The psychology of folk psychology

Alvin I. Goldman
Department of Philosophy, University of Arizona, Tucson, AZ 85721
Electronic mail: goldmen@ccit.arizona.edu

Abstract: Folk psychology, the naive understanding of mental state concepts, requires a model of how people ascribe mental states to themselves. Competent speakers associate a distinctive memory representation (a category representation, CR) with each mentalistic word in their lexicon. A decision to ascribe such a word to oneself depends on matching to the CR an instance representation (IR) of one's current state. As in visual object recognition, evidence about a CR's content includes the IRs that are or are not available to trigger a match. This poses serious problems for functionalism, the theory-of-mind approach to the meaning of mental terms. A simple functionalist model is inadequate because (1) the relational and subjunctive (what would have happened) information it requires concerning target states is not generally available and (2) it could lead to combinatorial explosion. A modified functionalist model can appeal to qualitative (phenomenological) properties, but the earlier problems still reappear. Qualitative properties are important for sensations, propositional attitudes, and their contents, providing a model that need not refer to functional (causal-relational) properties at all. The introspectionist character of the proposed model does not imply that ascribing mental states to oneself is infallible or complete; nor is the model refuted by empirical research on introspective reports. Empirical research on "theory of mind" does not support any strict version of functionalism but only an understanding of mentalistic words that may depend on phenomenological or experiential qualities.

Keywords: folk psychology; functionalism; introspection; mental concepts; propositional attitudes; self-attribution of mental states; sensations; theory of mind

1. Introduction

The central mission of cognitive science is to reveal the real nature of the mind, however familiar or foreign that nature may be to naive preconceptions. The existence of naive conceptions is also important, however. Precisely for this reason we need cognitive science to discover what those contents are. The study of folk psychology, then, is part of the psychology of concepts. We can divide the psychology of concepts into two parts: (a) the study of conceptualization and classification in general, and (b) the study of specific folk concepts or families of folk concepts such as number concepts, material object concepts, and biological kind concepts. The study of folk psychology is a subdivision of (b), the one that concerns mental state concepts. It presupposes that mature speakers have a large number of mentalistic lexemes in their repertoire, such as happy, afraid, want, hope, pain (or hurt), doubt, intend, and so forth. These words are used in construction with other phrases and clauses to generate more complex mentalistic expressions. The question is: What is the meaning, or semantical content, of these expressions? What is it that people understand or represent by these words (or phrases)?

This target article advances two sorts of theses: methodological and substantive. The general methodological thesis is that the best way to study mental-state concepts is through the theoretico-experimental methodology of cognitive science. We should consider the sorts of data structures and cognitive operations involved in mental attributions (classifications), both attributions to oneself and attributions to others. Although this proposal is innocuous enough, it is not the methodology that has been followed, or even endorsed in principle, by philosophers who have given these questions the fullest attention. And even the cognitive scientists who have addressed these questions empirically have not used the specific methodological framework I shall recommend below.

© 1993 Cambridge University Press
Goldman: Folk psychology

In addition to methodological theses, this target article advances some substantive theses, both negative and positive. On the negative side, some new and serious problems are raised concerning the functionalist approach to mental-state concepts (as well as some doubts about pure computationalism). On the positive side, the article supports a prominent role for phenomenology in our mental-state concepts. These substantive theses are put forward tentatively because I have not done the kind of experimental work that my own methodological precepts would require for their corroboration; nor does existing empirical research address these issues in sufficient detail. Theoretical considerations, however, lend them preliminary support. I might state at the outset, however, that I am more confident of my negative thesis – about the problems facing the relevant form of functionalism – than of my positive theses, especially the role of phenomenology in the propositional attitudes.

2. Proposed methodology

Philosophical accounts of mental concepts have been strongly influenced by purely philosophical concerns, especially ontological and epistemological ones. Persuaded that materialism (or physicalism) is the only tenable ontology, philosophers have deliberately fashioned their accounts of the mental with an eye to safeguarding materialism. Several early versions of functionalism (e.g., Armstrong 1968; Lewis 1966) were deliberately designed to accommodate type-physicalism, and most forms of functionalism are construed as heavily physicalist in spirit. Similarly, many accounts of mental concepts have been crafted with epistemological goals in mind, for example, to avoid skepticism about other minds.

According to my view, the chief constraint on an adequate theory of our common sense understanding of mental predicates is not that it should have desirable ontological or epistemological consequences; rather, it should be psychologically realistic: Its depiction of how people represent and ascribe mental predicates must be psychologically plausible.

An adequate theory need not be ontologically neutral, however. As we shall see in the last section, for example, an account of the ordinary understanding of mental terms can play a significant role in arguments about eliminativism. Whatever the ontological ramifications of the ordinary understanding of mental language, however, the nature of that understanding should be investigated purely empirically, without allowing prior ontological prejudices to sway the outcome.

In seeking a model of mental-state ascription (translation), there are two types of ascriptions to consider: ascriptions to self and ascriptions to others. Here we focus primarily on self-ascriptions. This choice is made partly because I have discussed ascriptions to others elsewhere (Goldman 1989; 1992a; 1992b), and partly because ascriptions to others, in my view, are "parasitic" on self-ascriptions (although this is not presupposed in the present discussion).

Turning now to specifics, let us assume that a competent speaker/hearer associates a distinctive semantical representation with each mentalistic word, whatever form or structure this representation might take. This (possibly complex) representation, which is stored in long-term memory, bears the "meaning" or other semantical properties associated with the word. Let us call this representation the category representation (CR), since it represents the entire category the word denotes. A CR might take any of several forms (see Smith & Medin 1981), including: (1) a list of features treated as individually necessary and jointly sufficient for the predicate in question; (2) a list of characteristic features with weighted values, where classification proceeds by summing the weights of the instantiated features and determining whether the sum meets a specified criterion; (3) a representation of an ideal instance of the category, to which target instances are compared for similarity; (4) a set of representations of previously encountered exemplars of the category, to which new instances are compared for similarity; or (5) a connectionist network with a certain vector of connection weights. The present discussion is intended to be neutral with respect to these theories.

What interests us, primarily, is the semantical "contents" of the various mentalistic words, or families of words, not the particular "form" or "structure" that bears these contents.

Perhaps we should not say that a CR bears the "meaning" of a mental word. According to some views of meaning, after all, naive users of a word commonly lack full mastery of its meaning; only experts have such mastery (see Putnam 1975). But if we are interested in what guides or underpins an ordinary person's use of mental words, we want an account of what he understands or represents by that word. (What the expert knows cannot guide the ordinary person in deciding when to apply the word.) Whether or not this is the "meaning" of the word, it is what we should be after.

Whatever form a CR takes, let us assume that when a cognizer decides what mental word applies to a specified individual, active information about that individual's state is compared or matched to CRs in memory that are associated with candidate words. The exact nature of the matching process will be dictated by the hypothesized model of concept representation and categorization. Because our present focus is self-ascription of mental terms, we are interested in the representations of one's own mental states that are matched to the CRs. Let us call such an active representation, whatever its form or content, an instance representation (IR). The content of such an IR will be something like, "A current state (of mine) has features \(\phi_1, \ldots, \phi_n\)." Such an IR will match a CR having the content: \(\phi_1, \ldots, \phi_n\). Our aim is to discover, for each mental word \(M\), its associated CR; or more generally, the sorts of CRs associated with families of mental words. We try to get evidence about CRs by considering what IRs are available to cognizers, IRs that might succeed in triggering a match.

To make this concrete, consider an analogous procedure in the study of visual object-recognition; we will use the work of Biederman (1987) as an illustration. Visual object-recognition occurs when an active representation of a stimulus that results from an image projected to the retina is matched to a stimulus category or concept category, for example, chair, giraffe, or mushroom. The psychologist's problem is to answer three coordinated questions: (1) What (high-level) visual representations (corresponding to our IRs) are generated by the retinal
image? (2) How are the stimulus categories represented in memory (these representations correspond to our CRs)? (3) How is the first type of representation matched against the second so as to trigger the appropriate categories?

Biederman (1987) hypothesizes that stimulus categories are represented as arrangements of primitive components, namely, volumetric shapes such as cylinders or bricks, which he calls geons (for “geometrical ions”). Object recognition occurs by recovering arrangements of geons from the stimulus image and matching these to one of the distinguishable object models, which is paired with an entry-level term in the language (such as lamp, chair, giraffe, and so forth). The theory rests on a range of research supporting the notion that information from the image can be transformed (via edge extraction, etc.) into representations of geons and their relations. Thus, the hypothesis that, say, chair is represented in memory by an arrangement (or several arrangements) of geons is partly the result of constraints imposed by considering what information could be (a) extracted from the image (under a variety of viewing circumstances), and (b) matched to the memory representation. In similar fashion I wish to examine hypotheses about the stored representations (CRs) of mental-state predicates by reflecting on the instance representations of mental states that might actually be present and capable of producing appropriate matches.

Although we have restricted ourselves to self-ascriptions, there are still at least two types of cases to consider: ascriptions of current mental states — “I have a headache (now)” — and ascriptions of past states — “I had a headache yesterday.” Instance representations in the two cases are likely to be quite different, obviously, so they need to be distinguished. Ascriptions of current mental states, however, have primacy, so these will occupy the center of our attention.

3. Problems for functionalism

In the cognitive scientific as well as the philosophical community, the most popular account of people’s understanding of mental-state language is the “theory of mind” theory, according to which naive speakers, even children, have a theory of mental states and understand mental words solely in terms of that theory. The most precise statement of this position is the philosophical doctrine of functionalism, which states that the crucial or defining feature of any type of mental state consists of its causal relations to (1) environmental or proximal inputs, (2) other types of mental states, and (3) behavioral outputs. (Detailed examples are presented below.) Since what is at stake is the ordinary understanding of the language of the mental, the doctrine is generally called analytic, or commonsense, functionalism. Another doctrine that fits under the label “functionalism” is scientific functionalism (roughly what Block [1978] calls “psychofunctionalism”), according to which it is a matter of scientific fact that mental states are functional states: That is, mental states have functional properties (causal relations to inputs, other mental states, and outputs) and should be studied in terms of these properties. I shall have nothing to say against scientific functionalism. I do not doubt that mental states have functional properties; nor do I challenge the proposal that mental states should be studied (at least in part) in terms of these properties. But this doctrine does not entail that ordinary people understand or represent mental words as designating functional properties and functional properties only. States designated by mental words might have functional properties without ordinary folk knowing this, or without their regarding it as crucial to the identity of the states. But since we are concerned here exclusively with the ordinary person’s representations of mental states, only analytic functionalism is relevant to our present investigation.

Philosophers usually discuss analytic, or commonsense, functionalism quite abstractly, without serious attention to its psychological realization. I am asking us to consider it as a psychological hypothesis, that is, a hypothesis about how the cognizer (or his cognitive system) represents mental words. It is preferable, then, to call the type of functionalism in question representational functionalism (RF). This form of functionalism is interpreted as hypothesizing that the CR associated with each mental predicate M represents a distinctive set of functional properties, or functional role, \( F_M \). Thus, RF implies that a person will ascribe a mental predicate M to himself when and only when an IR occurs in him bearing the message, “role \( F_M \) is now instantiated.” That is, ascription occurs precisely when there is an IR that matches the functional-role content of the CR for M. (This may be subject to some qualification. Ascription may not require perfect or complete matching between IR and CR; partial matching may suffice.) Is RF an empirically plausible model of mental self-ascription? In particular, do subjects always get enough information about the functional properties of their current states to self-ascribe in this fashion (in real time)?

Before examining this question, let us sketch RF in more detail. The doctrine holds that folk wisdom embodies a theory, or a set of generalizations, which articulates an elaborate network of relations of three kinds: (a) relations between distal or proximal stimuli (inputs) and internal states, (b) relations between internal states and other internal states, and (c) relations between internal states and items of overt behavior (outputs). Here is a sample of such laws from Churchland (1979). Under heading (a) (relations between inputs and internal states) we might have:

- When the body is damaged, a feeling of pain tends to occur at the point of damage.
- When no fluids are imbibed for some time, one tends to feel thirsty.
- When a red apple is present in daylight (and one is looking at it attentively), one will have a red visual experience.

Under heading (b) (relations between internal states and other internal states) we might have:

- Feelings of pain tend to be followed by desires to relieve that pain.
- Feelings of thirst tend to be followed by desires for potable fluids.

If one believes that \( P \), where \( P \) elementarily entails \( Q \), one also tends to believe that \( Q \).

Under heading (c) (relations between internal states and outputs) we might have:
Goldman: Folk psychology

Sudden sharp pains tend to produce wincing. States of anger tend to produce frowning. An intention to curl one's finger tends to produce the curling of one's finger.

According to RF, each mental predicate picks out a state with a distinctive collection, or syndrome, of relations of types (a), (b), and/or (c). The term pain, for example, picks out a state that tends to be caused by bodily damage, tends to produce a desire to get rid of that state, and tends to produce wincing, groaning, and so on. The content of each mental predicate is given by its unique set of relations, or functional role, and nothing else. In other words, RF attributes to people a purely relational concept of mental states.

There are slight variations and important additional nuances in the formulation of functionalism. Some formulations, for example, talk about the causal relations among stimulus inputs, internal states, and behavioral outputs. Others merely talk about transitional relations, that is, one state following another. Another important wrinkle in an adequate formulation is the subjunctive or counterfactual import of the relations in question. For example, part of the functional role associated with desiring water would be something like this: If a desire for water were accompanied by a belief that a glass of water is within arm's reach, then (other things being equal) it would be followed by extending one's arm. To qualify as a desire for water, an internal state need not actually be accompanied by a belief that water is within reach, nor need it be followed by an extending of the arm. It must, however, possess the indicated subjunctive property: If it were accompanied by this belief, the indicated behavior would occur.

We are now in a position to assess the psychological plausibility of RF. The general question I wish to raise is, Does a subject who self-ascribes a mental predicate always (or even typically) have the sort of instance information required by RF? This is similar to an epistemological question sometimes posed by philosophers, namely, whether functionalism can give an adequate account of one's knowledge of one's own mental state. But the present discussion does not center on knowledge; it merely asks whether the RF model of the CRs and IRs in mental self-ascription adequately explains this behavior. Does the subject always have functional-role information about the target states — functional-role IRs — to secure a "match" with functional-role CRs?

There are three sorts of problems for the RF model. The first is ignorance of causes and effects (or predecessor and successor states). According to functionalism, what makes a mental state a state of a certain type (e.g., a pain, a feeling of thirst, a belief that 7 + 5 = 12, and so forth) is not any intrinsic property it possesses but its relations to other states and events. What makes a state a headache, for example, includes the environmental conditions or other internal states that actually cause or precede it, and its actual effects or successors. There are situations, however, in which the self-ascription of headache occurs in the absence of any information (or beliefs) about relevant causes or effects, predecessors or successors. Surely there are cases in which a person wakes up with a headache and immediately ascribes this type of feeling to himself. Having just awakened, he has no information about the target state's immediate causes or predecessors, nor need he have any information about its effects or successors. The classification of the state occurs "immediately," without waiting for any further effects, either internal or behavioral, to ensue. There are cases, then, in which self-ascription occurs in the absence of information (or belief) about critical causal relations.

It might be replied that a person need not appeal to actual causes or effects of a target mental state to type-identify it. Perhaps he determines the state's identity by its subjunctive properties. This brings us to the second problem confronting the RF model: ignorance of subjunctive properties. How is a person supposed to determine (form beliefs about) the subjunctive properties of a current state (instance or "token")? To use our earlier example, suppose the subject does not believe that a glass of water is within arm's reach. How is he supposed to tell whether his current state would have produced an extending of his arm if this belief were present? It is extremely difficult to get information about subjunctive properties, unless the RF model is expanded in ways not yet intimated (a possible expansion will be suggested in sect. 4). The subjunctive implications of RF, then, are a liability rather than an asset. Each CR posited by RF would incorporate numerous subjunctive properties, each presumably serving as a necessary condition for applying a mental predicate. How is a cognizer supposed to form IRs containing properties that match those subjunctive properties in the CR? Determining that the current state has even one subjunctive property is difficult enough; determining many such properties is formidably difficult. Is it really plausible, then, that subjects make such determinations in type-identifying their inner states? Do they execute such routines in the brief time-frames in which self-ascriptions actually occur? This seems unlikely. I have no impossibility proof, of course; but the burden is on the RF theorist to show how the model can handle this problem.

The third difficulty arises from two central features of functionalism: (1) The type-identity of a token mental state depends exclusively on the type-identity of its relata, that is, the events that are (or would be) its causes and effects, its predecessors and successors, and (2) the type-identity of an important subclass of a state's relata, namely, other internal states, depends in turn on their relata. To identify a state as an instance of thirst, for example, one might need to identify one of its effects as a desire to drink. Identifying a particular effect as a desire to drink, however, requires one to identify its relata, many of which would also be internal states whose identities are a matter of their relata; and so on. Complexity ramifies very quickly. There is no claim here of any vicious circularity, or vicious regress. If functionalism is correct, the system of internal state-types is tacked down definitively to independently specified external states (inputs and outputs) via a set of lawful relations. Noncircular definitions (so-called Ramsey definitions) can be given of each functional state-type in terms of these independently understood input and output predicates (see Block 1978; Lewis 1970; 1972; Loar 1981; Putnam 1967). The problem I am raising, however, concerns how a subject can determine which functional type a given state-token
instantiates. There is a clear threat of combinatorial explosion: Too many other internal states will have to be type-identified in order to identify the target state.

This problem is not easily quantified with precision, because we lack an explicitly formulated and complete functional theory, hence we do not know how many other internal states are directly or indirectly invoked by any single functional role. The problem is particularly acute, though, for beliefs, desires, and other propositional attitudes, which under standard formulations of functionalism have strongly "holistic" properties. A given belief may causally interact with quite a large number of other belief tokens and desire tokens. To type-identify that belief, it looks as if the subject must track its relations to each of these other internal states, their relations to further internal states, and so on until each path terminates in an input or an output. When subjunctive properties are added to the picture the task becomes unbounded, because there is an infinity of possible beliefs and desires. For each desire or goal-state there are indefinitely many beliefs with which it could combine to produce a further desire or subgoal. Similarly, for each belief there are infinitely many possible desires with which it could combine to produce a further belief. If the type-identification of a target state depends on tracking all of these relations until inputs and outputs are reached, clearly it is unmanageably complex. At a minimum, we can see this as a challenge to an RF theorist, a challenge that no functionalist has tried to meet, and one that looks forbidding.

Here the possibility of partial matching may assist the RF theorist. It is often suggested that visual object identification can occur without the IR completely matching the CR. This is how partially occluded stimuli can be categorized. Biederman (1987; 1990) argues that even complex objects, whose full representation contains six or more geons, are recognized accurately and fairly quickly with the recovery of only two or three of these geons from the image. Perhaps the RF theorist would have us appeal to a similar process of partial matching to account for mental-state classification.

Although this might help a little, it does not get around the fundamental difficulties raised by our three problems. Even if only a few paths are followed from the target state to other internal states and ultimately to inputs and/or outputs, the demands of the task are substantial. Nor does the hypothesis of partial matching address the problem of determining subjunctive properties of the target state. Finally, it does not help much when classification occurs with virtually no information about neighboring states, as in the morning headache example. Thus, the simple RF model of mental self-ascription seems distinctly unpromising.

4. A second functionalist model

A second model of self-ascription is available to the RF theorist, one that assumes, as before, that for each mental predicate there is a functional-role CR. (That is what makes an RF model functional.) This second model, however, tries to explain how the subject determines which functional role a given state-token exemplifies without appealing to on-line knowledge of the state's current relations. How, after all, do people decide which objects exemplify other dispositional properties, for example being soluble in water? They presumably do this by inference from the intrinsic and categorical (i.e., non-dispositional) properties of those objects. When a person sees a cube of sugar, he may note that it is white, hard, and granular (all intrinsic properties of the cube), infer from background information that such an object must be made of sugar, and then infer from further background information that it must be soluble in water (because all sugar is soluble in water). Similarly, the RF theorist may suggest that a subject can detect certain intrinsic and categorical properties of a mental state, and from this he can infer that it has a certain functional property, that is, a suitable relational and dispositional property.

Let us be a bit more concrete. Suppose that the CR for the word headache is the functional-role property $F$. Further suppose that there is an intrinsic (nonrelational) property $E$ that mental states have, and the subject has learned that any state which has $E$ also has the functional-role property $F$. Then the subject is in a position to classify a particular headache as a headache without any excessively demanding inference or computation. He just detects that a particular state-token (his morning headache, for example) has property $E$, and from this he infers that it has $F$. Finally, he infers from its having $F$ that it can be labeled headache.4

Although this may appear to save the day for RF, it actually just pushes the problem back to what we may call "the learning stage." A crucial part of the foregoing account is that the subject must know (or believe) that property $E$ is correlated with property $F$ – that whenever a state has $E$ it also has $F$. But how could the subject have learned this? At some earlier time, during the learning stage, the subject must have detected some number of mental states, each of which had both $E$ and $F$. But during this learning period he did not already know that $E$ and $F$ are systematically correlated. So he must have had some other way of determining that the $E$-states in question had $F$. How did he determine that? The original difficulties we cited for identifying a state's functional properties would have been at work during the learning stage, and they would have been just as serious then as we saw them to be in the first model. So the second model of functionalist self-ascription is not much of an improvement (if any) over the first.

In addition, the second model raises a new problem (or question) for RF. What are the intrinsic properties of mental states that might play the role of property $E$? At this point let us separate our discussion into two parts, one dealing with what philosophers call sensation predicates (roughly, names for bodily feelings and percepts), and the other dealing with propositional attitudes (believing that $p$, hoping that $q$, intending to r, etc.). In this section we restrict our attention to sensation predicates; in section 8 we shall turn to predicates for propositional attitudes.

What kinds of categorical, nonrelational properties might fill the role of $E$ in the case of sensations? In addition to being categorical and nonrelational, such properties must be accessible to the system that performs

Goldman: Folk psychology

BEHAVIORAL AND BRAIN SCIENCES (1993) 16:1
Goldman: Folk psychology

the self-ascription. This places an important constraint on the range of possible properties.

There seem to be two candidates to fill the role of $E$: (1) neural properties and (2) what philosophers call “qualitative” properties (the “subjective feel” of the sensation). Presumably any sensation state or event has some neural properties that are intrinsic and categorical, but do these properties satisfy the accessibility requirements? Presumably not. Certainly the naive subject does not have “personal access” (in the sense of Dennett 1969; 1978b) to the neural properties of his sensations. That would occur only if the subject were, say, undergoing brain surgery and watching his brain in a mirror. Normally people do not see their brains; nor do they know much, if anything, about neural hardware. Yet they still identify their headaches without any trouble.

It may be replied that although there is no personal access to neural information in the ordinary situation, the system performing the self-ascription may have subpersonal access to such information. To exclude neural properties (i.e., neural concepts) from playing the role of $E$ we need reasons to think that self-ascription does not use these properties of sensations. It goes without saying that neural events are involved in the relevant information processing; all information processing in the brain is, at the lowest level, neural processing. The question, however, is whether the contents (meanings) encoded by these neural events are about neural properties. This, to repeat, seems implausible. Neural events process visual information, but cognitive scientists do not impute neural contents to these neural events. Rather, they consider the contents encoded to be structural descriptions, things like edges and vertices (in low-level vision) or geons (in higher-level vision). When connectionists posit neurally inspired networks in the analysis of language, they do not suppose that configurations of connection weights encode neural properties (e.g., configurations of connection weights), but rather things like phonological properties.

There is more to be said against the suggestion that self-ascription is performed by purely subpersonal systems, which have access to neural properties. Obviously, a great deal of information processing does occur at subpersonal levels within the organism. But when the processing is purely subpersonal, no verbal labels seem to be generated that are recognizable “mental.” There are all sorts of homeostatic activities in which information is transmitted about levels of certain fluids or chemicals; for example, the glucose level is monitored and then controlled by secretion of insulin. But we have no folk psychological labels for these events or activities. Similarly, there are information-processing activities in low-level vision and in the selection and execution of motor routines. None of these, however, are the subjects of primitive (pretheoretic) verbal labeling, certainly not “mentalistic” labeling. This strongly suggests that our spontaneous naming system does not have access to purely subpersonal information. Only when physiological or neurological events give rise to conscious sensations such as thirst, felt heat, or the like does a primitive verbal label get introduced or applied. Thus, although there is subpersonal detection of properties such as “excess glucose,” these cannot be the sorts of properties to which the mentalistic verbal-labeling system has access.

We seem to be left, then, with what philosophers call “qualitative” properties. According to the standard philosophical view, these are indeed intrinsic, categorical properties that are detected “directly.” Thus, the second model of functional self-ascription might hold that in learning to ascribe a sensation predicate like itch, one first learns the functional role constitutive of that word’s meaning (e.g., being a state that tends to produce scratching, and so forth). One then learns that this functional role is realized (at least in one’s own case) by a certain qualitative property: itchiness. Finally, one decides that the word is self-ascribable whenever one detects in oneself the appropriate qualitative property, or quale, and infers the instantiation of its correlated functional role. This model still depicts the critical IR as a representation of a functional role, and similarly depicts the CR to which the IR is matched.

We have found a kind of property, then, which might plausibly fill the role of $E$ in the second functionalist model. But is this a model that a true functionalist would welcome? Functionalists are commonly skeptical about qualia (e.g., Dennett 1988; 1991a; Harman 1990). In particular, many of them wish to deny that there are any qualitative properties if these are construed as intrinsic, nonrelational properties. But this is precisely what the second model of RF requires: that qualitative properties be accepted as intrinsic (rather than functional) properties of mental states. It is not clear, therefore, how attractive the second model would be to many functionalists.

5. A qualitative model of sensation representation

Furthermore, although the second functionalist model may retain some appeal (despite its problems with the learning stage), it naturally invites a much simpler and more appealing model: one that is wholly nonfunctionalist. Once qualitative properties are introduced into the psychological story, what need is there for functional-role components? Why not drop the latter entirely? Instead, we hypothesize that both the CR and the IR for each sensation predicate are representations of a qualitative property such as itchiness (or, as I suggest below, some microcomponents of the property of itchiness). This vastly simplifies our story, for both the learning phase and the ascription phase itself. All one learns in the learning phase is the association between the term itch and the feeling of itchiness. At the time of self-ascription, all one detects (or represents) is itchiness, and then matches this IR to the corresponding CR. This is a very natural model for the psychology of sensation self-ascription, at least to anyone free of philosophical prejudice or preconception.

Of course, some philosophers claim that qualitative properties are “ queer,” and should not be countenanced by cognitive science. There is nothing objectionable about such properties, however, and they are already implicitly countenanced in scientific psychology. One major text, for example, talks of the senses producing sensations of different “quality” (Gleitman 1981, p. 172). The sensations of pressure, A-flat, orange, or sour, for example, are sharply different in experienced quality (as Gleitman puts it). This use of the term quality refers to differences across the sensory domains, or sense modal-
Person introspectively into its several components or constitutional decisions to apply or withhold the predicate for imaginary cases. This supports the contention that no purely functional content exhausts the meaning of these sensation words; qualitative character is an essential part of that content.

On a methodological note, I should emphasize that the use of thought experiments, so routine in philosophy, may also be considered (with due caution) a species of psychological or cognitivist methodology, complementary to the methodology described earlier in this article. Not only do applications of a predicate to actual cases provide evidence about the correlated CR, but so do decisions to apply or withhold the predicate for imaginary cases. In the present context, reactions to hypothetical cases support our earlier conclusion that qualitative properties are the crucial components of CRs for sensation words.

Quite a different question about the qualitative approach to sensation concepts should now be addressed, namely, its compatibility with our basic framework for classification. This framework says that self-ascription occurs when a CR is matched by an IR, where an IR is a representation of a current mental state. Does it make sense, however, to regard an instance of a qualitative property as a representation of a mental state? Is it not more accurate to say that it is a mental state, not a representation thereof? If we seek a representation of a mental state, should we not look for something entirely distinct from the state itself (or any feature thereof)?

Certainly the distinction between representations and what they represent must be preserved. The problem can be avoided, however, by a minor revision in our framework. On reflection, self-ascription does not require the matching of an instance representation to a category representation; it can involve the matching of an instance itself to a category representation. The term instance representation was introduced because we wanted to allow approaches like functionalism, in which the state itself is not plausibly matched to a CR, only a representation of it. Furthermore, we had in mind the analogy of perceptual categorization, where the cognizer does not match an actual stimulus to a mental representation of the stimulus category but an inner representation of the stimulus to a category representation.

In this respect, however, the analogy between perceptual recognition and sensation recognition breaks down. In the case of sensation there can be a matching of the pain itself, or some features of the pain, to a stored structure containing representations of those features. Thus, we should revise our framework to say that categorization occurs when a match is effected between (1) a category representation and (2) either (a) a suitable representation of a state or (b) a state itself. Alternative (b) is especially plausible in the case of sensations, because it is easy to suppose that CRs for sensations are simply mem-

Goldman: Folk psychology

BEHAVIORAL AND BRAIN SCIENCES (1993) 16:1 21
ory “traces” of those sensations, which are easily activated by reoccurrences of the same (or similar) sensations.

This new picture might look suspicious because it seems to lead to the much-disparaged doctrines of infallibility and omniscience about one’s own mental states. If a CR is directly matched to an instance of a sensation itself, is all possibility of error not precluded? And would it not be impossible to be unaware of one’s sensations, because correct matching is inevitable? Yet surely both error and ignorance are possible.

The proposed change implies neither infallibility nor omniscience. The possibility of error is readily guaranteed by introducing an assumption (mentioned earlier) of partial matching. If a partial match suffices for classification and self-ascription, there is room for inaccuracy. If we hypothesize that the threshold for matching can be appreciably lowered by various sorts of “response biases” (such as prior expectation of a certain sensation), this makes error particularly easy to accommodate. Ignorance can be accommodated in a different way, by supplementary assumptions about the role of attention. When attentional resources are devoted to other topics, there may be no attempt to match certain sensations to any category representation. Even an itch or a pain can go unnoticed when attention is riveted on other matters. Mechanisms of selective attention are critical to the full story of classification, but this large topic cannot be adequately addressed here.

Even with these points added, some readers might think that our model makes insufficient room for incidental cognitive factors in the labeling of mental states. Does not the work on emotions by Schachter and Singer (1962), for example, show that such incidental factors are crucial? My first response is that I am not trying to address the complex topic of emotions but restricting attention to sensations (in this section) and propositional attitudes (in sect. 8). Second, there are various ways of trying to accommodate the Schachter-Singer findings. One possibility, for example, is to say that cognitive factors influence which emotion is actually felt (e.g., euphoria or anger) rather than the process of labeling or classifying the felt emotion (see Wilson 1991). So it is not clear that the Schachter-Singer study would undermine the model proposed here, even if this model were applied to emotions (which is not my present intent).

6. The classical functionalist account of self-ascription

Classical functionalists such as Putnam (1960), Sellars (1956), and Shoemaker (1975) have not been oblivious to the necessity of making room in their theory for self-ascriptions or self-reports of mental states. They thought this could be done without recourse to anything like qualitative properties. How, then, does their theory go, and why do I reject it?

According to the classical account, it is part of the specification of a mental state’s functional role that having the state guarantees a self-report of it; or, slightly better, that it is part of the functional specification of a mental state (e.g., pain) that it gives rise to a belief that one is in that state (Shoemaker 1975). If one adopts the general framework of RF that I have presented, however, it is impossible to include this specification. Let me explain why.

According to our framework, a belief that one is in state \( M \) occurs precisely when a match occurs between a CR for \( M \) and a suitable IR. (Since we are now discussing functionalism again, we need not worry about “direct” matching of state to CR.) But classical functionalism implies that part of the concept of being in state \( M \) (a necessary part) is having a belief that one is in \( M \). Thus, no match can be achieved until the system has detected the presence of an \( M \)-belief. However, to repeat, what an \( M \)-belief is, according to our framework, is the occurrence of a match between the CR for \( M \) and an appropriate IR. Thus, the system can only form a belief that it is in \( M \) (achieve an IR-CR match) by first forming a belief that it is in \( M \! \). Obviously this is impossible.

What this point shows is that there is an incompatibility between our general framework and classical functionalism. They cannot both be correct. But where does the fault lie? Which should be abandoned?

A crucial feature of classical functionalism is that it offers no story at all about how a person decides what mental state he is in. Being in a mental state automatically entails, or gives rise to, the appropriate belief. Precisely this assumption of automaticity has until now allowed functionalism to ignore the sorts of questions raised in this paper. In other words, functionalism has hitherto tended to assume some sort of “nonrecognitional” or “noncriterial” account of self-reports. Allegedly, you do not use any criterion (e.g., the presence of a qualitative property) to decide what mental state you are in. Classification of a present state does not involve the comparison of present information with anything stored in long-term memory. Just being in a mental state automatically triggers a classification of yourself as being in that state.

It should be clear, however, that this automaticity assumption cannot and should not be accepted by cognitive science, for it would leave the process of mental-state classification a complete mystery. It is true, of course, that we are not introspectively aware of the mechanism by which we classify our mental states. But we are likewise not introspectively aware of the classification processes associated with other verbal labeling, the labeling of things such as birds or chairs, leavings or stockings. Lack of introspective access is obviously no reason for cognitive scientists to deny that there is a microstory of how we make — or how our systems make — mental classifications.

There must be some way a system decides to say (or believe) that it is now in a thirst state rather than a hunger state, that it is hoping for a rainy day rather than expecting a rainy day. That is what our general framework requires. In short, in a choice between our general framework and classical functionalism (with its assumption of automatic self-report), cognitive science must choose the former. Any tenable form of functionalism, at least any functionalism that purports to explain the content of naive mental concepts, must be formulated within this general framework. This is just how RF has been formulated. It neither assumes automaticity of classification nor does it create a vicious circularity (by requiring the prior detection of a classification event — a belief — as a necessary condition for classification). So RF is superior to classical functionalism for the purposes at hand. Yet RF, we have seen, has
7. Dual representations

Thus far we have assumed that there is a single CR (however complex) paired with each mental word. This assumption, however, is obviously too restrictive. Indeed, there are many lines of research that suggest the possibility of dual representations for a single word. For example, Biederman (1987; 1990) hypothesizes that a single word (e.g., piano) is associated with two or more visual object-models, (e.g., a model for a grand piano and a model for an upright). In addition, however, there may be a wholly nonvisual representation associated with piano (e.g., keyboard instrument that has a certain characteristic sound). Some neuropsychological evidence for dual representations has been provided by Warrington and her coworkers. Warrington and Taylor (1978) studied patients with right-hemisphere and left-hemisphere lesions. They were given two sorts of tasks: one involving perceptual categorization by physical identity and the other involving semantic categorization by functional identity. The right-hemisphere group showed impairment on the perceptual categorization task. The left-hemisphere group showed no impairment on this task but they did show impairment on the second task. For example, they could not pair (pictures of) two deck chairs that were physically dissimilar but had the same function. Warrington and Taylor accordingly postulate two anatomically separate, postsensory categorical stages in object recognition, which occur in sequence. In a somewhat similar vein, Cooper, Schacter, and their colleagues have found evidence of distinct forms of representation of visual objects: a purely structural representation stored in implicit memory and a semantical/functional representation available to explicit memory (Cooper 1991; Schacter et al. 1990; 1991). Other types of evidence or arguments for dual representations have been given by Landau (1982), McNamara and Sternberg (1983), Putnam (1975), and Rey (1983). [See also Bradshaw & Nettleton: “The Nature of Hemispheric Specialization in Man” BBS 4(1) 1981; and Shallice’s précis of From Neuropsychology to Mental Structure, BBS 14(3) 1991.]

If the idea of dual representation is applied to sensation terms, or mental terms generally, it might support a hybrid feature that features both a qualitative representation and an independent, nonqualitative representation, which might be functional. The qualitative representation could accommodate self-ascriptions and thereby avert the problems we posed for pure functionalism. However, it is not clear how happy functionalists would be with this solution. As we have remarked, most functionalists are skeptics about qualia. Any dual-representation theory that invokes qualitative properties (not functionally reconstructed) is unlikely to find favor with them.

Furthermore, functionalist accounts of mental concepts face another difficulty even if the functional representation is only one member of a subject’s pair of representations. According to functional-style definitions of mental terms, the meaning of each term is fixed by the entire theory of functional relations in which it appears (see Block 1978; Lewis 1972; Loar 1981). This implies that every little change anywhere in the total theory—every addition of a new law or revision of an old law—entails a new definition of each mentalistic expression. Even the acquisition of a single new mental predicate requires amending the definition of every other such predicate, since a new predicate introduces a new state-type that expands the set of relations in the total theory. This holistic feature of functionalism entails all-pervasive changes in one’s repertoire of mental concepts. Such global changes threaten to be as computationally intractable as the familiar “frame problem” (McCarthy & Hayes 1969), especially because there is a potential infinity of mental concepts, owing to the productivity of “that” clauses and similar constructions. The problem of conceptual change for theory-embedded concepts is acknowledged and addressed by Smith, Carey, and Wiser (1985), but they are concerned with tracking descent for individual concepts, whereas the difficulty posed here is the computational burden of updating the vast array of mental concepts implicated by any theoretical revision.

8. A phenomenological model for the attitudes?

Returning to my positive theory, I have thus far only proposed a qualitative approach to sensation concepts. Let us turn now from sensations to propositional attitudes, such as believing, wanting, and intending. This topic can be divided into two parts (Fodor 1987): the representation of attitude types and the representation of attitude contents. Wanting there to be peace and believing there will be peace are different attitudes because their types (wanting and believing) are different. Intending to go shopping and intending to go home are different attitudes because, although their type is the same, their contents differ. In this section we consider how attitude types are represented; in the next we consider attitude contents.

Philosophical orthodoxy favors a functionalist approach to attitude types. Even friends of qualia (e.g., Block 1990a) feel committed to functionalism when it comes to desire, belief, and so forth. Our earlier critiques of functionalism, however, apply with equal force here. Virtually all of our antifunctionalist arguments (except the absent-qualia arguments) apply to all types of mental predicates, not just to sensation predicates. So there are powerful reasons to question the adequacy of functionalism for the attitude types. How, then, do people decide whether a current state is a desire rather than a belief, a hope rather than a fear?

In recent literature some philosophers use the metaphor of “boxes” in the brain (Schiffer 1981). To believe something is to store a sentence of mentalese in one’s “belief box”; to desire something is to store a sentence of mentalese in one’s “desire box”; and so on. Should this metaphor be taken seriously? I doubt it. It is unlikely that there are enough “boxes” to have a distinct one for each attitude predicate. Even if there are enough boxes in the brain, does the ordinary person know enough about these neural boxes to associate each attitude predicate with one of them (the correct one)? Fodor (1987) indicates that box-talk is just shorthand for treating the attitude types in functional terms. If so, this just reintroduces the forbidding problems already facing functionalism.
Goldman: Folk psychology

Could a qualitative or phenomenological approach work for the attitude types? The vast majority of philosophers reject this approach out of hand, but this rejection is premature. I shall adduce several tentative(!) arguments in support of this approach.

First a definitional point. The terms *qualia* and *qualitative* are sometimes restricted to sensations (percepts and somatic feelings), but we should not allow this to preclude the possibility of other mental events (beliefs, thoughts, etc.) having a phenomenological or experiential dimension. Indeed, at least two cognitive scientists (Baars 1988; Jackendoff 1987) have defended the notion that "abstract" or "conceptual" thought often occupies awareness or consciousness, even if it is phenomenologically "thinner" than modality-specific experience. Jackendoff appeals to the tip-of-the-tongue phenomenon to argue that phenomenology is not confined to sensations. When one tries to say something but cannot think of the word, one is phenomenologically aware of having a requisite conceptual structure, that is, of having a determined thought-content one seeks to articulate. What is missing is the phonological form: the sound of the sought-for word. The absence of this sensory quality, however, does not imply that nothing (relevant) is in awareness. Entertaining the conceptual unit has a phenomenology, just not a sensory phenomenology.

Second, in defense of phenomenal "parity" for the attitudes, I present a permutation of Jackson's (1982; 1986) argument for qualia (cf. Nagel 1974). Jackson argues that qualitative information cannot be captured in physicalist (including functionalist) terms. Imagine, he says, that a brilliant scientist named Mary has lived from birth in a cell where everything is black, white, or gray. (Even she herself is painted all over.) By black-and-white television she reads books, engages in discussion, and watches experiments. Suppose that by this means Mary learns all physical and functional facts concerning color, color vision, and the brain states produced by exposure to colors. Does she therefore know all facts about color? There is one kind of fact about color perception, says Jackson, of which she is ignorant: what it is like (i.e., what it feels like) to experience red, green, and so on. These qualitative sorts of facts she will come to know only if she actually undergoes spectral experiences.

Jackson's example is intended to dramatize the claim that there are subjective aspects of sensations that resist capture in functionalist terms. I suggest a parallel style of argument for attitude types. Just as someone deprived of any experience of colors would learn new things upon being exposed to them, namely, what it feels like to see red, green, and so forth, so (I submit) someone who had never experienced certain propositional attitudes, for example, doubt or disappointment, would learn new things on first undergoing these experiences. There is "something it is like" to have these attitudes, just as much as there is "something it is like" to see red. In the case of the attitudes, just as in the case of sensations, the features to which the system is sensitive may be microfeatures of the experience. This still preserves parity with the model for sensations.

My third argument is from the introspective discriminability of attitude strengths. Subjects' classification abilities are not confined to broad categories such as belief, desire, and intention; they also include intensities thereof. People report how *firm* is their intention or conviction, how *much* they desire an object, and how satisfied or dissatisfied they are with a state of affairs. Whatever the behavioral predictive power of these self-reports, their very occurrence needs explaining. Again, the functionalist approach seems fruitless. The other familiar device for conceptualizing the attitudes—namely, the "boxes" in which sentences of mentalese are stored—would also be unhelpful even if it were separated from functionalism, since box storage is not a matter of degree.

The most natural hypothesis is that there are dimensions of awareness over which scales of attitude intensity are represented.

The importance of attitude strength is heightened by the fact that many words in the mentalistic lexicon ostensibly pick out such strengths. *Certain*, *confident*, and *doubtful* represent positions on a credence scale; *delighted*, *pleased*, and *satisfied* represent positions on a liking scale. Since we apparently have introspective access to such positions, self-ascription of these terms invites an introspectivist account (or a quasi-introspectivist account that makes room for microfeatures of awareness).

One obstacle to a phenomenological account of the attitudes is that stored (or dispositional) beliefs, desires, and so on are outside awareness. However, there is no strain in the suggestion that the primary understanding of these stems from their activated ("occurrent") incarnations; the stored attitudes are just dispositions to have the activated ones.

A final argument for the role of phenomenology takes its starting point from still another trouble with functionalism, a trouble not previously mentioned here. In addition to specific mental words like *hope* and *imagine*, we have the generic word *mental*. Ordinary people can classify internal states as mental or nonmental. Notice, however, that many nonmental internal states can be given a functional-style description. For example, *having measles* might be described as a state that tends to be produced by being exposed to the measles virus and tends to produce an outbreak of red spots on the skin. So having measles is a functional state; although, it clearly is not a *mental* state. Thus, functionalism cannot fully discharge its mission by saying that mental states are functional states; it also needs to say *which* functional states are *mental*. Does functionalism have any resources for marking the mental/nonmental distinction? The prospects are bleak. By contrast, a plausible-looking hypothesis is that mental states are states having a phenomenology, or an intimate connection with phenomenological events. This points us again in the direction of identifying the attitudes in phenomenological terms.

Skepticism about this approach has been heavily influenced by Wittgenstein (1953; 1967), who questioned whether there is any single feeling or phenomenal characteristic common to all instances of an attitude such as intending or expecting. (A similar worry about sensations is registered by Churchland & Churchland 1981.) Notice, however, that our general approach to concepts does not require a single "defining characteristic" for each mental word. A CR might be, for example, a list of exemplars (represented phenomenologically) associated with the word, to which new candidate instances are compared for similarity. Thus, even if Wittgenstein's (and Churchland & Churchland) worries about the phenomenological
9. Content and computationalism

The commonsense understanding of the contents of the propositional attitudes is an enormous topic; we shall touch on it here only lightly. The central question concerns the "source" of contentfulness for mental states. Recent theories have tended to be externalist, claiming that content arises from causal or causal-historical interactions between inner states (or symbols in the language of thought) and external objects. It is highly doubtful, however, whether any of the most developed externalist theories gives an adequate account of the naive cognizer's understanding or representation of content.

Fodor currently advocates a complex causal-counterfactual account (Fodor 1987, 1990). Roughly, a mental symbol C means cow if and only if (1) C-tokens are reliably caused by cows, and (2) although noncows (e.g., horses) also sometimes cause C-tokens, noncows would not cause C-tokens unless cows did, whereas it is false that cows would not cause C-tokens unless noncows did. Clause (2) of this account is a condition of "asymmetric dependence" according to which there being non-cow-caused C-tokens depends on there being cow-caused C-tokens, but not conversely. It seems most implausible, however, that this sort of criterion for the content of a mental symbol is what ordinary cognizers have in mind. Similarly implausible for this purpose are Millikan's evolutionary account of mental content (Millikan 1984; 1986) and Dretske's (1988) learning-theoretic (i.e., operant conditioning) account of mental content. Most naive cognizers have never heard of operant conditioning, and many do not believe in evolution; nevertheless, the same subjects readily ascribe belief contents to themselves. So did our sixteenth-century ancestors, who never dreamt of the theory of evolution or operant conditioning. (For further critical discussion see Cummins 1989.) Perhaps Millikan and Dretske do not intend their theories as accounts of the ordinary understanding of mental contents. Millikan (1989), for one, expressly disavows any such intent. But then we are left with very few detailed theories that do address our question. Despite the popularity of externalist theories of content, they clearly pose difficulties for self-ascription. Cognizers seem able to discern their mental contents — what they believe, desire, or plan to do — without consulting their environment.6

What might a more internalist approach to contents look like? Representationalism, or computationalism, maintains that content is borne by the formal symbols of the language of thought (Fodor 1975; 1981; 1987; Newell & Simon 1976). But even if the symbolic approach gives a correct de facto account of the working of the mind, it does not follow that the ordinary concept of mental content associates it with formal symbols per se. I would again suggest that phenomenological dimensions play a crucial role in our naive view. Only what we are aware or conscious of provides the primary locus of mental content.

For example, psycholinguists maintain that in sentence processing there are commonly many interpretations of a sentence that are momentarily presented as viable, but we are normally aware of only one — the one that gets selected (Garrett 1990). The alternatives are "filtered" by the processing system outside of awareness. Only in exceptional cases such as "garden path" sentences (e.g., "Fatty weighed three hundred and fifty pounds of grapes") do we become aware of more than one considered interpretation. Our view of mental content is, I suggest, driven by the cases of which we are aware, although they may be only a minority of the data structures or symbolic structures that occupy the mind.

Elaboration of this theme is not possible in the present paper, but a brief comment about the relevant conception of "awareness" is in order. Awareness, for these purposes, should not be identified with accessibility to verbal report. We are often aware of contents that we cannot (adequately) verbalize, either because the type of content is not easily encoded in linguistic form or because its mode of cognitive representation does not allow full verbalization. The relevant notion of awareness, or consciousness, then, may be that of qualitative or phenomenological character (there being "something it is like") rather than verbal reportability (see Block 1990b; 1991).

The role I am assigning to consciousness in our naive conception of the mental bears some similarity to that assigned by Searle (1990). Unlike Searle, however, I see no reason to decree that cognitive science cannot legitimately apply the notion of content to states that are inaccessible (even in principle) to consciousness. First, it is not clear that the ordinary concept of a mental state makes consciousness a "logical necessity" (as Searle puts it). Second, even if mental content requires consciousness, it is inessential to cognitive science that the nonconscious states to which contents are ascribed should be considered mental. Let them be "psychological" or "cognitive" rather than "mental"; this does not matter to the substance of cognitive science. Notice that the notion of content in general is not restricted to mental content; linguistic utterances and inscriptions are also bearers of content. So even if mental content is understood to involve awareness, this places no constraints of the sort Searle proposes on cognitive science.

10. Empirical research on the theory-theory

The idea underlying functionalism, that the naive cognizer has a "theory" of mind, goes increasingly by the label "the theory-theory" (Morton 1980). Although originally proposed by philosophers, this idea is now endorsed by a preponderance of empirical researchers, especially developmental psychologists and cognitive anthropologists. Does their research lend empirical support to the functionalist account of mental concepts?

Let us be clear about exactly what we mean by functionalism, especially the doctrine of RF (representational functionalism) that concerns us here. There are two crucial features of this view. The first feature is pure relationalism. RF claims that the way subjects represent mental predicates is by relations to inputs, outputs, and other internal states. The other internal-state concepts are similarly represented. Thus, every internal-state concept is ultimately tied to external inputs and outputs. What is deliberately excluded from our understanding of
mental predicates, according to RF, is any reference to the phenomenology or experiential aspects of mental events (unless these can be spelled out in relationalist terms). No "intrinsic" character of mental states is appealed to by RF in explaining the subject's basic conception or understanding of mental predicates. The second crucial feature of RF is the appeal to nomological (lawlike) generalizations in providing the links between each mental-state concept and suitably chosen inputs, outputs, and other mental states. Thus, if subjects are to exemplify RF, they must mentally represent laws of the appropriate sort. Does empirical research on "theory of mind" support either of these two crucial features? Let us review what several leading workers in this tradition say on these topics. We shall find that very few of them, if any, construe theory of mind in quite the sense specified here. They usually endorse vaguer and weaker views.

Premack and Woodruff (1978), for example, say that an individual has a theory of mind if he simply imputes mental states to himself and others. Ascriptions of mental states are regarded as "theoretical" merely because such states are not directly observable (in others), and because such imputations can be used to make predictions about the behavior of others. This characterization falls short of RF, because it does not assert that the imputations are based on lawlike generalizations and does not assert that mental-state concepts are understood solely in terms of relations to external events.

Wellman's (1988; 1990) concept of the theory-theory (TT) is also quite a weak one. A body of knowledge is theory-like, he says, if it has (1) an interconnected ("coherent") set of concepts, (2) a distinctive set of ontological commitments, and (3) a causal-explanatory network. Wellman grants that some characterizations of theories specify commitments to nomological statements, but his own conception explicitly omits that provision (Wellman 1990, Chap. 5). This is one reason why his version of TT falls short of RF. A second reason is that Wellman explicitly allows that the child's understanding of mind is partly founded on firsthand experience. "The meaning of such terms/images as belief, desire and dream may be anchored in certain firsthand experiences, but by age three children have not only the experiences but the theoretical constructs" (Wellman 1990, p. 195). Clearly, then, Wellman's view is not equivalent to TT, and the evidence he adduces for his own version of TT is not sufficient to support RF.

Similarly, Rips and Conrad (1989) present evidence that a central aspect of people's beliefs about the mind is that mental activities are interrelated, with some activities being kinds or parts of others. For example, reasoning is a kind of thinking and reasoning is a part of problem solving. The mere existence of taxonomies and paronyms (part-whole hierarchies), however, does not support RF, since mental terms could still be represented in introspective terms, and such taxonomies may not invoke laws.

D'Andrade (1987) also describes the "folk model of the mind" as an elaborate taxonomy of mental states, organized into a complex causal system. This is no defense of functionalism, however, since D'Andrade expressly indicates that concepts such as emotion, desire, and intention are "primarily defined by the conscious experience of the person" (p. 139). The fact that laymen recognize causal relations among mental events does not prove that they have a set of laws. Whether or not belief in causal relations requires belief in laws is a controversial philosophical question. Nor does the fact that people use mental concepts to explain and predict the behavior of others imply the possession of laws, as we shall see below.

The TT approach to mental concepts is, of course, part of a general movement toward understanding concepts as theory-embedded (Carey 1985; 1988; Premack & Woodruff 1978; Karmiloff-Smith & Inhelder 1975; Keil 1989; Murphy & Medin 1985). Many proponents of this approach, however, acknowledge that their construal of "theory" is quite vague, or remains to be worked out. For example, Murphy and Medin (1985, p. 290) simply characterize a theory as "a complex set of relations between concepts, usually with a causal basis"; and Keil (1989, pp. 279-80) says: "So far we have not made much progress on specifying what naive theories must look like or even what the best theoretical vocabulary is for describing them." Thus, commitment to a TT approach does not necessarily imply commitment to RF in the mental domain; nor would evidential corroboration of a TT approach necessarily corroborate RF.

A detailed defense of TT is given by Gopnik (this issue), who specifically rejects the classical view of direct or introspective access to one's own psychological states. However, even Gopnik's view is significantly qualified and her evidential support far from compelling. First, although her main message is the rejection of an introspective or "privileged access" approach to self-knowledge of mental states, she acknowledges that we use some mental vocabulary "to talk about conscious experiences with a particular kind of phenomenology, the Joycean or Woolfian stream of consciousness, if you will." This does not sound like RF. Second, Gopnik seems to concede privileged access, or at least errorless performance, for subjects' self-attributions of current mental states. At any rate, all of her experimental data concern self-attributions of past mental states; nowhere does she hint that subjects make mistakes about their current states as well. How can errorless performance be explained on her favored inferential model of self-attribution? If faulty theoretical inference is rampant in children's self-attribution of past states, why do they not make equally faulty inferences about their current states? Third, there is some evidence that children's problems with reporting their previous thoughts are just a result of memory failure. Mitchell and Lacoheé (1991) found that such memory failure could be largely alleviated with a little help. Fourth, Gopnik's TT does not explain satisfactorily why children perform well on self-attributions of past pretend and imaging. Why are their inferences so much more successful for those mental states than for beliefs? Finally, how satisfactory is Gopnik's explanation of the "illusion" of first-person privileged access? If Gopnik were right that this illusion stems from expertise, why should we not have the same illusion in connection with attribution of mental states to others? If people were similarly positioned vis-à-vis their own mental states and those of others, they would be just as expert for others as for themselves and should develop analogous illusions; but there is no feeling of privileged access to others' mental states.

At this point the tables might be turned on us. How are
The positive approach to mental concepts I have ten-
tatively endorsed has much in common with the classical introspectionist, to my knowledge, has ever asserted. Moreover, Nisbett and Wilson explicitly con-
cede direct access to many or most of the private states that concern us here and that concern philosophy of mind in general.

We do indeed have direct access to a great storehouse of private knowledge. . . . The individual knows a host of personal historical facts; he knows the focus of his attention at any given point of time; he knows what his current sensations are and has what almost all psycholo-
gists and philosophers would assert to be "knowledge" at least quantitatively superior to that of observers concerning his emotions, evaluations, and plans. (Nisbett & Wilson, 1977, p. 255)

Their critique of introspectionism, then, is hardly as encompassing as it first appears (or as citations often suggest). As White (1988) remarks, "causal reports could turn out to be a small island of inaccuracy in a sea of insight" (p. 37).

Nisbett and Wilson’s paper reviewed findings from several research areas including attribution, cognitive dissonance, subliminal perception, problem solving, and bystander apathy. Characteristically, the reported findings were of manipulations that produced significant differences on behavioral measures but not on verbal self-report measures. In Nisbett and Wilson’s position-effect study, for example, passersby appraised four identical pairs of stockings in a linear display and chose the pair they judged of best quality. The results showed a strong preference for the rightmost pair. Subjects did not report that position had influenced their choice and vehemently denied any such effect when the possibility was mentioned.

However, as Bowers (1984) points out, this sort of finding is not very damaging to any sensible form of introspectionism. As we have known since Hume (1748), causal connections between events cannot be directly observed; nor can they be introspected. A sensible form of introspectionism, therefore, would not claim that people have introspective access to causal connections, but this leaves it open that they do have introspective access to the mere occurrence of certain types of mental events.

Other critics, for example Ericsson and Simon (1980), complain that Nisbett and Wilson fail to investigate or specify the conditions under which subjects are unable to make accurate reports. Ericsson and Simon (1980; 1984) themselves develop a detailed model of the circumstances in which verbal reports of internal events are likely to be accurate. In particular, concurrent reports about information that is still in short-term memory (STM) and fully attended are more likely to be reliable than retrospective reports. In most of the studies reviewed by Nisbett and Wilson, however, the time lag between task and probe was sufficiently great to make it unlikely that relevant information remained in STM. A sensible form of introspectionism would restrict the thesis of privileged access to current states and not extend it to past mental events. Of course, people often do have long-term memories of their past mental events. But their direct access is then to these memories, not to the original mental events themselves.

In a more recent work, one of the two authors, T. D. Wilson, has been very explicit in accepting direct access. He writes: "People often have direct access to their
Goldman: Folk psychology

mental states, and in these cases the verbal system can make direct and accurate reports. When there is limited access, however, the verbal system makes inferences about what these processes and states might be” (Wilson 1985, p. 16). He then explores four conditions that foster imperfect access, with the evident implication that good access is the default situation. This sort of position is obviously quite compatible with the one advocated in the present target article.

With its emphasis on conscious or phenomenological characteristics, the present article appears to be challenged by Velmans (1991a). Velmans raises doubts about the role of consciousness in focal-attentive processing, choice, learning and memory, and the organization of complex, novel responses. His target article seems to conjecture that consciousness does not enter causally into human information processing at all.

However, as many of his commentators point out, Velmans’s evidence does not support this conclusion. Block (1991) makes the point particularly clear. Even if Velmans is right that consciousness is not required for any particular sort of information processing, it does not follow that consciousness does not in fact figure causally. Block also sketches a plausible model, borrowed from Schacter (1989), in which consciousness does figure causally. In the end, this debate may be misplaced, since Velmans (1991), in his response to commentators, says that he did not mean to deny that consciousness has causal efficacy. Velmans’s views aside, the position sketched in the present article invites questions about how, exactly, qualitative or phenomenological properties can figure in the causal network of the mind. This raises large and technical issues pertaining not only to cognitive science but also to philosophical questions about causation, reduction and identity, supervenience, and the like. Such issues require an entirely separate paper (or more than one), and cannot be addressed here.

12. Ontological implications of folk psychology

These technical issues aside, the content of folk psychology may have significant ontological implications about the very existence of the mental states for which common speech provides labels. Eliminativists maintain that there are no such things as beliefs, desires, thoughts, hopes, and other such intentional states (Churchland 1981; Stich 1983). They are all, like phlogiston, caloric, and witches, the mistaken posits of a radically false theory, in this case a commonsense theory. The argument for eliminativism proceeds from the assumption that there is a folk theory of mind. This article counters the argument by denying (quite tentatively, to be sure) that our mental concepts rest on a folk theory.

It should be stressed that the study of folk psychology does not by itself yield ontological consequences. It just yields theses of the form: Mental (or intentional) states are necessarily conceptualized as states of kind K. This sort of thesis, however, may appear as a premise in arguments with eliminativist conclusions, as we have just seen. If kind K features nomological relations R, for example, one can defend eliminativism by holding that no states actually instantiate relations R. On the other hand, if K just includes qualitative or phenomenological properties, it is harder to see how a successful eliminativist argument could be mounted. One would have to hold that no such properties are instantiated or even exist. Although qualitative properties are indeed denied by some philosophers of mind (e.g., Dennett 1989; 1991; Harman 1990), the going is pretty tough (for replies to Dennett and Harman, respectively, see Flanagan 1992, Ch. 4, and Block 1990a).

The general point, however, should be clear. Although the study of folk psychology does not directly address ontological issues, it is indirectly quite relevant to such issues.

Apart from ontological implications, the study of folk psychology also has intrinsic interest as an important subfield of cognitive science. This, of course, is the vantage-point from which the present discussion has proceeded.

ACKNOWLEDGMENTS

I am grateful to the following people for very helpful comments and discussion: Kent Bach, Paul Bloom, Paul Boghossian, Robert Cummins, Carl Ginet, Christopher Hill, John Kihlstrom, Mary Peterson, Sydney Shoemaker, and Robert Van Gulick, as well as the BBS referees. I also profited from discussions of earlier versions of the paper at the Society for Philosophy and Psychology, the Creighton Club, CUNY Graduate Center, Cornell University, Yale University, University of Connecticut, and Brown University.

NOTES

1. Type-physicalism is the doctrine that every mental state type is identical with some neurophysiological state type.

2. Eliminativism is the philosophical view that certain types of mental states invoked by everyday language, especially the propositional attitudes, would not appear in a fully mature science of the mind and should therefore be “eliminated” from our ontology.

3. Since a person does not have direct introspective access to the contents of his CRs, he may be unable to report or even recognize F,M as the content paired with M. This, at any rate, is what a defender of RF might point out. We do seem to have limited access, however, to the meanings or contents of our words. Hence, if ordinary people have no recognition whatsoever of functional roles as the meanings or contents of their mental predicates, this (in my opinion) would be some evidence, though not decisive evidence, against the hypothesis that these are their meanings or contents.

4. Note that our framework allows the RF theory to exclude E from the content of the CR for headache despite the fact that E is used as “evidence” for a headache. This shows that our framework is not verificationist: It does not equate something’s being evidence for a type of mental state with its being in the CR for that type of mental state.

5. Block describes an imaginary scenario in which a system involving the entire Chinese nation (one billion people strong) is functionally equivalent for an hour to a single human mind. Although each Chinese person in this system has inner states with qualitative character, it is doubtful that the entire system in question would be conscious, or have inner states with qualitative character, despite the system’s functional equivalence to an individual human mind.

6. Solutions to this apparent puzzle are proposed by Davidson (1987) and Burge (1988); but the former is less than transparent and the latter is convincingly shown by Boghossian (1989) to be inadequate. The considerations of this target article support Boghossian’s contention, in contrast to Burge’s, that knowledge of one’s own mental contents cannot be merely automatic or “cognitively insubstantial” (as Boghossian puts it).