Philosophical Perspectives, 8, Logic and Language, 1994

THE TRANSPARENCY OF MENTAL CONTENT[1]

Paul A. Boghossian
New York University

Introduction

Michael Dummett once wrote:

> It is an undeniable feature of the notion of meaning—obscure as that notion is—that meaning is *transparent* in the sense that, if someone attaches a meaning to each of two words, he must know whether these meanings are the same.[2]

As the surrounding discussion makes clear, Dummett's claim isn't merely the fairly weak thesis that, for any two expressions in a person's repertoire, it must be possible for that person to come to know whether or not they mean the same. Rather, the claim is that it must be possible for that person to come to know such a fact purely *introspectively*, without the benefit of further empirical investigation--*a priori*, as I shall also occasionally put it.

Dummett doesn't say much about the basis for his conviction that 'transparency' in this sense—*epistemic transparency* as I shall sometimes call it, to distinguish it from the *referential* variety—is "an undeniable feature of the notion of meaning." He gives the impression of finding the claim too obvious to need arguing. In this respect, I believe, he follows in the footsteps of Frege and Russell, both of whom gave epistemic transparency a pivotal, if unargued, role in their respective theories of linguistic and mental content.

In another paper, I hope to discuss this historical point. In this paper, however, my concern will be purely philosophical and, then, only with *mental* content. I believe that the notion of epistemic transparency does play an important role in our ordinary conception of mental content and I want to say what that role is. Unfortunately, the task is a large one; here I am able only to begin on its outline.

I shall proceed somewhat indirectly, beginning with a discussion of *externalist* conceptions of mental content. I shall show that such conceptions violate epistemic transparency to an extent that has not been fully appreciated. Subsequently, I shall look at the implications of this violation and at the reconstructive project that a rejection of transparency entails. I am inclined to think it unlikely that we will get a conception of propositional content that underwrites epistemic transparency. But I am concerned that we have not fully appreciated the role that transparency currently plays and the work that would need to be done were we to discard it.

Externalism, Self-Knowledge and Transparency

The minimal intended significance of the Twin Earth thought experiments may be put like this: Either mental contents are not individuated individualistically, or they are not individuated in terms of their truth conditions.

Thus, consider Oscar and twin Oscar; they are alike in all intrinsic, non-relational respects. But whereas Oscar grows up in a physical environment which contains H_2O, twin Oscar grows up in an environment which contains XYZ. And widespread intuition appears to have it that whereas Oscar's utterance of 'Water is wet' expresses a thought that is true if and only if H_2O is wet, twin Oscar's utterance of that sentence expresses a thought that is true if and only if XYZ is wet. Hence, either the contents of the two thoughts are not individuated individualistically or they are not individuated in terms of their truth conditions.[3]

Thus, consider also Bert and twin Bert; they are alike in all intrinsic, non-relational respects. But whereas Bert grows up in, and defers to, a community in which the word 'arthritis' is used to apply only to a certain kind of inflammation of the joints, twin Bert grows up in, and defers to, a community in which 'arthritis' is used to apply to all rheumatoid ailments, even to those that might occur in the thigh. And widespread intuition has it that whereas Bert's utterance of 'I have arthritis in my thigh' expresses a thought that is true if and only if he has arthritis in his thigh (and, hence, can never be true), twin Bert's utterance of that sentence expresses a thought that is true if and only if he has (some other disease) tharthritis in his thigh (and, hence, may occasionally be true). As before, the moral would appear to be: Either the mental contents of the two thoughts are not individuated individualistically or they are not individuated in terms of their truth conditions.[4] Let us call the former view *externalism*, and the latter *individualism.*

According to externalism, then, the psychologically relevant contents of a subject's intentional states are individuated in part by certain sorts of facts about the physical and/or social environment in which he happens to be situated.[5] And although this would appear to undermine the claim that we can have purely introspective, non-empirical access to the contents of our thoughts, we are assured by an increasing number of philosophers that this appearance is illusory:

there is in fact no conflict between an externalist conception of mental content and introspective self-knowledge. I want to begin with the question, On what is this confidence based?

It is based on a thought that is succinctly expressed by Davidson:

> Showing that there is no conflict [between externalism and knowledge of content] is basically simple. It depends on realizing that whatever is responsible for the contents of our thoughts, whether known or not, is also responsible for the content of the thought that we have the thought.[6]

How exactly should we understand Davidson's thought here?

Let's assume, without loss of generality, that we think in a language of thought (in fact, I'll pretend that we think in English. This will make the exposition easier but is not required for anything that's to follow: similar claims could have been made without the assumption that psychological states possess syntactic structure.) With the convenient assumption in place, we may say that Peter believes that **p** just in case Peter has a token of a sentence S which means that **p**, in his belief box; and Peter desires that **q** just in case Peter has a token of a sentence Q, which means that **q**, in his desire box; and so on for the other attitudes.

It is now easy to give a completely general elaboration of Davidson's thought. For since externalism is guaranteed to be entirely neutral on the matter, we may safely assume that there are *syntactic* mechanisms which ensure that when, and only when, a token of the sentence S enters Peter's belief box, then *ceteris paribus*, so does a token of the sentence 'I believe S'. And so, provided that the embedded and unembedded tokens of S may always be counted upon to express the same content, Peter's second-order belief will always report correctly on the content of his first-order belief, the externalist nature of content individuation notwithstanding. Hence, and in this sense, externalism is shown to be consistent with authoritative first-person knowledge of thought content. This is, I think, the best and simplest way to develop the idea that Davidson expresses.[7]

Now, there are, I think, a number of questions that might be raised about this proposal. For instance: can we in fact always count on the embedded and unembedded occurrences of S to express the same content? (As we shall presently see, there may be special reasons for doubting the reliability of this covariation under externalist assumptions.) Also: the envisaged proposal would appear at best to ensure the availability of *reliable* beliefs about content, not *knowledge*; and there are some tricky questions about justification that need answering. But I am not going to pursue either of these lines of questioning here; for present purposes, I wish simply to grant the claim that externalism is consistent with the sort of knowledge of content envisaged by this elaboration of Davidson's idea. My point will rather be that the kind of self-knowledge that is thereby secured falls short—way short—of the kind of self-knowledge we normally think of ourselves as possessing.

Let us grant the claim that, for any sentence S in my belief box, I can form a correct second-order belief about its content by prefixing the words 'I believe...' to that very sentence. It is important to observe, however, that no other epistemic capacity is thereby grounded. For instance, suppose I have two sentences P and Q in my belief box that are *identical* in content; it needn't follow that I will know, or be able to *a priori* come to know, that they are, and this despite my being said to know what the contents of P and Q are individually. Nor, similarly, does it follow that if I have two sentences Q and R in my belief box that are *distinct* in content, that I will know, or be able to *a priori* come to know, that they are; and this despite my being said to know what the contents of Q and R are individually. In fact, it doesn't even follow from the sort of knowledge that is underwritten by the proposal, that I will know, or be able to *a priori* come to know, whether two distinct tokens of the *same* sentence type S express the same or distinct contents. To put the point concisely, the proposal does nothing to reconstruct the intuitive idea that we know the contents of our thoughts in the sense that we can introspectively *discriminate* between them, that we can tell them apart from each other: it doesn't by itself underwrite the epistemic transparency of mental content.

Externalism, Transparency and the Semantics of Travel

The thesis of the epistemic transparency of content may be usefully broken up into two parts: (a) If two of a thinker's token thoughts possess the same content, then the thinker must be able to know *a priori* that they do; and (b) If two of a thinker's token thoughts possess distinct contents, then the thinker must be able to know *a priori* that they do. Call the first the thesis of the *transparency of sameness* and the second the thesis of the *transparency of difference.*

Now, it is fairly easy to show that externalist contents do not satisfy the transparency of sameness. Kripke's notorious Frenchman, Pierre, already shows this for the special case of *Millian* contents (themselves, of course, a species of externalist content).[8] The details are well-enough known that I need not rehearse them here. A plausible story culminates in Pierre having two expressions—'Londres' and 'London'—which refer to the same city without, however, his realizing that they do. The thought he expresses with

> London is pretty

is the same as the thought he expresses with

> Londres is pretty;

but this fact is unavailable to introspection alone.

We need a different sort of case, however, to illustrate the general claim that even a non-Millian, referentially opaque content would fail the transparency of sameness, if it were individuated externalistically. Brian Loar has described a case which may be adapted to just this purpose.[9]

Paul, an English speaker, has been raised by a French nanny in a sheltered way. She speaks English with Paul, but amuses herself by referring to the cats around them as "chats" and never as "cats." Paul, satisfying the usual competence and deference constraints on concept possession, is to be credited with beliefs involving the concept **cat**. In particular, the belief he would express with "All chats have tails" would be the belief that all cats have tails.

Now, Paul occasionally sees his parents, who speak of animals called "cats." Because on those occasions, no cats or pictures of cats are ever present, Paul does not realize that cats are his familiar "chats." However, Paul's parents tell him quite a lot about cats and in particular that all cats have tails. On this basis, it is again true of Paul that the belief he would express with "All cats have tails" would be the belief that all cats have tails.[10]

Intuitively, however, it is quite clear that Paul will not be able to tell *a priori* that the belief he expresses with "All chats have tails" is the same belief that he expresses with "All cats have tails."

So much, then, by way of illustrating externalism's violation of the transparency of sameness. What has not been generally appreciated, however, is that externalism also allows for a coherent description of the *converse* case: two tokens of a thinker's language of thought belonging to the *same* syntactic type have *distinct* meanings, but the thinker is in principle not in a position to know *a priori* that they do. These converse cases, illustrating externalism's violation of the transparency of difference, are, I believe, important even for issues beyond the ones presently at hand.[11]

To see *one* way in which such converse cases might arise we need to think—harder perhaps than is common in the literature—about the semantics of so-called 'one-world traveling cases'. Suppose that Earth and Twin Earth are part of the actual world and that Peter, a normal, competent adult Earthling, is suddenly and unwittingly transported to Twin Earth. An accidental tourist, Peter goes to sleep one night at home and wakes up in twin home in twin bed. He suffers no discernible disruption in the continuity of his mental life. Here on Twin Earth, Peter happily lives out the rest of his days, never discovering the relocation that he has been forced to undergo. How should we think about the semantics of Peter's thoughts?

Well, one intuition that is shared by practically everyone who has thought about these cases is that, after a while (how long is unclear), some tokens of, say, 'water' in Peter's language of thought will shift from meaning **water** and will come to mean **twater**. Thus, to quote just one example, Tyler Burge writes:

> The thoughts would not switch as one is [quickly] switched from one actual situation to another twin actual situation. The thoughts would switch only if

> one remained long enough in the other situation to establish environmental relations necessary for new thoughts. So quick switching would not be a case in which thoughts switched...
>
> But slow switching could be such a case.[12]

Viewed from an externalist perspective, it's hard to deny the force of the intuition that Burge is expressing here. We imagine Peter moving to Twin Earth and staying there for a while. Over time he thinks to himself a variety of thoughts that he expresses with sentences like:

> I won't go swimming today; the water is too cold.

Or:

> There's too much water in this whiskey.

It seems very plausible that these expressions of beliefs about one's present environment, expressions of current desires and current intentions, are expressions of thoughts that involve the concepts **twater**, not **water**.

So it seems right to say that some tokens of 'water' will shift from meaning **water** to meaning **twater**. But it seems to me equally compelling to say that certain *other* tokens of 'water' *won't* shift. Specifically, it seems to me that those tokens of 'water' occurring in *memories*, and in *beliefs about the past* based upon them, will retain their Earthly interpretations, despite being tokened on Twin Earth. Such thoughts, unlike, for instance, beliefs with undated general contents, or thoughts about one's present surroundings, are caused and sustained by *previous* perceptions long gone. In the normal case, they owe little, if anything, to current perceptions and cognitive transactions with one's environment. From a purely intuitive standpoint, they would be expected to retain their Earthly interpretations, despite the admitted shift in their syntactic cousins. After all, why should mere *travel* from one place to another obliterate one's capacity to have certain memories. Normal tourism doesn't have such an effect; why should travel to another planet with somewhat unusual geographical properties?

Consider an example. While still on Earth, Peter goes hiking in the mountains of northern New Zealand. Here he comes across Lake Taupo and is startled to see the famous tenor Luciano Pavarotti floating on its pristine waters. They talk amiably for a while and Peter goes off flushed with excitement.

Understandably enough, this experience of Peter's gives rise to many subsequent memories on his part, and to beliefs based upon them. Consider the ones that are tokened while he is still on Earth, many years prior to his trip to Twin Earth. There can be no question about their content. They are about Pavarotti and Lake Taupo and water.[13] In fact, I take it, they would be about Pavarotti even in a possible world in which Pavarotti has a twin living in Los Angeles who Peter doesn't know about but whom he subsequently meets. Even after he does so, the memories of the encounter on Lake Taupo, and the beliefs based upon them, would continue to be about Pavarotti, not Twin Pavarotti.

Well, some years go by and Peter is moved surreptitiously to Twin Earth and becomes happily ensconced there. Eventually, some of the tokens of his mental names come to refer to the twin counterparts of the familiar Earthly figures, and some of the tokens of his general terms come to express twearthly properties: some tokens of 'Pavorotti' will come to refer to twin Pavarotti and some tokens of 'water' will come to mean **twater**.[14]

One day, perhaps while reading a schedule of upcoming events, Peter is moved to reminisce about the occasion when he saw Pavorotti swimming in the waters of Lake Taupo. He calls up vivid and accurate representations of the scene. Of course, he takes himself to be remembering scenes involving the singer he is reading about now. But he isn't. His memories, intuitively, are about the Earthly Pavarotti, the Earthly Lake Taupo, and Earthly water, previous perceptions of which are the sustaining cause of his later ability to recall what the scene looked like and how it felt.

In the situation described, Peter's externally individuated thought tokens are not epistemically transparent to him. In particular, Peter's language of thought contains token expressions that possess different semantic values, despite being of the same syntactic type. And yet, clearly, Peter does not know that they do. Tokens of 'Pavarotti', 'water', and 'Lake Taupo', in sentences expressing memories and beliefs about that memorable occasion, will mean Pavarotti, water and Lake Taupo, respectively; whereas other tokens of that type, in sentences expressing beliefs about his current environment, or current desires, will intuitively mean twater and twin Pavarotti. From the inside, however, there will be no indication of this: as far as Peter is concerned, they will appear to express precisely the same contents.[15]

Externalism, then, would appear to be robustly inconsistent with the epistemic transparency of thought contents: thoughts that have the same content may look to introspection to have distinct contents (as in Pierre and Paul), and thoughts that have distinct contents may look to introspection to have the same content (as in Peter). Externalist contents fail, in other words, to possess the feature that Dummett proclaimed "undeniable." What problem, if any, does this pose?

I shall argue that the problem is this. We don't just ascribe thoughts to a person in order to say something descriptively true of him. We use such ascriptions for two related purposes: on the one hand, to enable assessments of his rationality and, on the other, to explain his behavior. As these matters are currently conceived, a thought must be epistemically transparent if it is to play these roles. Without transparency, our conceptions of rationality and of rational explanation yield absurd results. We manifest our recognition of this fact by barring *de re* thoughts—thoughts which intuitively lack epistemic transparency—from figuring in assessments of rationality and psychological explanation. However, if we abandon transparency even for *de dicto* thoughts, and hence in effect altogether, then we must either jettison the notion of rationality and with it the practice of psychological explanation that it underwrites, or we must show these notions can be refashioned so as not to yield absurd results. The problem is that the first suggestion is wild and there appears to be no obviously satisfactory way of implementing the second.

The Apriority of Logical Properties and De Re Thoughts

We may usefully begin with a discussion of *de re* beliefs. By a '*de re* belief' I shall mean, *by stipulation*, a belief that is individuated by the objects it is about. Such beliefs are typically reported with the use of an 'of-clause', rather than a 'that-clause', as in,

> Jane believes *of* the piano that it is ugly,

and their content is given by a *Millian* proposition. Clearly, and in contrast with fully conceptualized, referentially opaque *de dicto* beliefs, it is both necessary and sufficient for the distinctness of two *de re* beliefs applying the same predicate that they concern distinct objects.

I should emphasize here that I am not presupposing, for the purposes of this discussion, either that this account explicates the 'intuitive' notion of *de re* belief, if there is such a thing, or even that there actually are *de re* beliefs in this stipulated sense. I am only interested in the question what would be true of such beliefs, if there were any.

Well, one claim that is often made in connection with such thoughts is nicely expressed by Burge:

> individual entities referred to by [referentially] transparently occurring expressions, and, more generally, entities (however referred to or characterized) *of* which a person holds his beliefs do not in general play a direct role in characterizing the nature of the person's mental state or event. *The difference [between such entities] does not bear on Alfred's mind in any sense that would immediately affect explanations of Alfred's behavior or assessment of the rationality of his mental activity.* ... Moreover, it seems unexceptionable to claim that the obliquely occurring expressions in propositional attitude attributions are critical for characterizing a given person's mental state. Such occurrences are the stuff of which explanations of his actions and assessments of his rationality are made.[16]

Burge's striking claim here is that *de re* beliefs, in contrast with fully conceptualized, referentially opaque, *de dicto* beliefs, don't enter into *assessments of a subject's rationality or psychological explanations of his behavior*. Here, for the purposes of further illustration, is Jerry Fodor making a similar point (Fodor's emphasis is on psychological explanation):

> Suppose I know that John wants to meet the girl who lives next door, and suppose I know that this is true when "wants to" is construed opaquely. Then, given even rough-and-ready generalizations about how people's behaviors

> are contingent upon their utilities, I can make some reasonable predictions (guesses) about what John is likely to do: he's likely to say (viz., utter), "I want to meet the girl next door." He's likely to call upon his neighbor... .
>
> On the other hand, suppose that all I know is that John wants to meet the girl next door where "wants to" is construed [referentially] transparently; i.e., all I know is that it's true of the girl next door that John wants to meet her. Then there is little or nothing that I can predict about how John is likely to proceed. And this is *not* just because rough-and-ready psychological generalizations want *ceteris paribus* clauses to fill them in...[17]

What reasons do Burge and Fodor offer in support of their respective claims? Burge, actually, has very little to say on the matter (though in fairness to him, I should point out that the essay from which the citation is drawn is largely concerned with other questions). And the exact interpretation of what Fodor has to say would take us too far afield. So without concerning myself overmuch with why these authors believe that *de re* beliefs are unfit for the purposes of content-based psychology, let me offer my own explanation. Let's begin with the question about rationality.

Suppose that Jane sees a wholesome-looking apple. She thinks *de re* of the apple that it is wholesome. That is, she comes to believe the Millian proposition

<apple_1, is wholesome>.

She subsequently sees the same apple with its blemished side exposed. She thinks *de re* of this apple that it is not wholesome. That is, she comes to believe the Millian proposition

<apple_1, is not wholesome>.

Two things are true of Jane in this case. First, her *de re* beliefs about the apple logically contradict each other: the (Millian) proposition subtended by the one is **p** and the one subtended by the other is **not-p**. And, second, she cannot recover from this condition on an *a priori* basis; to discover that the beliefs contradict each other she would have to learn an empirical fact, namely, that the apple involved in the first thought is identical to the apple involved in the second. That the two thoughts logically contradict each other is not introspectively accessible to her.[18]

Clearly, a similar case can be described to illustrate the fact that the logical *consistency* of Millian propositions is also not necessarily introspectively accessible. In general, then, the point is that the logical properties of *de re* propositions are not knowable *a priori*. The question is: How might this fact help explain why *de re* beliefs are unfit for the purpose of assessments of rationality?

The Apriority of Logical Properties and Norms of Good Reasoning

The answer derives from our conception of the nature of rationality and, in particular, of what it is for someone to be a good *reasoner*. What does a person have to do in order to count a good reasoner? Clearly, it is not at all a question of knowing empirical facts, of having lots of justified true beliefs about the external world. Rather, it is a matter of being able, and of being disposed, to make one's thoughts conform to the principles of logic on an *a priori* basis.[19] A surreptitiously envatted brain—transplanted from its normal adult body into a vat and attached to a computer that seamlessly duplicated and continued its previous course of experience—could be as good a *reasoner* as it ever was, despite the sharp escalation in the number of its false beliefs about the external world. Or so, at any rate, our conception of rationality requires us to think.

So, rationality is a function of a person's ability and disposition to conform to the norms of rationality on an *a priori* basis; and the norms of rationality are the norms of logic. We may, if we wish, put matters in a far less committal way: let's say that being *minimally* rational is a matter of being able to *avoid* obvious violations of the principles of logic, given enough time to reflect on the matter and so on. But even relative to this very minimal notion of rationality we would appear to have stumbled onto a problem. For according to this view, our Jane, who innocently believed of one and the same apple that it is both wholesome and not wholesome, would appear to stand convicted of irrationality: she believes a pair of Millian propositions that contradict each other, but she is unable to recover from this predicament on an *a priori* basis, no matter how long she may be given to reflect on the matter. But, intuitively, there is nothing irrational about her. Therefore, either our conception of rationality is mistaken or we have to find some other systematic and non-arbitrary way of absolving Jane's cognitive behavior. Obvious conservative solution: Bar an agent's *de re* thoughts from entering into assessments of her rationality. The ban on contradictory belief and invalid inference is preserved, but only in application to *de dicto* beliefs, just as Burge says.

The Apriority of Logical Properties and Psychological Explanation

Precisely parallel considerations explain why *de re* thoughts are also unfit for the purposes of *psychological explanation*. Since rationality is taken to consist in the ability and disposition to conform to the principles of logic on an *a priori* basis, *any* rational subject, regardless of his external conditions, may be expected to obey certain laws (or counterfactual-supporting generalizations): namely, those generalizations that mirror the introspectively obvious logical consequences of a person's propositional attitudes. Thus, our ordinary psychological practice of explaining and predicting behavior is built upon appeal to such laws as this:

> If S occurrently believes **p** and occurrently intends to **F** if **p**, and if S has no independent reason for not **F**'ing, then S will intend to **F** or, at the very least, will be disposed to intend to **F**.
> If S intends to **F** iff **p**, but does not believe **p**, but merely **q** instead, (where **p** and **q** are logically independent propositions), then S will not intend to **F**.

The trouble is that perfectly rational subjects will *not* obey these generalizations, when they are construed as quantifying over *de re* thoughts.

Thus, suppose that Jack intends to call the FBI whenever he is within 50 feet of a spy. And let's suppose that he believes of the dean of his College that he is a spy. The dean in fact is currently seated next to him at the beach, disguised as a lifeguard. So, *de re*, he believes of the 'lifeguard' seated next to him, that he is a spy. Yet he doesn't budge. Yet Jack would appear to be a perfectly rational person.

The elementary generalizations upon which the practice of psychological explanation depends fail to hold for *de re* thoughts. Therefore, either there is something wrong with our conception of rationality and, hence, with the generalizations that it underwrites, or we must find some way of justifying our belief in these generalizations, consistent with their failing to hold in this case. Obvious conservative solution: Except a subject's *de re* thoughts from psychological explanations of his behavior. The elementary generalizations on which psychological practice depends are preserved, but only in application to *de dicto* beliefs, just as Burge and Fodor say.

It seems to me, then, that we have before us the general answer to the question: What considerations might underwrite Burge's claim that *de re* thoughts are unfit for the purposes of assessments of rationality and psychological explanation? The answer is that both these enterprises require that the logical properties of the propositional attitudes they manipulate be knowable *a priori*; and the logical properties of *de re* thoughts aren't. (Again, I don't claim that this is the answer that Burge himself would give.)

Reference and Epistemic Transparency

Why, though, are the logical properties of referentially individuated thoughts—whether these be *de re* thoughts normally so-called, or thoughts involving Millian proper names—not knowable *a priori*? What makes these thoughts special in this regard?

The answer, of course, is that the logical properties of referentially individuated thoughts fail to be *a priori* just because these thoughts fail to be epistemically transparent. It is precisely because—and only because—a subject is unable to tell *a priori* whether the thought he expresses with

> Tully is bald

is the same as the thought he would express with

> Cicero is bald

that he is unable to tell whether the thought he would express with

> Tully is bald and Cicero isn't

is or is not a contradiction, on a Millian construal. Referential transparency entails epistemic opacity.

It is worth noticing in this connection that violations of the transparency of sameness and violations of the transparency of difference induce different sorts of logical defect. A thinker for whom the transparency of sameness is false may well fail, as we have seen, to be able to tell *a priori* that a given proposition is in fact a logical consequence of other propositions that he believes. That it is such a consequence is a fact he will be able to discover only *a posteriori*, by learning some empirical facts.

Thus, Pierre both believes that

> He lives in London

and believes that

> If he lives in London ('Londres'), he lives in the same city as Oscar Wilde lived.

Yet he is unable to draw the conclusion that he lives in the same city as Oscar Wilde lived in.

Such a thinker, however, need not be supposed ever to actually reason invalidly. It is consistent with the falsity of the transparency of sameness, in other words, that all the simple inferences that look *a priori* to such a thinker to be valid, are valid; what is falsified is the claim that all the simple inferences that are valid, will necessarily so look.

On the other hand, a thinker who suffers from the converse defect—failure of the transparency of difference—will suffer from the converse failing. For such a thinker, certain inferences may well look valid, when they in fact are not. And that they are not is a fact he will be able to discover only *a posteriori*, by learning some empirical facts.

Peter provides an appropriate example. Since his language of thought contains token expression that differ in semantic value *despite* being of the same syntactic type, he will be tempted to think that certain inferences are instances of, e.g., *modus ponens*, when they in fact aren't. Thus, he might muse to himself as follows:

> Whoever floats on water, gets wet.

This thought, by virtue of expressing a general quantified proposition, is to be regarded as having Twearthly content, i.e. as being about twin-floating and twin water. Now, however, he combines it with a memory belief that he would express with the sentence

> Pavarotti once floated on water

to conclude

> Pavarotti once got wet.

The inference will seem valid to him; but it arguably isn't. The second premise, by virtue of expressing a memory belief that is rooted in an Earthly experience, will be about Earthly floating and Earthly water. True premises, aided by a failure of univocity that Peter is in principle not in a position to introspect, will combine to produce a false conclusion.

Returning to the main line of argument, we see, then, that the *fundamental* answer to the question: Why are *de re* thoughts *unsuitable* for the purposes of assessments of rationality and psychological explanations, is this: It's because *de re* contents (and Millian propositions quite generally) are not epistemically transparent. By contrast, fully conceptualized *de dicto* thoughts, nowhere subject to co-referential substitution, are supposed to be, as Burge rightly says, "the stuff of which explanations of his actions and assessments of his rationality are made."

If the diagnosis on offer is correct, however, this contrast can hold up only if fully conceptualized *de dicto* thoughts *are* transparent. But as Loar's example of Paul and my example of Peter show, fully conceptualized *de dicto* contents will themselves fail to be epistemically transparent—and hence will themselves fail to be suitable for the purposes of psychological explanation and assessments of rationality—if they are individuated externalistically in the manner that Burge advocates. On what basis, then, does Burge distinguish between them? Why are externalistically individuated, and, hence, *non-transparent, de dicto* thoughts held to be suitable for the purposes of rational psychology, when *de re* thoughts are conceded not to be?

The Apriority of Logical Properties and the Introduction of Sense

It seems to me that once epistemic transparency is identified as a semantically significant thesis, its role in a variety of important disputes in the philosophy of language and mind becomes obvious. I shall discuss two of these: the role it has played in the canonical argument for the thesis that names have *sense* and not merely reference, and the role it plays in generating Kripke's puzzle about belief.

Beginning with the former, many philosophers would probably resist the claim that transparency plays a part in the canonical argument motivating a non-Millian view of names, because they would resist the claim that there is any such *argument*. Most philosophers write as if it's merely obvious—and, hence, in need of no argument—that someone might be in a state truly described by

(1) Mary believes that Ali was a champ

but not thereby in a state truly described by

(2) Mary believes that Clay was a champ.

It's worth seeing, however, that it isn't merely obvious. It becomes compelling, as I shall now argue, only when one makes the assumption, left implicit by Frege, that beliefs involving proper names are fit for the purposes of assessments of rationality and hence must be epistemically transparent. It is only under the terms of this assumption that one gets an argument for the referential opacity of proper names in the first place.[20]

For consider how a belief in referential opacity is typically motivated. We are given a case which goes like this: Mary sincerely asserts 'Ali was a champ'. She also sincerely asserts 'Clay was not a champ'. She asserts these sentences even though it is clear that, as she is using the names 'Ali' and 'Clay' they refer to one and the same legendary boxer. Now, given the following principle for reporting beliefs

Jones' sincere assertion of 'p' expresses his belief that p

we may conclude that

(1) Mary believes that Ali was a champ

and that

(3) Mary believes that Clay was not a champ.

It is important to notice, however, that nothing *so far* bars us from supposing that beliefs involving names *are* referentially transparent, and, hence, that (1) is equivalent to

(2) Mary believes that Clay was a champ.

For all that this would entail is that Mary has contradictory beliefs, a state of affairs that is, presumably, perfectly possible. We need to be given a reason why an ascription of contradictory beliefs is unacceptable in the present instance.

Otherwise, we would have no case illustrating, and consequently no argument for, the referential opacity of beliefs involving names. What is that reason?

We get such a reason only if we insist that beliefs involving proper names must be fit for the purposes of assessments of rationality and psychological explanation, and hence must have logical properties that are knowable *a priori* – must, that is, be transparent. Armed with such an assumption the argument for referential opacity is finally enabled. For the assumption insists that the attribution of a contradictory pair of beliefs involving proper names is acceptable only if the fact that they contradict each other is *a priori* available to the subject to whom they are attributed. Yet it seems perfectly clear in this case that no matter how much Mary might search her own mind, she won't discover that the belief she expresses with 'Ali was a champ' is referentially (and, hence, on a Millian construal, logically) inconsistent with the belief that she expresses with 'Clay was not a champ'. With the insistence in place, then, it follows, that (1) and (3) could not be attributing logically contradictory beliefs and, hence, that (1) is not equivalent to (2). If we ignore, as Frege himself evidently did, views that attempt to accommodate this result by analyzing belief in terms of a three-place relation, we get Frege's conclusion: there must be a level of semantic description of beliefs involving names other than the referential.

Kripke's Puzzling Pierre

We may observe the very same dialectic at work in Kripke's famous Pierre case. In one scenario (I won't discuss the other), Pierre assents both to

Londres is pretty

and to

London is not pretty.

Kripke argues that, in this case, there is no saying what it is that Pierre believes, no satisfactory belief ascription. His overall idea is to protect *Millian* theories from the charge that they generate absurd belief ascriptions, by showing that parallel absurdities can be generated solely from principles constitutive of belief as such, and without reliance on distinctively Millian principles (for example, substitutivity):

> When we enter into the area exemplified by...Pierre, we enter into an area where our normal practices of interpretation and attribution of belief are subjected to the greatest possible strain, perhaps to the point of breakdown. So is the notion of the *content* of an assertion, the *proposition* it expresses.[21]

But is it really true that we are unable to say what it is that Pierre believes in the case as described? Why can't we say, applying our disquotational principle, that Pierre believes *both* that London is pretty and that London is not pretty? Kripke writes:

> ...there seem to be insuperable difficulties with this alternative as well. We may suppose that Pierre...is a leading philosopher and logician. He would *never* let contradictory beliefs pass. And surely anyone, leading logician or no, is in principle in a position to notice and correct contradictory beliefs if he has them... .[Pierre] cannot be convicted of inconsistency: to do so is incorrect.[22]

Here we see the assumption of epistemic transparency playing an explicit role: Pierre cannot be ascribed contradictory beliefs because "anyone...is in principle in a position to notice and correct contradictory beliefs if he has them"; and yet in this case it is clear that he will not be able to do so. But the blame ought not to be placed on the very idea of belief or propositional content as such, at least not in the first instance. Rather, the source of the problem lies in the fact that Kripke is working both with the requirement that content be transparent and with a notion of propositional content that falsifies that assumption. The impression of a puzzle is generated by keeping two conflicting elements at play at the same time. No wonder, then, that no satisfactory belief ascription to Pierre is forthcoming.

A Simple Solution?

Our story thus far has unfolded as follows. We have the view that rationality, or at any rate, good reasoning, is the disposition to conform to the principles of logic on an *a priori* basis. This view, in conjunction with a non-transparent conception of propositional content, yields highly counterintuitive results. Our willingness to exclude *de re* thoughts in particular, and Millian thoughts in general, from the province of rationality-based psychology manifests our recognition of this fact. It commits us to the assumption, unless we are to forego rationality-based psychology altogether, that *de dicto* contents are epistemically transparent.

Yet most of us adhere to conceptions of propositional content, chief among them *externalist* conceptions of propositional content, that have it as a consequence that *de dicto* contents are not transparent. We cannot have it both ways. We must either reject such conceptions of propositional content, or we must show how to refashion the idea of reasoning so that it no longer consists in the disposition to conform to logic on an *a priori* basis.

But isn't there in fact a simple way of refashioning the idea of reasoning that will take care of the problem cases?[23] Well, how would it go?

We can't just say: A person is absolved for believing a contradiction provided that he doesn't—or can't—'see' that it's a contradiction. Irrationality often consists in precisely such failure. We want in some sense to capture the fact that Pierre or Peter or Paul are *blameless* for not seeing the contradictions that the stories attribute to them. The question is how that is to be done.

The only proposal I can think of is this: A thinker is to be absolved for believing a contradiction, provided that the contradictory character of the proposition he believes is inaccessible to mere *a priori* reflection on his part.

The trouble is that, against the background of a non-transparent conception of propositional content, any contradictory proposition will satisfy that description. Since on a non-transparent conception, it is precisely not available to mere *a priori* reflection that a given belief is a belief in a contradiction, practically any contradictory belief will be absolvable under the terms of this proposal. The only exceptions will be those beliefs of which it is simultaneously true (i) that they are beliefs in contradictions, (ii) that the subject *believes* them to be beliefs in contradictions, and (iii) which he nevertheless refuses—mysteriously enough—to change his mind about. Any such subject would undoubtedly count as exemplifying a form of extreme irrationality; but he could hardly be considered the normal case.

Conclusion

There is a pervasive tension between our conception of rationality and the practice of psychological explanation it underwrites, on the one hand, and currently dominant conceptions of mental content, on the other. The former presuppose what the latter deny. One or the other conception must be reconsidered.[24]

Notes

1. This paper is extracted from a longer unpublished manuscript, much of which dates back to 1987. Some portions of the present paper overlap with portions of "Externalism and Inference" in E. Villanueva (ed.): *Rationality in Epistemology* (Atascadero: Ridgeview Press, 1992). For helpful reactions either to this paper, or to the material in the larger manuscript, I am grateful to Barry Loewer, Stephen Schiffer, Jennifer Church, Crispin Wright and David Velleman; participants in seminars at Princeton and Michigan; and audiences at St. Andrews, Stony Brook, Campinas and Oxford.
2. *Truth and Other Enigmas*, (Cambridge, Mass: Harvard University Press, 1978), p. 131.
3. See Hilary Putnam: "The Meaning of 'Meaning,'" in his *Mind, Language and Reality* (Cambridge: Cambridge University Press, 1978).
4. See Tyler Burge: "Individualism and the Mental," in *Midwest Studies in Philosophy*, 1979. Burge, of course, regards externalism to be the correct moral of the thought experiments.
5. Throughout this paper, I shall use 'externalism about mental content' to refer only to the sorts of dependence on external factors that are encouraged by the Putnam and Burge thought experiments. Any number of other sorts of dependence are also imaginable, but I take it that no case has been made for taking them seriously as models of *our* propositional attitudes.
6. "Reply to Burge," unpublished ms.
7. Though I certainly don't wish to *assume* that Davidson himself would develop it

along these lines. For a similar proposal see Ernie Lepore and Barry Loewer: "Solipistic Semantics," in *Midwest Studies in Philosophy*, 1980.

[8] See Kripke: "A Puzzle About Belief," reprinted in Salmon and Soames (eds.): *Propositions and Attitudes* (Oxford: Oxford University Press, 1988).

[9] "Social Content and Psychological Content," in Grimm and Merrill (eds.): *Thought and Content* (Tucson: University of Arizona Press, 1987). As I explain below, Loar himself uses his example to illustrate a different thesis.

[10] Loar himself uses the case to argue for individualism about thought content. His idea is to concede that the beliefs would be *reported* as the externalist claims, while still maintaining that their psychologically relevant contents would be individualistically individuated.

[11] The assumption that there could not be disambiguated tokens of an ambiguous word type in the language of thought is widespread in the philosophy of mind. See, for example, Jerry Fodor: *Representations* (Cambridge: MIT Press, 1981), p. 227.

[12] "Individualism and Self-Knowledge," *Journal of Philosophy*, December 1989, p. 652.

[13] In this example, I am not taking a stand on whether the intuitive semantics of proper names is Millian or not. I intend the example to work on either reading.

[14] Obviously, my Twin Earth is richer than Putnam's in containing counterparts for many more objects, properties and people than his.

[15] Ideally, this example would be described at much greater length and would include a discussion of various objections and replies. For some further discussion see my "Externalism and Inference," Stephen Schiffer's "Boghossian on Externalism and Inference" and my "Reply to Schiffer" all in *Philosophical Issues*, *op. cit.*

[16] "Other Bodies," in Woodfield (ed.): *Thought and Content* (New York: Oxford University Press, 1982), p. 99, my emphasis.

[17] "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology," reprinted in *Representations* (Cambridge: MIT Press, 1983), p. 234-5.

[18] I am assuming here, and in the rest of this paper, that logical properties are, in the first instance, properties of *propositions*, not of the sentences which express them. I cannot defend this assumption here, except to say that it seems to me incredible to suppose that we cannot, in point of principle, evaluate a languageless creature's logical abilities. I hope to discuss this issue at greater length elsewhere.

[19] Notice that I am talking primarily about 'good reasoning,' rather than 'rationality' more widely construed. I take good reasoning to involve norms concerning the manipulation of propositions already at hand; I leave it open whether rationality involves something more.

[20] Here I follow a point made by Kripke in "A Puzzle About Belief."

[21] p. 135.

[22] p. 122.

[23] In discussions of this paper, I have often encountered the impression that there may be.

[24] Since writing this paper, I have come across two recent discussions of epistemic transparency—Ruth Millikan: "White Queen Psychology," unpublished ms; and Joseph Owens: "Cognitive Access and Semantic Puzzles," in Anderson and Owens: *Propositional Attitudes* (Chicago: CSLI, 1990). Both writers agree that the notion has played an important role in traditional discussions, yet both advocate that it be eliminated: they concur in seeing a belief in transparency as a mere residue of a 'Cartesian' conception of mind, Millikan calling it the "last myth of the given." However, neither writer satisfactorily explains, in my view, how the crucial notions of reasoning and rationality are to be plausibly refashioned in its absence.