FOLK PSYCHOLOGY AND MENTAL CONCEPTS

Alvin I. Goldman

University of Arizona

I.

By "folk psychology" I mean the commonsense understanding and deployment of mentalistic concepts, especially the propositional attitudes. Three (or four) principal questions are of interest in the study of folk psychology: (Q1) How do ordinary people understand, or represent to themselves, the various mental states? That is, what are the contents of their concepts of the mental states? (Q2) How do they go about attributing these states? This question decomposes into two subquestions: (Q2)(A) how do people attribute such states to others, and (Q2)(B) how do they attribute such states to themselves? (3) How do people acquire their concepts of mental states and their skill at applying these concepts? The study of folk psychology is the attempt to answer these questions, an attempt which should be (and is) a combined effort among philosophers and cognitive scientists. Large literatures on the topic already exist in both philosophy and developmental psychology, and other branches of cognitive and neuroscience also have much to contribute.

The study of folk psychology might profitably be viewed as a branch of "descriptive epistemology", since epistemology is (partly) concerned with how beliefs are formed (Goldman 1986) and folk psychology involves the formation of beliefs about a special subject-matter: mental states. By contrast with other portions of philosophy of mind, the study of folk psychology is not a first-order examination of the nature and constitution of the attitudes (and other mental states). Rather, it is interested in second-order attitudes: beliefs (and concepts) of the folk about the attitudes. These two subjects, of course, are connected. A first step in trying to identify the true nature of mental states, it may well be suggested, is to identify the ordinary understanding of mental states, or what ordinary people mean by mental-state terms. What needs

emphasis, however, is that the study of folk psychology is not directly or primarily concerned with the "proper" metaphysical status of mental states, only with the folk's understanding (or misunderstanding) of mental-state phenomena. For example, it is a metaphysical question whether mental states are identical to neural states. But if mental-neural identity is not part of the folk's understanding of mental states, then this matter does not come directly under the purview of the study of folk psychology.[i]

Admittedly, metaphysical questions have traditionally been paramount in philosophy of mind. Are mental states identical with brain states? Do they supervene on the physical states of the agent, or on a wider class of physical states? Jerry Fodor surveys a large set of metaphysical questions in "Fodor's Guide to Mental Representation" (Fodor 1985): Are propositional attitudes "real"? If so, are they functional states? Do they essentially involve relations to other entities such as propositions, sentences of natural language, or sentences of a language of thought? These metaphysical questions will not occupy me here, because they do not belong to the study of folk psychology as I construe it. To be sure, I adopt the working assumption that beliefs exist (in some sense) because the subject-matter of folk psychology presupposes that ordinary people have beliefs about the attitudes; but this assumption is in principle defeasible.

Three general approaches to folk psychology will be examined and assessed in the present discussion: the theory theory (TT), the simulation theory (ST), and the rationality (or charity) theory (RT). The first two of these approaches will take center stage, partly because they have been most prominent in recent years. Philosophical formulations of TT and RT are not usually presented in exactly the guise they are considered here. For example, analytical functionalism, which is a version of TT, is typically presented only as an answer to question (Q1), not as an answer to either (Q2) or (Q3). But I am interested, at least in principle, in how TT would answer all three (or four) of the questions we have posed. Our three questions are not

[i] Again, there are important connections between the the study of folk psychology and metaphysics. If mental-state concepts are theoretical concepts, for example, it seems easier to mount an argument for (metaphysical) eliminativism by arguing that the supposed laws believed by the folk are not genuine laws at all, and no states instantiate them.

wholly unrelated, of course. The content of mental-state concepts may constrain how these concepts can be plausibly deployed in attribution, and evidence about their contents can be obtained by inquiring into their manner of deployment. Which attributional operations are used is (evidentially) relevant to what the contents of the concepts might be.

II.

The first problem I shall explore is how the three positions on the table should be defined. This is a non-trivial exercise. Within each position--especially TT and ST--different adherents characterize their approaches differently, even when they share the same label. It is standardly assumed that the three approaches are disjoint, but some definitions might imply otherwise. Finally, the possible features that might be used in defining any of the positions are quite varied. A given approach might be defined in terms of its answer to (Q1), its answer(s) to (Q2) (both (A) and (B)), its answer to (Q3), or by some combination of such answers.

Theory theory. At least two distinct forms of TT are on offer: a strict and a loose version. The strict version--essentially equivalent to analytical functionalism--says that the commonsense understanding of mental states is wholly in terms of their place in a folk-psychological theory. A theory is construed as a substantial set of laws or generalizations, in the present case, laws connecting various mental states with other mental states, with external circumstances, and with overt behaviors. This version of TT is presented, for example, by David Lewis (1972) and Paul Churchland (1988). Churchland gives a sampling of the folk generalizations or platitudes of the type in question, including: "Persons tend to feel pain at points of recent bodily damage," "Persons denied fluids for some time tend to feel thirst," "Persons in pain tend to want to relieve that pain," and "Persons who want that P, and believe that Q would be sufficient to bring about P, and have no conflicting wants or preferred strategies, will try to bring it about that Q." (1988: 58-59) TT suggests that ordinary people's concepts of mental states are exhausted by their understanding of the roles these postulated states play within such a set of laws.[ii]

[ii] Notice that TT cannot be distinctively characterized as the metaphysical thesis that mental states have functional properties, or even have them essentially. It is mainly the conceptual (or semantic) thesis--that the concepts of mental states are specified by their functional roles--that

Defenders of ST have challenged the strict version of TT by raising doubts about whether all mentalistic attributors possess the sorts of generalizations or platitudes that are usually invoked (Gordon 1986; Heal 1986; Goldman 1989). Mentalistic attributors include children under age (say) six, whose putative knowledge (whether explicit or implicit) of such laws is especially open to question. In light of these challenges, perhaps, certain defenders of TT soften their commitments by omitting any requirement of laws or generalizations. Stich and Nichols (1992) opt for a loose version of TT (what they call a "wide" interpretation), formulated as follows: "just about any internally stored body of information about a domain [is] an internally represented theory of that domain" (1992: 46). This is a somewhat tendentious definition of TT, however, because it implies that a position ascribing to attributors any information about mental states is an instance of TT. This definition threatens to swallow up virtually all opposition, by counting only extremely radical positions as part of the opposition at all. This is not a fruitful way to divide up the available options, because it lumps almost every position into a single category.

If we want to avoid the definition-by-laws characterization of TT, but not be so loose as the Stich-Nichols definition, how might TT be characterized? The historical motivation behind TT was to provide an account of mental concepts that fits with traditional empiricist notions of cognitive or semantic respectability.[iii] The construal of theoretical terms in science was the model on which TT was patterned. So one component of the core conception of TT is that the content of mental concepts should be empirically definable, where this means that the basic terms of the definitions should refer to publicly observable events.[iv] Indeed, it is common for

---

distinguishes TT from alternative positions. Thus, Jerry Fodor embraces the metaphysical thesis of functionalism but rejects conceptual, or semantic, functionalism: "[D]enying, as a point of semantics, that "believe" has a functional definition is compatible with asserting, as a point of metaphysics, that belief has a functional essence." (1998: 8)

[iii] It is noteworthy that the first statement of TT was in Wilfrid Sellars's paper "Empiricism and the Philosophy of Mind" (Sellars 1963).

[iv] I here make the working assumption that the contents of people's mental-state concepts come in the form of definitions. This assumption is quite controversial, of course, in the philosophical and cognitive science literatures. However, I only make this as a working assumption, and mean

proponents of TT to remark that mental states are "unobservables" and must be understood by their relations to observables. The psychologists Alison Gopnik and Henry Wellman write: "All these characteristics of theories ought also to apply to children's understanding of mind, if such understandings are theories of mind. That is, such theories should involve appeal to abstract unobservable entities, with coherent relations among them" (1992: 148, emphasis added).

As our discussion of functionalism indicates, TT standardly appeals to specific sorts of publicly observable events to which an understanding of mental states is tied, viz., the subject's behavior and stimulus conditions in her environment. These are, of course, the sorts of observables to which (logical) behaviorism traditionally appealed, and functionalism is a direct descendant of behaviorism. Thus, TT endorses the idea that an understanding of mental states is founded on appropriate logical and/or epistemological relations to publicly observable peripheral events. TT thereby avoids any reliance on "qualitative" or introspectible properties, which scientific empiricists characteristically regard as either epistemologically suspect, metaphysically dangerous (as a threat to physicalism), or both. Tying mental-state concepts to purely peripheral events effects a purely third-person conception of mentality, a perspective congenial to traditional empiricism because it avoids epistemologically problematic issues of "privacy".

If mentalistic concepts are grasped in terms of peripheral observables and their relata, there follows a natural-seeming answer to question (Q2). People go about the business of attributing mental states by inferring such states from observed peripheral events in accordance with the theoretical connections, i.e., the laws allegedly known to folk attributors. This is the standard TT story about mental attributions to others; and paradigmatic proponents of TT endorse the same story about self-attribution of mental states (Gopnik 1993). So we might define TT in terms of its answers to two questions. In answer to (Q1), it says that mental-state (MS) concepts are defined wholly in terms of their relationships to peripheral observables. In

---

to be open to other formats for the contents of concepts (see Goldman 1993). The general problems of what concepts are and what it means to possess them transcend the scope of this paper. See Peacocke 1992 and Fodor 1998.

answer to (Q2), it says that beliefs about MSs are formed by making inferences to their occurrence from observations of peripheral events, using the posited nomological relationships.

There are several problems with this way of defining TT. Is TT committed to holding that all cases of MS attribution feature nomological inferences of the types in question? Fuller (1995) points out that TT might accept the simulation method of MS attribution as an "epistemological tool", or shortcut. This possibility would be excluded by the second part of the currently proposed definition. One could, of course, simply delete this second element from the definition. The trouble with the resulting proposal is that even ST might accept the first part alone, viz., that MS types are defined in terms of relationships to peripheral events. At least one proponent of ST, viz., Gordon, might be willing to accept this constraint, yet he is no theory-theorist.

A different way to configure TT is to drop the second part and replace it with some constraint on the account of concept acquisition. TT might be characterized as holding that the child's formulation and grasp of MS concepts results from general-purpose scientizing procedures (in this case, processes of concept formation). This characterization would fit with the views of Gopnik and Meltzoff (1997), for example, and Gopnik and Wellman (1992). This definition, however, would exclude developmentalists of the modularist or nativist persuasion (Leslie 1991; Leslie and German 1995; Baron-Cohen 1995), who have typically presented themselves as proponents of TT. These psychologists reject the "child scientist" account of MS-concept acquisition.[v]

Rationality theory. RT partly derives from Quine's (1960) articulation of a principle of charity in the context of his approach to "radical translation". Variants of his charity principle

[v] Another interesting question in the definition of TT is whether an approach that drops the peripheralist-empiricist constraint but retains the nomological inference account of attribution should qualify as a version of TT. Consider an account that configures MS-concepts in terms of directly introspectible (or internally detectable) features rather than publicly observable, peripheral evemts. Using these concepts people acquire nomological beliefs (based solely on evidence about their own case) relating introspectible events to one another and to behavior. Using these laws, they make inferences (whether justified or not) to mental states of other people. Would such a position qualify as a specimen of TT?

have been endorsed by others, most prominently by Davidson (1984) and Dennett (1987), as an account of how attributors "interpret" others, i.e., assign intentional states to them. The core idea is that attributors make assumptions about the normative propriety of having certain intentional states under various conditions. Rationality is assumed to impose certain requirements on the propositional attitudes a person should have. For example, (1) rationality requires agents to believe truths; (2) it requires their belief-sets to be coherent; (3) it requires their belief-sets to be closed under entailment; and (4) it requires their desires to be aimed at things it is good for them to have. Interpretation proceeds by making the charitable assumption that people usually comply with these normative principles. That is, attributors allegedly assign intentional states on the assumption that their targets mostly conform with these dictates of rationality.[vi] Thus, according to RT, normative principles of rationality guide the process of intentional attribution in roughly the way nomological generalizations allegedly guide attribution according to (strict versions of) TT. According to Dennett, for example, to adopt the "intentional stance" toward people (or other systems) consists in assigning intentional states to them in accordance with a default assumption that they conform with principles of rationality.

On the surface RT is a theory of attribution, a response to question (Q2).[vii] Is there also a complementary answer to question (Q1) that RT can offer? What, according to RT, do the folk mean or understand by a desire for q or a belief that p? Adherents of RT are often silent or evasive on this question, but here are two possible approaches. First, RT might portray the folk understanding of these states in terms of the specific rules of rationality it imputes to attributors. For example, attributors might understand a belief that p as a state one ought to be in if one has other beliefs that entail p and one ought not be in if one has other beliefs inconsistent with p. A desire for q might be understood as a state one ought to be in if q is good for one. And so forth.

[vi] For a reading of Davidson, Lewis (1974), and Dennett as endorsing this approach, see Fodor and LePore 1992: 142-144.

[vii] That Dennett intends his theory as a theory of how people actually deploy their intentional concepts is clear in the following passage: "Do people actually use this strategy [the strategy of the intentional stance]? Yes, all the time." (1987: 21)

A second possible answer might appeal to the general idea of the intentional strategy. This is suggested by Dennett in the following passage:

> [A]ll there is to being a true believer is being a system whose behavior is reliably predictable via the intentional strategy, and hence all there is to really and truly believing that p ... is being an intentional system for which p occurs as a belief in the best (most predictive) interpretation. (1987: 29)

A possible implication of this passage is that what the folk understand by a belief is a state that enables an attributor to make reliable predictions of its possessor's behavior via the intentional strategy. The adequacy of both this approach and the preceding approach to answering (Q1) will be addressed in section III.

How is RT related to TT and ST? Presumably, RT contrasts with TT in its emphasis on normativity. Of course, if one adopts the loosest possible definition of TT, i.e., the one cited above from Stich and Nichols 1992, then RT qualifies as a species of TT. But if more stringent conditions are imposed for being an instance of TT, e.g., conditions that exclude normativity from being a defining trait of intentional states, then RT and TT will be mutually exclusive categories.

What about the relationship between RT and ST? This relationship is often quite intimate. Some theorists who characterize themselves as simulationists endorse rationality postulates as a core component of MS attribution. Heal (1998) is a clear example of this. Coming from the opposite direction, when defenders of RT are pressed on the question of exactly which rationality postulates are utilized, they sometimes fall back into a position indistinguishable from simulationism. I'll return to this point below.

What about ST itself? What are its core commitments? The distinctive element of simulationism is its answer to (Q2)(A), concerning third-person MS attribution. The standard ST position is that one makes attributions by putting oneself in the target's shoes and modeling her resultant mental activity (Gordon 1986; Goldman 1989). More precisely, one creates "pretend" states of oneself intended to correspond to initial states of the target. Next one feeds these

pretend states into a cognitive mechanism of one's own--e.g., a practical or factual reasoning mechanism--and lets it operate on inputs so as to produce a new state, for example, a decision or a belief. Finally, one attributes this newly outputted state to the target. Some simulationists, however, take exception to this depiction. Specifically, Heal (1998) prefers to characterize the simulation account as holding that the attributor must "co-cognize" with the target, but that co-cognition need not involve feeding pretend states into cognitive mechanisms.

Simulationists disagree even more sharply about the method of first-person MS attribution. I endorse an introspectivist or direct monitoring approach to first-person (contemporaneous) MS attribution (Goldman 1993, forthcoming). Gordon rejects such an account and opts for an ascent-routine story (Gordon 1992, 1995, 1996), which will be examined below. Analogous differences persist on the topic of the contents of MS concepts. Thus, simulationists share common ground only on the third-person attribution question, and even on that question, unanimity is far from perfect.

III.

In this section I shall briefly present some serious problems facing TT and RT. These criticisms do not purport to be definitive--that would be too much to expect for the complexity of the subject and the limits of space. But they are major stumbling blocks for the rivals of ST, which help tilt me and should tilt others toward ST. I begin with problems for RT.

The first problem facing RT is its failure to address an entire subclass of the mental states, viz., the sensations. The sensations are surely understood and attributed by the folk; but since there appear to be no questions of "rationality" pertaining to pains or tickles, for example, RT comes up empty in trying to explain how attributors ascribe them.

A second problem for RT is the implausibility of either of its approaches to attitude concepts. The first approach, recall, is to analyze these concepts in terms of the posited norms of rationality. An initial problem with this approach is a threat of circularity. To say that believing p is a state one ought not be in if one holds other beliefs inconsistent with p is to explain one type of belief-state in terms of other belief-states, which looks flagrantly circular. This might be

avoided by positing a recursive specification of beliefs in which the foregoing serves as a recursive clause but other formulas provide base-clauses, e.g., believing p is a state one ought to be in if state-of-affairs p obtains in one's immediate environment and one is looking in the appropriate direction. The question is whether this recursive approach can provide a sufficiently complete characterization of our attitude concepts. Isn't there more to our concepts of beliefs and desires than when we ought to have them?

A second problem is that, in general, analyzing the concept of an attitude or action in terms of norms for its adoption or performance seems to get things backwards. We ordinarily understand norms for an action of type X only if we independently understand what it is to do X. To grasp a library's rules for when to return books, one must already understand what a book is and what counts as returning one. It doesn't seem possible to reverse the order of explanation and explain what it means to return a book solely in terms of rules that specify when to return them.

Many theorists, of course, hold that meanings involve norms.[viii] But the normative approach to meaning poses two problems in this context. First, what attitudes or actions do the norms of meaning govern? Are they propositional attitudes like belief, for example? If so, there is another threat of circularity, because the targets of our theorizing--attitude concepts like belief--are presupposed in the theoretical account of them. Second, even if it is granted that the meanings of concepts in general are to be explained in terms of norms, a second level of norms would have to be invoked to explain the specific meanings of the attitude concepts. This second level is what would distinguish the contents of the attitude concepts from the contents of other (descriptive) concepts such as the concepts of "chair" or "mountain". It is at this level that my previous worries arise.

The second possible RT approach to attitude concepts, floated in the previous section, is to say that the attitudes are understood in terms of their predictive utility when adopting the

[viii] For a recent example, see Brandom (1994).

intentional stance. According to this account, however, possessing the belief concept or the desire concept would require one to possess the concept of the intentional strategy. Is this plausible? Does everyone who attributes belief or desire (and therefore possesses these concepts) also possess the concept of the intentional strategy? Notice that this concept is a highly reflective one. It requires a possessor not only to be a user of the intentional strategy but to understand, reflectively, what that strategy consists in. But according to developmental psychologists (e.g., Bartsch and Wellman 1995), two-year-olds use the desire concept and three-year-olds use the belief concept. So, according to the present approach, two- and three-year-olds must possess the concept of making-reliable-predictions-via-the-intentional-strategy. This consequence is hard to swallow.

A different type of problem for RT concerns the specifics of its rationality norms and the questionable grasp of those norms by ordinary people, even children. As mentioned above, quite young children show substantial mastery of attribution skills in their attitude ascriptions. According to RT, then, these children must understand and grasp the canons of rationality that RT postulates. Is it really plausible to suppose that they grasp the general notions of logical consistency and deductive closure? Actually, it is doubtful whether even untrained adults grasp these notions. Many scientific studies of deductive reasoning challenge the notion that untrained adults approach such tasks with abstract semantical or proof-theoretic concepts of the sorts used in formal logic (Cheng and Holyoak 1985, Cosmides 1989). Similarly, psychological studies of decision and choice challenge the notion that naive people utilize standard normative models (Tversky and Kahneman 1986).

When challenged by worries of these sorts, Dennett often backs away from the logic-based rules that he and other RT proponents had previously postulated. One should not "cling", he says, "to the ideals of Intro Logic for one's model of rationality" (1987: 96). What canons, then, <u>do</u> guide intentional-state attribution? Dennett replies: "When considering what we <u>ought to do</u>, our reflections lead us eventually to a consideration of what we <u>in fact do</u>; this is inescapable, for a catalogue of our considered intuitive judgments on what we ought to do is both

a compendium of what we do think, and a shining example ... of how we ought to think" (1987: 98).

As Dennett himself immediately recognizes, this response makes it look as if his theory collapses into something similar to ST.[ix] He briefly tries to distance himself from this threatened collapse (1987: 99-100), but the attempt is not very convincing.[x] So the question is: Can RT, without collapsing into ST, identify a set of normative rules of rationality that attributors actually employ? Thus far, the prospects look bleak. I turn now to TT. The first problem I wish to raise for TT is the difficulties it faces in trying to give a plausible account of first-person (current) mental attribution. According to TT, what are the "data" from which a person might infer one of her current mental states? Is it her own behavior, some current external stimulus, or a combination of the two? This seems hopeless as a full account of how people make self-attributions. I currently believe that I have the intention to cook shrimp for dinner tonight. But there is no external stimulus from which this intention state is inferable (I just thought of cooking shrimp without seeing any pictures of shrimp, or the like), and I have not yet taken any behavioral steps from which this intention could be inferred. So how can TT accommodate my epistemic access to this intention state?

The problem is daunting when one considers what (strict) TT requires. As formulated in a suitably sophisticated functionalism, the concepts of mental-state types are concepts of functional roles: dispositions to be caused by specified external events, to causally interact in specified ways with other internal events (which themselves have certain functional roles), and to cause specified behaviors. Strictly speaking, what has functional roles are state-types of a person

[ix] Specifically, Dennett worries whether his own view collapses into Stephen Stich's view, a "projectivist" view very similar to ST that Stich endorsed in the early 1980s (including Stich 1983) but which he has subsequently criticized in many papers.

[x] Dennett goes on (1987: 100-101) to consider an even clearer formulation of ST (using the very term "simulation"), and argues that it collapses into TT. His argument rests on the claim that simulation could only work by relying on a theory. But this is incorrect, as I argue elsewhere (Goldman 1989). While simulation could, of course, be "theory driven", it also could be "process driven", i.e., driven by psychological processes rather than a theory. The interesting and proper interpretation of ST is that it postulates intensive use of process-driven simulation.

or organism (see Schiffer 1987: 20-22). So if a person wishes to answer the question, "Am I now in mental-state M?", she needs to determine whether there is any state token she is currently in which instantiates a type that realizes the functional role associated with "M". To determine this, she will have to determine the dispositions associated with her current state-tokens, dispositions that are highly complex matters.

The dispositions are complex for three reasons, each of which poses epistemic problems (see Goldman 1993). First, the dispositions involve relationships with other events, to which the person may not have current access. For example, they may involve propensities to cause later mental states and/or behavior, but the person will have no current epistemic access to later events. (Of course, later events might be inferred; but such inference is only possible after the current state is type-identified.) Second, functional roles involve subjunctive relationships. They specify interactions that would occur if certain other conditions obtained. For example, having desire D might involve a disposition to form plan P if one has belief B. But perhaps one does not have belief B. How is one to determine whether or not a present state-token would produce P if B, contrary to fact, were present? Third, there is a problem of combinatorial explosion for inferences addressed to propositional attitudes. Each attitude type is associated with a functional role that implicates an indefinitely large number of other attitudes types, each of which is in turn associated with a comparably complex functional role. The upshot for epistemic purposes is an inferential holism posing a severe threat of computational intractability (Goldman 1993).

To avert these problems, a more plausible theory would say that self-attribution occurs by internally detecting some properties of mental-state tokens that are (A) categorical (i.e., non-dispositional) and (B) non-relational (at least not massively relational in the way that functional-role properties are). There are two types of candidates for such properties: phenomenological and non-phenomenological properties. The language of thought, for example, presumably has non-phenomenological characteristics that meet conditions (A) and (B), and these characteristics

might be the ones directly detected in identifying the content components of mental states.[xi] In the past (Goldman 1993) I flirted with the idea that the pertinent properties are phenomenological--at least for the non-content components of mental states.[xii] In the present paper, I mean to remain neutral on this issue. I mean to endorse only the direct detection of some such properties, whether or not they are phenomenological.

It is noteworthy that some long-time proponents of TT, viz., Nichols and Stich (forthcoming), now endorse a similar model, thereby abandoning the core TT answer to question (Q2)(B). They call their theory the "Monitoring Mechanism Theory", and contrast it with TT. However, they don't seem to appreciate the full implications of this approach. They adopt the standard boxological story of the attitudes, according to which beliefs are depicted as states "residing" in certain boxes. The standard lore on boxes (which they follow) is that box talk is merely short-hand for talk about functional roles. If this is correct, the properties that qualify a state as a belief, a desire, or an intention are not categorical or non-relational. So the Nichols-Stich story does not escape the computational problems mentioned above.

What about TT's approach to third-person attribution? I'll be quite brief here, and I'll center my comments around the "frame problem" that Jane Heal (1996) has helpfully introduced in this context.[xiii] Begin with the assumption that adults are pretty proficient at predicting the reasoning of other adults on the persistence or nonpersistence of various states of affairs over time. If a familiar type of change takes place in the world, adults will draw conclusions about what will change and what will remain the same in the next time period. Furthermore, if these are prosaic real-world scenarios, Sandy will be reasonably competent at predicting the

---

[xi] The question of content attribution is greatly complicated by issues concerning externalism about content. There is no room in this paper to take up those issues, which in any case are somewhat tangential to the problems addressed here.

[xii] My discussion may have encouraged an interpretation under which attitude contents were also supposed to be phenomenologically detectable; but that was not my intention.

[xiii] However, Heal's discussion formulates the frame problem in very general terms, whereas I shall try to stick to its original, somewhat narrower, set of concerns.

conclusions Mandy will draw about such scenarios. Is it plausible that Sandy's mindreading of Mandy is guided by a theory of such reasoning that Sandy knows?

If Sandy really has such knowledge, she is way ahead of, or at least in the same league with, the most astute practitioners of Artificial Intelligence, who have wrestled with this knotty problem without reaching much consensus (Lormand 1999). Workers in AI have tried to handle nonchanges by letting a system assume by default that a state persists unless there is an axiom specifying that it is changed by an occurrence, given surrounding conditions. But new evidence often requires such an assumption to be retracted. When should such retraction take place?[xiv] Ordinary people seem to be able to determine which changes are relevant to which other changes. But how, in general, is such relevance determined? Some theories of reasoning about change implausibly posit the use of enormous numbers of axioms about nonchanges; others avoid this cost but posit implausibly bold axioms about nonchanges.

TT implies that ordinary folk have developed (presumably tacitly) an excellent theory about other people's reasoning about change. That is how TT would explain their substantial success in predicting other people's inferences about changes and nonchanges. It's not just that one or two ordinary folk have made this theoretical breakthrough; most ordinary people appear to have done so. However, this seems to beggar belief, given the limited success that apparently brilliant minds have enjoyed (at the explicit level) when addressing the very same problem. To be sure, young children develop theories of a language's grammar in a way that surpasses what linguists have yet accomplished. But there it is plausible to suppose that a dedicated, special-purpose module is at work. Is it comparably plausible that there is a special-purpose module for developing theories specifically about the causal reasoning powers of other people?

An entirely different account of mindreading in this domain seems far more promising,

[xiv] In the "Yale Shooting Problem," for example, Hanks and McDermott (1986) explore the following case. Let a system assume by default that (1) live creatures remain alive and (2) loaded guns remain loaded. Confront it with this information: Fred is alive, then a gun is loaded, then, after a delay, the gun is fired at Fred. If assumption (2) is in force through the delay, Fred probably vioates (1). But equally, if assumption (1) is in force after the shooting, the gun probably violates (2). Why is (2) the more natural assumption to enforce?

viz., that when Sandy seeks to predict Mandy's inferences about changes and nonchanges, she (Sandy) goes through the target inferences on her own and expects Mandy to reach the same conclusions as she does. This, of course, is precisely what ST postulates. Sandy does these reasoning tasks as well as Mandy, and presumably in the very same fashion. If she uses the simulation heuristic to predict Mandy's results, she doesn't need a theory about how the task is executed.

The same point can be made with a related example: counterfactual reasoning. Philosophers first identified the problem of counterfactual reasoning only in the 1940s or so, and only by 1970 (roughly) were moderately adequate theories developed (e.g., Lewis 1973). Presumably, however, ordinary people living in earlier periods were reasonably competent at predicting and explaining other people's feats of counterfactual reasoning. Doesn't it stretch credulity to suppose that they had a folk theory of such reasoning at times when philosophers had little if any inkling that there was even a problem in this territory?

IV.

In this final section I turn to my own favored approach, ST, and to some slight specifics of its configuration, especially as pertains to question (Q1). There is broad agreement among simulationists as to how ST answers (Q2)(A), the question of third-person attribution. There is far less agreement about how it should answer (Q1) or (Q2)(B). It seems natural, moreover, for positions on the latter two questions to be intertwined. For example, if one accepts the direct detection account of first-person (current) attribution, it is not unnatural to suppose that mental-state concepts involve some sort of categorical features or characteristics that are epistemically identifiable via direct, internal detection. This is the kind of view I have advocated elsewhere (Goldman 1993, forthcoming). This view, with its strong Cartesian or semi-Cartesian flavor, contrasts with another major tradition in philosophy of mind, rooted in the writings of Wittgenstein, which places overt behavior at the center of all comprehension of the mind.

Another simulationist, Robert Gordon, takes a very different tack from mine, clearly along behaviorist lines. Gordon (1996, unpublished) contends that self-attribution relies, not on

introspection or inner detection, but on what he calls ascent routines. The way adults ordinarily determine whether or not they believe that p, he says, is simply to ask themselves the question whether or not p. If someone were to ask me whether I believe Mickey Mouse has a tail, I would ask myself "Does Mickey Mouse have a tail?". If my answer to this latter question is "yes", then my presumptive answer to the belief-question will also be "yes". Gordon calls this an "ascent routine" because it answers a question at the mental level by answering another question pitched at what he considers a "lower" level (Gordon 1996: 15). He claims that this enables one to answer a seemingly "internal" question by means of anwering an "outward-looking" question (unpublished: 3). To find out about one's own mental states, he thinks, one does not consult "internal" information but "external" information.

> For example, when I tell the clerk, "I want two of these and three of those, "it is very likely that I do want possession of the items I point to. To attain reliability with regard to what I want, I don't look inward, I look outward at the display case. (Unpublished: 3.)

I find several problems with Gordon's ascent-routine approach. First, how does he hope to generalize from cases of belief to other propositional attitudes? In the case of belief, one substitutes for the question, "Do I believe p?" the question "Is p true?". But what is the story for the other propositional attitudes? If I ask myself the question, "Do I want two croissants from the display case?", what ("outward-looking") question am I to substitute for that? Perhaps Gordon will reply. "Are the two croissants in the display case appetizing-looking?" But there are several problems here. First, how am I to tell whether they are appetizing looking--to me--without (tacitly) consulting an internal state of mine? Second, consider two different questions I might ask about my attitude toward the croissants: "Do I want two of the crossants in the display case?" and "Do I believe there are two croissants in the display case?" In both cases, I might be assisted in answering the question by looking at the display case, but the difference between what I need to know in the two cases is not available from the outward object, only from my mental condition. Third, consulting the display case only seems relevant to making up one's mind what to believe or desire, not to determining an attitude already possessed. Fourth, what

about other mental attitudes? Which outward-looking questions should I substitute for the following questions, "Do I now intend to bring it about that p?", "Am I now wondering whether p?", "Am I now recalling p?", and "Am I now imagining that p?"?

A different kind of problem with Gordon's approach is that there should be a tight connection or resemblance between information relevant to questions about past mental states and information relevant to present mental states. If Gordon is right that questions about present first-person mental states can be answered by information about outward objects, then questions about past first-person mental states should also be answered by information about outward objects. But that is surely wrong. If I ask myself, "Did I intend to bring about p (at such-and-such a time)?" or "Was I then imagining that p?", I need not retrieve anything about outward objects to answer these questions. I can just recall my (internal) mental state at the time.

This discussion confirms my earlier proposal, viz., that the story of first-person present mental-state attribution should assign a salient role to the direct inner detection of some sorts of internal features or characteristics. If we now try to apply this result to question (Q1) on our list, it looks as if the concepts of mental states should somehow be composed of the relevant sorts of internally detectable properties that would account for first-person attribution. This proposal immediately runs into trouble on both empirical and philosophical grounds. There is empirical evidence, which fits with suggestions of Wittgenstein, that infants (under one year) have a sensitivity to certain patterns of movement associated with intentionality, agency, or goal-directedness. According to Premack (1990) and Gergeley et al. (1995), the core notion of intentionality is triggered in young children by perceptual patterns suggestive of self-propelled motion or motion with respect to a potential goal. Baron-Cohen (1995) seconds Premack's idea by positing an "intentionality detector", alleged to be an innate endowment for reading mental states off of behavior (1995: 32). As described by Baron-Cohen, the intentionality detector interprets almost anything that propels itself or makes a non-random sound as an agent with goals and desires (1995: 34).[xv] But if goal possession is stimulated by external motion or

[xv] Some aspects of these ideas are challenged by a study of Woodward (forthcoming).

behavior of certain types, then the concepts of purposiveness or goal-possession do not, after all, seem to consist in inwardly detectable features.

I shall approach this problem, in an admittedly speculative fashion, by advancing a dual-representation hypothesis about MS concepts. The focus will be on goals, desires, or plans, though similar proposals might apply to other types of MSs. The rough idea is that people develop two sorts of mental representations for at least some MSs. Representations associated with behavioral characteristics are an initial entry-way into a conception of desire and intentionality; but a full conception of desire and intentionality also involves representations associated with internal characteristics (of a phenomenological or sub-phenomenological variety). In tendering this hypothesis, I do not mean to propose that people have one type of representation for third-person desires and a different type of representation for first-person desires. That would yield the traditional puzzle in philosophy of mind of how people can conceptualize both themselves and others as having the same sorts of states. Rather, people coordinate representations of certain behavior and representations of certain inner features as representations of one and the same (sort of) state. My story is intended to allow for changes and development in the grasp of mentality at different (early) ages, but the changes I have in mind do not coincide with those standardly discussed in the developmental literature.

The idea of dual, or multiple, representations of a single type of object, state, or category is quite common in cognitive science. For example, people might represent a single sortal in terms of both shape and function; and they may deploy representations utilizing different cognitive codes or modalities. Thus, Biederman (1987) proposes that visual object-classification commonly proceeds by means of shape-coded object types. In visually identifying something as

---

Woodward found that 9-month-olds differentiated between movements of grasping a toy and movements of merely letting the back of one's hand drop onto a toy. In other words, by nine months of age infants selectively encode certain goal-related properties and not others; and this selective encoding appears to begin at roughly five months. So infants don't seem to attribute goal-directedness indiscriminately to any motion of a self-propelled entity. However, these sorts of qualifications still leave in place the fundamentally behavioral character of the cues used for third-person goal-state detection.

a piano or a lamp, one exploits a stored model or prototype of what pianos or lamps look like. In addition, there are separate, modality-neutral codes representing pianos and lamps in terms of their functions, e.g., "keyboard musical instrument" and "portable source of (artificial) light".

Some pairs of matching representations are especially striking. Meltzoff and Moore (1983) discovered that infants as young as forty-two minutes can imitate the facial gestures of another person. These newborns apparently represent their own facial movements proprioceptively--after all, they have not previously seen their own faces. They produce facial movements of their own that mimic those of a target, so they must somehow match representations of their own movements to representations of the target's facial gestures (Meltzoff 1999).[xvi] But since they don't see their own movements, it must be proprioceptive representations of their own movements that they link to visual representations of the target's movements. Meltzoff and Moore therefore postulate an innate cross-modal matching between visual and proprioceptive representations of the same movements.[xvii]

In a somewhat analogous fashion, I conjecture that children learn to match inner features detectable in their own goal or desire experiences with behavioral cues utilized in identifying goals or desires. In principle, this could be achieved by noting their own goal-driven behavior and its regular association with these inner features. But this would leave us with the traditional question of why they should think that other people ever undergo analogous inner experiences. This would be a particularly vexing problem for ST, because ST postulates that third-person attributors make an inference from the output states of their own mental simulations to corresponding inner states of their targets. Why should they think that other people undergo such internal states at all? Philosophers of mind have shied away from internal features in their theorizing precisely because of this looming problem of other minds.

[xvi] In this case, of course, the infants' "matching" representations are not representations of numerically the same action, only similar actions (their own and that of the observed target).

[xvii] Elisabeth Pacherie (1998) argues that the cross-modal matching must be between motor and visual representations rather than proprioceptive and visual representations. This fits the hypothesis I shall advance below even better.

To develop my conjecture further, I now turn to the recently discovered phenomenon of so-called "mirror neurons".[xviii] An interesting class of premotor neurons were initially discovered in macaque monkeys, neurons that discharge both when the monkey performs an action and when it observes another individual making a similar action (Rizzolatti et al. 1988). The discharge of these neurons is not correlated with individual movements, but rather with general types of actions, especially grasping, holding, tearing, poking, and manipulating. So these neurons seem to code motor representations of goal-directed actions involving objects. Mirror-neuron activity, then, is apparently a neural correlate of plans or goals for action. Mirror neurons are also triggered, however, when an observer watches a target agent act toward a goal. This correspondence, or "mirroring", between observed and executed actions is sometimes described in terms of the metaphor of physical "resonance" (Rizzolatti et al., in press). It is as if neurons in certain motor areas start to "resonate" with their cousins as soon as the appropriate visual input is presented. This resonance phenomenon is not restricted to monkeys, but is found in humans as well, as confirmed by transcranial magnetic stimulation (TMS), MEG/EEG recordings, and brain imaging techniques. Experiments demonstrate that motor centers of adult humans resonate during movement observation. For example, Fadiga et al. (1995) stimulated the left motor cortex of normal subjects using TMS while they were observing arm movements and hand grasping movements performed by the experimenter. Motor evoked potentials were recorded from various arm and hand muscles. A selective increase of motor evoked potentials was found in those muscles that the subjects normally use for producing the observed movements. One can make casual observations of the same underlying mechanisms in everyday life. When you see a cyclist about to have an accident, your own muscles tense in what feels like the way they would tense if you were on the bicycle about to have an accident.[xix] Findings in clinical populations

[xviii] The reader is reminded that I am not addressing the traditional problems of normative epistemology in the mental domain. I am not asking what confers justification on people's beliefs about other minds. I seek only a descriptive account of the generation of their beliefs, which is what the study of folk psychology, as here construed, aspires to provide.

[xix] Thanks to Bill Child for this example.

further support these ideas. Lhermitte et al. (1986) document a phenomenon of "imitation behavior", in which patients with prefrontal lesions compulsively imitate gestures or even complex actions performed in front of them by an experimenter. This behavior is explained as arising from an impairment of inhibitory control that normally governs motor plans. Apparently, when observing someone else perform an action, a normal human generates a "resonant" plan or image of doing that action himself. This plan is normally inhibited so that it does not yield motor output; but such inhibition is impaired in the relevant patient population.

Vittorio Gallese and I (Gallese and Goldman 1998) have cited resonance phenomena as possible evidence for, or a precursor of, mental simulation. Here I mean to invoke them for the twin purposes sketched above: (1) to explain how there could be a "matching" between internal and external representations of (what is conceived of as) the same state, and (2) to explain how a child might come to interpret others as undergoing internal experiences of desiring or planning similar to her own.

It is easy to see how mirror-neuron (MN) activity could facilitate the establishment of a correspondence between internal and behavioral formats for representing mental states. MN activity involves inner events associated with goal possession, desire, or planning. Externally triggered MN activity involves the observation of behavior and environmental circumstances (i.e., the presence of a goal object) associated with goal possession. So when one undergoes externally triggered MN activity, there is a (roughly) simultaneous occurrence of observed behavior and detectable inner events that might come to be linked or associated with one another.

To be sure, the observed behavior is the behavior of another creature. How might a learner come to link certain inner experience in himself with the behavior of another? The learner is in a position to notice that the same sorts of inner experience occurs when he himself prepares to execute an action. This is because MN activity serves the primary function (along with other neuron groups in the premotor cortex) of preparing to execute an action. When MN activity is externally driven, the action is not actually executed; it is inhibited. But the appropriate sort of action <u>is</u> executed by the person being observed. So the learner might get the

idea--perhaps is even "hard-wired" to get the idea--that inner events of the type he undergoes while watching the other's behavior are also undergone by the observed actor. This would be no more remarkable than the matching of visual and proprioceptive (or motor) representations of behavior by Meltzoff and Moore's neonates.

Although the dual-representation hypothesis I am floating may be appealing, some might argue that it constitutes a major concession, or even capitulation, to TT. If behavioral features are admitted into the set of representations associated with desire and goal-possession, doesn't this amount to belated agreement with TT on the proper answer to question (Q1)? This isn't so, for two reasons. First, only a limited number of mental-state types, I suspect, have relatively simple and tight connections with stereotypical behavior. The story being told here about the concepts of desire, intention, or goal-possession probably does not generalize to all mental-state concepts (e.g., belief). Second, nothing in the story sketched above suggests that children acquire any far-flung theory of systematic mental-state connectedness of the sort postulated, for example, by functionalism. And if they do develop a grasp of mental-state/mental-state connections, this may only be with the help of internally detected features, which does not fit the standard mold of the TT approach. So although the dual-representation answer to (Q1) takes notice of some of the motivations behind traditional versions of TT, it keeps a considerable distance from all of the detailed versions of TT that have actually been proposed.[xx]

[xx] I am grateful to Elisabeth Pacherie and Brad Thompson for very helpful comments on this paper.

REFERENCES

Baron-Cohen, S. (1995). Mindblindness. Cambridge, MA: MIT Press.

Bartsch, K. and Wellman, H. (1995). Children Talk about the Mind. New York: Oxford University Press.

Biederman, I. (1987). Recognition by components: a theory of human image understanding. Psychological Review 94: 115-147.

Brandom, R. (1994). Making It Explicit. Cambridge, MA: Harvard University Press.

Cheng, P. and Holyoak, K. (1985). Pragmatic reasoning schemas. Cognitive Psychology 17: 391-416.

Churchland, P. M. (1988). Mind and Consciousness. Cambridge, MA: MIT Press.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. Cognition 31: 187-276.

Davidson, D. (1984). Inquiries into Truth and Interpretation. Oxford: Oxford University Press.

Dennett, D. (1987). The Intentional Stance. Cambridge, MA: MIT Press.

Fadiga, L., Fogassi, L., Pavesi, G., and Rizzolatti, G. (1995). Motor facilitation during action observation: A magnetic stimulation study. Journal of Neurophysiology 73: 2608-2611.

Fodor, J. (1985). Fodor's guide to mental representation: The intelligent auntie's vade-mecum. Mind 94: 55-97.

Fodor, J. (1998). Concepts: Where Cognitive Science Went Wrong. Oxford: Oxford University Press.

Fodor, J. and LePore, E. (1992). Holism: A Shopper's Guide. Cambridge, MA: Blackwell.

Fuller, G. (1995). Simulation and psychological concepts. In M. Davies and T. Stone, eds., Mental Simulation. Oxford: Blackwell.

Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. Trends in Cognitive Sciences 2: 493-501.

Gergeley, G. et al. (1995). Taking the intentional stance at 12 months of age. Cognition 56: 165-193.

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

Goldman, A. (1989). Interpretation psychologized. Mind and Language 4: 161-185.

Goldman, A. (1993). The psychology of folk psychology. Behavioral and Brain Sciences 16: 15-28.

Goldman, A. (forthcoming). The mentalizing folk. In D. Sperber, ed., Metarepresentation. New York: Oxford University Press.

Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge. Behavioral and Brain Sciences 16: 1-14.

Gopnik, A. and Meltzoff, A. (1997). Words, thoughts and theories. Cambridge, MA: MIT Press.

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

Gordon, R. (1986). Folk psychology as simulation. Mind and Language 1: 158-171.

Gordon, R. (1995). Simulation without introspection or inference from me to you. In M. Davies and T. Stone, eds., Mental Simulation. Oxford: Blackwell.

Gordon, R. (1996). 'Radical' simulationism. In P. Carruthers and P. Smith, eds., Theories of Theories of Mind. Cambridge: Cambridge University Press.

Gordon, R. (unpublished). Self-ascription: Jonesian conditioning vs. ascent routines.

Hanks, S. and McDermott, D. (1986). Default reasoning, non-monotonic logic, and the

frame problem. Proceedings of the American Association for Artificial Intelligence 328-333.

Heal, J. (1986). Replication and functionalism. In J. Butterfield, ed., Language, Mind, and Logic. Cambridge: Cambridge University Press.

Heal, J. (1996). Simulation, theory, and content. In P. Carruthers and P. Smith, eds., Theories of Theories of Mind. Cambridge: Cambridge University Press.

Heal, J. (1998). Co-cognition and off-line simulation. Mind and Language 13: 477-498.

Leslie, A. (1991). The theory of mind impairment in autism: Evidence for a modular

mechanism of development? In A. Whiten, ed., Natural Theories of Mind. Oxford: Blackwell.

Leslie, A. and German, T. (1995). Knowledge and ability in "theory of mind": One-eyed overview of a debate. In M. Davies and T. Stone, eds., Mental Simulation. Oxford: Blackwell.

Lewis, D. (1972). Psychophysical and theoretical identifications. Australasian Journal of Philosophy 50: 249-258.

Lewis, D. (1973). Counterfactuals. Oxford: Blackwell.

Lewis, D. (1974). Radical interpretation. Synthese 23: 331-344.

Lhermitte, F., Pilon, B. and Serdaru, M. (1986). Human autonomy and the frontal lobes: I. Imitation and utilization behavior: a neuropsychological study of 75 patients. Annals of Neurology 19: 326-334.

Lormand, E. (1999). Frame problem. In R. Wilson and F. Keil, eds., MIT Encyclopedia of the Cognitive Sciences 326-327.

Meltzoff, A. (1999). Imitation. In R. Wilson and F. Keil, eds., MIT Encyclopedia of the Cognitive Sciences. Cambridge, MA: MIT Press.

Meltzoff, A. and Moore, M. (1983). Newborn infants imitate adult facial gestures. Child Development 54: 702-709.

Nichols, S. and Stich, S. (forthcoming). Reading one's own mind: A cognitive theory of self-awareness.

Pacherie, E. (1998). Representations motrices, imitation et theorie de l'esprit. In H. Grivois and J. Proust, eds., Subjectivite et Conscience d'Agir: Approches Cognitive et Clinique de la Psychose. Paris: Presses Universitaires de France.

Peacocke, C. (1992). A Study of Concepts. Cambridge, MA: MIT Press.

Premack, D. (1990). The infant's theory of self-propelled objects. Cognition 36: 1-16.

Quine, W. V. (1960). Word and Object. Cambridge, MA: MIT Press.

Rizzolatti, G. et al. (1988). Functional organization of inferior area 6 in the macaque monkey: II. Area F5 and the control of distal movements. Experimental Brain Research 71: 491-507.

Rizzolatti, G., Fadiga, L. Fogassi, L., and Gallese, V. (in press). From mirror neurons to imitation: Facts and speculations. In W. Prinz and A. Meltzoff, eds., The Imitative Mind: Development, Evolution and Brain Bases. Cambridge: Cambridge University Press.

Schiffer, S. (1987). Remnants of Meaning. Cambridge, MA: MIT Press.

Sellars, W. (1963). Empiricism and the philosophy of mind. In Science, Perception, and Reality. New York: Humanities Press.

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

Stich, S. and Nichols, S. (1992). Folk psychology: Simulation or tacit theory? Mind and Language 7: 35-71.

Tversky, A. and Kahneman, D. (1986). Rational choice and the framing of decisions. Journal of Business 59: 251-278.

Woodward, A. (forthcoming). Selectivity and discrimination in infants' encoding of human behavior.

NOTES