both to non-verbal behavior and (most importantly in this connection) to verbal episodes in which they are expressed. It is in virtue of this last fact about intentional states that the semantic categories that apply in the first instance to the linguistic episodes in which intentional states are expressed also apply by analogical extension to those intentional states themselves. Thus, Smith's inner state that would typically express itself in an assertion that it is cold outside is a belief with that same content, that it is cold outside.

Attributing meaning or content to an intentional state is analogous, according to Sellars/Jones, to giving the meaning of an utterance. So what is it to give the meaning of an utterance, according to Sellars? His analysis of meaning sentences is a very important theme in Sellars's work, but has not been given very much attention. In his view, a meaning sentence such as,

‘Gelb’ (in German) means yellow

should not be understood as stating a kind of relation between a German word type and *yellow*, (which is what: an abstract object (the property yellowness), the set of yellow objects (actual and/or possible), the scattered object that is the sum of all yellow things?). Rather, according to Sellars, ‘means’ functions in such assertions as a specialized form of the copula. That means that he construes meaning statements as statements in which we *classify* a symbol type. How does this form of classification work? Well, Sellars argues that in the above meaning statement the symbol ‘yellow’ occurs in a very special way. It certainly is not straightforwardly *used* – nothing is being described as yellow. Neither is it straightforwardly *mentioned* – the sentence is not about the English symbol ‘yellow’. There is a third way in which a symbol can occur in a sentence like a means sentence that is neither use nor mention, though it has some affinities with each. Sellars called such terms “illustrating sortals” and developed his somewhat notorious dot-quote device to capture the intention. But Sellars's extensive technical discussions of distributive singular terms, illustrating sortals, and dot-quotes did not, apparently, get his point across very well, for as I have said, his analysis of meaning has not received much attention. So let me try a different approach here.

Jane Heal (1997, 2001, unpublished) has lately been developing a related conception that many might find more intuitive than Sellars's somewhat baroque terminology. Heal points out that, though indexical reference has been given a great deal of attention, there is no reason to believe that indexicality is

confined to the phenomenon of reference alone. She therefore develops in a most interesting fashion a notion of *indexical predication*. And (as is often the case with a nice piece of philosophy), we can immediately recognize a familiar phenomenon. Suppose one is redecorating. Having chosen the central furniture group, one then seeks coordinating drapes. Some people might be able to describe in language alone the exact color and pattern to be matched, but most of us would need to bring along a sample of the fabric or a color swatch we have matched to the fabric directly. Then we can search for drapes that are *that* color, where we exhibit an example of the relevant property in conjunction with our assertion or thought. Similarly, one might describe what someone did by saying “He did this,” accompanied by an exhibition of the relevant action or motion.

Armed with this notion, we can say that Sellars thinks that meaning statements are indexical predications the subject of which is the symbol on the left-hand side (‘Gelb’ in the example above) and the predicate of which involves an indexical, illustrative, or exemplary use of something with the relevant property on the right-hand side, in this case the English word ‘yellow’. According to Sellars, the above meaning statement conveys the information that in the German language the symbol type ‘Gelb’ has a (set of) function(s) relevantly similar to those possessed by the symbol ‘yellow’ in the language of the speaker. It *conveys* this information, but it does not literally *say* it. It does not mean *in the German language the symbol type ‘Gelb’ has a (set of) function(s) relevantly similar to those possessed by the symbol ‘yellow’ in the language of the speaker*, any more than ‘He did this’ combined with a picking of one’s nose literally says that he picked his nose.¹³ When we say what an expression means, we do so by exhibiting another expression in our language with the same meaning (or as close as we can get). Although Sellars thinks that having a meaning is playing a functional role in a complex “linguistic economy” of language-entry moves, intralinguistic transitions, and language-exit moves, and that *which* meaning an expression has is a matter of which functional role it plays, we do not specify the meaning of an expression by describing which role it plays. We give the meaning by exhibiting another expression that plays a relevantly similar role. Could we say which functional

¹³ A quick test: translate both “‘Gelb’ (in German) means yellow” and “In the German language the symbol type ‘Gelb’ has a (set of) function(s) relevantly similar to those possessed by the symbol ‘yellow’ in the language of the speaker” into French. They are clearly not the same, since in the former ‘yellow’ is translated, while the ‘yellow’ that occurs in the latter would not be.

role a particular expression plays, and thus avoid the indexical predication?¹⁴ Perhaps in some cases, but if so, only a few.¹⁵

5.2. *Classificatory vs. Causal Theories*

In Sellars's view, then, when we attribute intentional states to people, we are effectively *classifying* their states. Consider a standard propositional attitude attribution, e.g.

Ralph believes that it is raining.

In Sellars's view this functions much like a piece of indirect discourse such as

Ralph asserts that it is raining.

In indirect discourse (roughly), as in meaning statements, we are classifying some item (in this case Ralph's assertion) using a kind of indexical predication in which we exhibit an item in our background language that possesses the relevantly similar property. Since Sellars thinks that having a specific meaning or content is possessing a specific functional role, such a classificatory statement will be true iff Ralph's internal state is involved in a complex functional system in which it plays a role analogous to the role assertions of 'it is raining' plays in English. Such attributions presume a complex system of causal relations, but they do not say what those causal relations are. The hard work is done (or perhaps it is avoided) by counting on our linguistic competence. Because I am an accomplished speaker and interpreter of English, I have an inarticulable sense or command of the role 'It is raining' plays. (I may be able to say many things about the proper use of the phrase, but could not finally articulate all the "rules" at work in my know-how concerning even this simple phrase.) I can employ my *linguistic* know-how to generate *psychological* insight, for my knowledge of the appropriate situations for asserting that, denying that, apologizing for, or praying that it is raining can be utilized to recognize appropriate situations for believing that, disbelieving that, regretting that, hoping that it is raining, etc., and the consequences thereof.

It is Jones's theory of intentional states that he and his comrades possess internal states with semantic properties (such as meaning) that are highly

¹⁴ Heal names kinds our principal cognitive relations to which are narrowly indexical in this way "Lagadonian," a reference to Swift and *Gulliver's Travels*. Meanings, in Sellars's view, are Lagadonian kinds.

¹⁵ Truth tables give the function the connectives have. But, notoriously, the closest English equivalents are not simply and straightforwardly the truth-functional connectives. A complete specification of the functional role of 'and' in English would be very complex.

analogous to the semantic properties of overt utterances. So he can commandeer (or perhaps better, transmute) metalinguistic forms of speech to talk about the properties of these internal states. With such a move a great deal comes along for free. We not only get a method of distinguishing in a very fine-grained way among the contents attributed to these internal states, but the notion that we can take different attitudes towards a given content. We can, that is, not only differentiate between being in internal states with content p and content q , we can distinguish being in the kind of state that is normally expressed in an assertion that p and one that is expressed in a demand that p or a request that p or a question whether p . In what is close to a fell swoop, Jones acquires a rich nomenclature of internal states, which we can perfectly well call a classificatory theory of mental states.¹⁶

Does he acquire thereby a rich *causal theory* of internal states, that is, a set of connected causal generalizations that would have empirical consequences for the developmental and cognitive processing questions? I think the answer here is clearly *no*.

Jones and his protégées, while still in the awkward early stages of mastering his theory, can reason as follows:

Ralph is in a state that he would candidly express by asserting that p , so Ralph believes that p .

Ralph is also in a state that he would candidly express by asserting that *if p , then q* , so Ralph believes that *if p , then q* .

So, *ceteris paribus*, it is likely that Ralph believes that q (though I have no direct behavioral evidence of that).

Jones's reasoning here is perfectly good and is familiar to all of us. But it is not explicitly a form of causal reasoning. Jones and company get the connection between these states of mind for free via their semantic characterization. They (and we) can assume that there are causal connections among our internal states that enable those internal states to relate to each other in ways that conform to their semantic characterization – but how those causal connections work, what kinds of mechanisms instantiate them, how they might develop and how they might break down, is left totally untouched. The “internal structures” recognized in Jonesian psychology are almost completely parasitic on the structure of the language Jones and company speak.

¹⁶ It seems worth calling it a “theory” because there are a wealth of implications about relations among the states we attribute to people with this nomenclature. We can therefore use it to generate new knowledge not otherwise available.

5.3. Constraints on “Theory of Mind”

In developing his theory, Jones has very cleverly exploited a model (language) from which he can import a vast array of distinctions, distinctions which presume causal underpinnings but make no direct commitment to them. Jones thereby jumps into a rich and complex classificatory theory of mental states that is fairly impoverished as a causal theory. However the causal story underlying our mental states gets filled in, however, it will be expected to preserve in fair degree the classificatory scheme and generic causal ties that provided the initial fruit of the theory. Thus Jones’s theory (and Sellars’s) does not have a developmental component built into it, nor a theory of cognitive processing; it is compatible with many such theories. For instance, the theory-theory would be a consistent extension and development of Jones’s theory, but it is not, as it were, an *automatic* extension of it, nor even a particularly natural extension of it.

One of the reasons Jones’s theory can be impoverished when considered as a causal theory is that we do not need to do a great deal of causal calculation in order to be able to use the theory effectively and efficiently. We can exploit the classificatory scheme the theory makes available without employing complex causal laws to connect the various states and state transitions because we can connect them *semantically*. Because of Jones’s cleverness, our mastery of our language does most of the work in our psychological competence. To flip Haugeland’s dictum around, by taking care of the semantics, we can let the syntax (the causal) take care of itself. Thus, we do not need or have (at least in the immediate post-Jonesian community) a separable psychological theory as an extensive set of lawlike causal generalizations governing interrelations among our psychological states. We learn rather to exploit our linguistic competence “in a new key.”¹⁷

If our linguistic competence is itself to be explained by the possession of a theory, then, of course, the theory theorist could still be happy. But it is not clear that Sellars believes that our mastery of language is to be explained as possession of a theory. A thoroughgoing investigation of Sellars’s understanding of linguistic competence goes beyond what we can hope to do here, but Sellars certainly rejects the Fodorian line that we have, essentially, an innate language that we command because we possess (innately) a theory of its

¹⁷ Indeed, the view we have discovered in Sellars explains why most of the time it certainly does not “feel like” we are applying a theory when we explain and predict the behavior of others: much of the work is done by whatever mechanisms are responsible for our linguistic processing, and language comprehension and performance does not “feel like” theory application either.

grammar and semantics. In the first place, any such position flies in the face of Sellars's critique of the given, for it presumes

. . . that the process of teaching a child to use a language is that of teaching it to discriminate elements within a logical space of particulars, universals, facts, etc., of which it is already indiscriminatingly aware, and to associate these discriminated elements with verbal symbols (1997, §30/p. 241).

Despite his trenchant critique of empiricism, Sellars has no intention running into the arms of an innatist rationalism, even with regard to the linguistic abilities he believes are the central core of our cognitive lives.

This point can be generalized a bit: Sellars would also have to reject the "modularist" position as it is most often conceived. Many modularists (e.g. Baron-Cohen, Swettenham) maintain that our psychological competence is to be explained by our possession of a theory, but this theory is not acquired by a complex process of theory revision. Rather, it is encapsulated in an innately specified module that develops in response to environmental stimuli. Sellars, I believe, would find modularism of this type a bit of a muddle. Such a psychology module does not develop by any standard form of learning, it *grows*; furthermore, it is a *special purpose device*, to some important degree independent of our other cognitive abilities. Yet insofar as it contains or applies a theory, it is supposed to operate by applying *concepts*, performing *inferences* and other logico-conceptual, that is, *rational* operations. But theories are subject-specific data-structures or knowledge-structures, possession of which allows us to bring our most powerful general tool, Reason, to bear upon that particular subject-matter. In Sellars's view there may be innate, genetically determined, special-purpose, relatively encapsulated capacities employed in recognizing and understanding psychological states, but it would muddle important issues to characterize these capacities as constituting a *theory*. If the "theory" that accounts for our psychological competence is not only implicit, but cut off from our general rationality both in its synchronic use and its diachronic development, what is the point of still calling it a "theory"? Of course, these remarks are consistent with the empirical truth of modularism, namely that our psychological abilities are best explained by the presence of an encapsulated, innately specified, mandatory set of processes. The point is only that explaining our psychological competence by such an innate module really abandons the idea that psychological concepts are – at the empirical as well as at the semantic or

philosophical level – significantly like theoretical concepts.¹⁸ A proponent of innate psychology modules talks of the “theory of mind” only by courtesy or convention.

Besides the oxymoronic character of talk of innate, modular “theories,” there is a deeper question: what is the relationship between innate, modular cognitive capacities – and there are without much doubt some such capacities, though which ones is still in question – and an epistemological *given*? But that is a topic that demands a paper on its own.

6. Bridging Another Gap: Theory and Simulation

From the 1980s the “theory-theory” has been contrasted to the so-called “simulation” theory, an approach that argues on philosophical grounds that we do not apply psychological concepts in any way similar to the application of theoretical concepts, because there is a distinctive kind of use of one’s own psychology to model the psychological structures attributed to others. Proponents of the simulation approach have at times attacked Sellars, hoping to discredit the father of the theory theory and similar approaches and thereby all his progeny as well (see Gordon 2000).

But if the arguments I have assembled here are right, Sellars’s own position is hardly a full-fledged theory-theory position. Rather, Sellars is committed only to the limited idea that, epistemologically, our psychological concepts are analogous to theoretical concepts; he remains steadfastly uncommitted on the causal mechanisms that underlie our use of those psychological concepts. Those causal mechanisms might be like those used in scientific theories, but they need not be. The simulationists, however, are concerned primarily about the mechanisms by which we apply our psychological concepts.

The simulationists maintain that in attributing psychological states to others we utilize our own psychological mechanisms to model the psychology of the other. As others have pointed out, this seems to assume prior and independent access to our own psychological states.¹⁹ And that smacks of a return to the inside-out Cartesian view that we have primitive, privileged access to our own psychological states as such, from which we can construct

¹⁸ For other, empirically based arguments against the notion that our psychological competence is to be explained by reference to an innate theory-constituting module, see Garfield *et. al.* (2001). They argue there that there is a perfectly acceptable sense in which our psychological competence may be modularized, but that it should not be thought of as an innate module.

¹⁹ Gordon works hard to avoid this commitment, but it is clearly evident in Goldman’s work.

other knowledge. Sellars would clearly not be happy with any such position, because he wants to be able to explain and not presuppose our access to our own psychologies, but we can nonetheless identify simulation-like elements in his view.

A Sellarsian story would not presume our ability to identify our own psychological state independently of the psychological states of others. Our ability to identify, understand, and attribute psychological states is intersubjective from the root up in a Sellarsian story. But, because many important relationships among our psychological states are mirrored in corresponding relationships between the speech acts in which their contents are expressed, language learning creates a competence that can also be recruited to perform psychological work, especially when combined with an increasingly sophisticated understanding of and competence in social interaction.²⁰

In Sellars's view, then, our attributions of psychological states rests on a kind of pre-theoretic²¹ know-how, but it is not knowing how to invoke our own psychology (which would have to be known to us in a different way) in order to understand another's, it is knowing how language and social situations work "from the inside." This knowledge can then be recruited to make sense out of both our own and other's behavior without having to take a detour through externalization in an explicit theory of psychology.²²

7. Conclusion

"Empiricism and the Philosophy of Mind" marked a major turn in the philosophy of mind. Some turned too far, for lack of a deeper understanding of the place psychological concepts play in our conceptual framework. It is unfortunate that Sellars's writing career was pretty much finished before most of the empirical research into "theory of mind" occurred and before the theory-theory vs. simulation debate really took hold. It is unlikely Sellars

²⁰ Heal's version of simulationism emphasizes its relevance to our understanding of psychological content. Given the similarities between Sellars's and Heal's treatments of content, it is not surprising to find them echoing each other in this regard.

²¹ Sellars can call these competencies "pre-theoretic" despite the Myth of Jones, because children acquire these competencies well before they acquire any conception of theory or observation. Remember, to call something "theoretical" is a methodological distinction, and so, therefore, is the notion of the pre-theoretical.

²² For a more detailed picture of how this view fits into the empirical data currently available, see the article cited above by Garfield *et. al.* (2001).

would have found any of the standard alternatives fully satisfying, and most likely he would have cast interesting new light on the debate. Thinking his position through forces us to be more careful about the alleged analogy between the theoretical and the psychological. Without such care, confusion threatens on many fronts. It is important to remember that Sellars's original philosophical point about the epistemological status of psychological concepts is separable from empirical debates about the mechanisms and the developmental process by which we come to employ those concepts.

Willem deVries
 Department of Philosophy
 University of New Hampshire
 Durham, NH 03824
 United States of America
e-mail: willem.devries@unh.edu

Visiting Fellow, Philosophy Programme
 University of London
 London WC1E 7HU
 United Kingdom

REFERENCES

- Astington, J. (1996). What Is Theoretical About the Child's Theory of Mind? A Vygotskian View of Its Development. In: P. Carruthers and P. Smith, (eds.), *Theories of Theories of Mind*, pp. 184-199. Cambridge: Cambridge University Press.
- Baron-Cohen, S. (1991). The Theory of Mind Deficit in Autism: How Specific Is It? *British Journal of Developmental Psychology* **9**, 301-314.
- Baron-Cohen, S., A. Leslie, and U. Frith (1985). Does the Autistic Child Have a Theory of Mind? *Cognition* **21**, 37-46.
- Botterill, G. (1996). Folk Psychology and Theoretical Status. In: P. Carruthers and P. Smith, (eds.), *Theories of Theories of Mind*, pp. 184-199. Cambridge: Cambridge University Press.
- Churchland, P. (1979). *Scientific Realism and the Plasticity of Mind*. Cambridge: Cambridge University Press.
- Davidson, D. (2001). *Subjective, Intersubjective, Objective*. Oxford: Oxford University Press.
- deVries, W.A. and T. Triplett (2000). *Knowledge, Mind, and the Given: Reading Wilfrid Sellars's 'Empiricism and the Philosophy of Mind'*. Indianapolis, IN: Hackett Publishing.
- Garfield, J.L., C.C. Peterson, and T. Perry (2001). Social Cognition, Language Acquisition and The Development of the Theory of Mind. *Mind & Language* **16**, 494-541.
- Gopnik, A. and H.M. Wellman (1992). Why the Child's Theory of Mind Really *Is* a Theory. *Mind and Language* **7**, 145-171.

- Gordon, R.M. (2000). Sellars's Ryleans Revisited. *Protosociology* **14**, 102-114.
- Heal, J. (1997). Indexical Predicates and their Uses. *Mind* **106**, 619-640.
- Heal, J. (2001). On Speaking Thus: The Semantics of Indirect Discourse. *The Philosophical Quarterly* **51**, 433-454.
- Heal, J. (unpublished). Lagadonian Kinds. Manuscript.
- Mitchell, P. (1997). *Introduction to Theory of Mind: Children, Autism, and Apes*. London: Arnold Publishers.
- Perner, J., T. Ruffman, and S. Leekam (1994). Theory of Mind Is Contagious: You Catch It From Your Sibs. *Child Development* **65**, 1228-1238.
- Peterson, C. and M. Siegal (2000). Insights into Theory of Mind from Deafness and Autism. *Mind and Language* **15**, 123-145.
- Premack, D. and G. Woodruff (1978). Does the Chimpanzee Have a Theory of Mind? *Behavioural and Brain Sciences* **1**, 515-526.
- Rorty, R. (1965). Mind-Body Identity, Privacy, and Categories. *The Review of Metaphysics* **19**, 24-54.
- Sellars, W. (1997). *Empiricism and the Philosophy of Mind* (with an introduction by Richard Rorty and a Study Guide by Robert Brandom). Cambridge, MA: Harvard University Press.
- Sellars, W. (1963). *Science, Perception and Reality*. London: Routledge and Kegan Paul.
- Wellman, H. M. (1990). *The Child's Theory of Mind*. Cambridge, MA: The MIT Press.