For a book symposium on R. Stalnaker's, *Our Knowledge of the Internal World,* forthcoming in *Philosophical Studies.*

The Transparency of Mental Content Revisited

Paul Boghossian
NYU

Our Residual Cartesianism

Robert Stalnaker has written a short but extraordinarily rich book, one that sheds light on a number of important and difficult issues in the philosophy of mind. Its overarching claim is that a Cartesian view of the mind continues to color our conception of a range of philosophical issues, even as mainstream thought in the philosophy of mind has tended to move away from Cartesianism. In his book, Stalnaker aims to provide a more thoroughly externalist view of the mental, one which, he claims, will defuse some of the puzzles to which our residual Cartesianism gives rise.

One of the puzzles to which Stalnaker applies his general strategy is the one generated by the need to have mental contents satisfy a thesis that I labeled (following Dummett's closely related thesis about linguistic meaning) *transparency.* This thesis has two parts, the "transparency of sameness" and the "transparency of difference."

> (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.[1]

I made two main claims about transparency. First, that when mental contents violate one or both of these transparency theses, we get cases in which a thinker who intuitively looks fully rational, and is merely missing some empirical information, is made to look as though he is committing simple logical fallacies in his reasoning. I claimed, in other words, that violations of transparency blur the distinction between errors of reasoning and errors of fact.

I also claimed that *anti-individualist* conceptions of mental content, in the manner of Burge and Putnam, violate both of these transparency theses.

[1] Boghossian, (1994), 36. I am using "a priori" in this connection to mean "independently of outer experience" and therefore in a way that's consistent with knowledge being a priori if it's based on inner experience or introspection.

If these two claims are correct, this poses a problem for anti-individualist conceptions of mental content, for any adequate theory of mental content should observe the distinction between errors of reasoning and errors of fact.

Stalnaker agrees with this desideratum:

> I agree … that we ascribe thoughts in order to explain action, and to assess the reasoning of thinkers, and that such explanations and assessments cannot turn on facts that are inaccessible to the subject. If we could be wrong, on empirical grounds, about the contents of our own thoughts, then we could be wrong, on empirical grounds, about the validity of our reasoning, and this seems incompatible with the idea that we can separate the assessment of the reasoning from assessment of the truth of the premises….(114)

He argues, though, that

> … an anti-individualist account of the facts can be reconciled with a suitably qualified version of a principle of epistemic transparency. I won't defend the principles of the transparency of sameness and difference exactly as Boghossian states them, nor will I provide an alternative formulation. My aim is just to show that an anti-individualist thesis about content attribution is compatible with an account of reasoning that is clear about the difference between errors of reasoning and errors of fact, as … any adequate account must be. (114-115)

And as I indicated above, Stalnaker's main idea is that

> the puzzles about knowledge of content involve a mixing of internal and external perspectives, and that if we give a thoroughly externalist account both of knowledge and of the role of content in the characterization of states of mind, we can give a plausible account of what we know and what we do not know about what we are thinking. (115)

Let us look at how this is supposed to work in detail.

Stalnaker on Slow Switching

The puzzle case I described involved a variant on a case first introduced by Burge, in which an earthling, Peter, is suddenly and unwittingly transported to Twin Earth in such a way that he suffers no discernible disruption in the continuity of his mental life and never discovers the relocation that he has undergone. Most philosophers agree that, initially, the thoughts Peter expresses with his word 'water' would involve the concept *water* and not *twater* (the concept that, as anti-individualists believe, Twin Earthlings have). However, it is also widely agreed that, after a certain period of time, at least some of Peter's thoughts will come to involve the concept *twater* and will no longer involve the concept *water.*

I argued for two main points. First, that while we would expect, on general anti-individualist grounds, that some of Peter's tokens of 'water' would express thoughts involving the concept *twater*, we would also expect others of his 'water' thoughts to retain their earthly contents and so to involve the concept *water*. For example, suppose Peter remembers some earthly experience in which he was looking out on a lake. It would be natural to describe this as a case in which Peter was thinking about water and not twater. Second, I argued that it followed that we could therefore imagine a case in which Peter combines a *water* thought with a *twater* thought to reason invalidly, as follows:

> (Water)
> 'There was water in that lake'
> 'There is water in this lake'
> Therefore
> 'There is water in both lakes.'

Intuitively, though, Peter is not irrational, he is merely lacking some empirical information.

As Stalnaker notes, philosophers have responded to this puzzle case in one of two main ways.

The first, by Michael Tye, in effect denies that Peter has both the *water* and the *twater* concepts available to him after the switch; rather, the *water* concept is displaced (or replaced) by the *twater* concept.[2]

The second way of responding, developed by Burge, concedes that Peter will have access to both concepts after the switch, but insists that no equivocation in reasoning can result because of the role of "preservative memory" in reasoning. The idea is that some appropriate principle applicable to reasoning will guarantee that the three tokens of 'water' in the (Water) inference will all express the same 'water' concept, whichever one that happens to be.[3]

I find both of these responses implausible. The first because it involves an implausible view of concept attribution and the second because, as some have pointed out, it requires saying that which concepts are deployed in the (Water) inference will depend on the *order* in which one thinks the premises.[4]

Interestingly, though, Stalnaker's view is that there is actually something right in each of these solutions, but that they will continue to seem implausible unless they are embedded in the more thoroughly externalist picture that it is his aim to set out and recommend. He says:

[2] Tye, (1998).

[3] Burge, (1998). See also Schiffer (1992), which first proposed this sort of line.

[4] See Brown, 2004, p. 178.

> Each of these responses to Boghossian's argument may appear strained, but I think the source of this appearance is that there is something misleading about locating the cognitive shift in Peter, rather than in the context in which knowledge is attributed to him, or in the circumstances that are appropriate for describing his cognitive situation. (122)

Instead of thinking of content attributions as a matter of trying to get something right about a thinker's internal states, Stalnaker says, we should think of them, rather, as parts of attempts to explain how thinkers can have a capacity to make their actions depend on the way the world is and a disposition to make their actions depend on the way they take the world to be. If we view content attribution in this way, we will not seek some overarching theory that will tell us in some attributor-context-independent way what a thinker is thinking. Rather, we will use different resources on different occasions, possibly depending both on the thinker's and the attributor's circumstances. Transparency will get satisfied not because thinkers are acquainted with objects whose essence they can't mistake, but because an apt explanation of a thinker's capacities and dispositions will often call for satisfying it.

Stalnaker gives a helpful summary of his preferred way of thinking about attributions of intentional content. It will be useful to quote him at length:

> As I understand it, Boghossian's view of the target of his criticism is an essentially internalist picture, with an externalist component grafted onto it. Our thoughts are something like internal sentences to which we have access because they are part of the internal mental world. But (this is the externalist part) these mental sentences, individuated by their content, have essential properties that are extrinsic to the mind, and so are not accessible to the person who is thinking the thought. But we shouldn't think of access to our thoughts as access to an internal vehicle of representation. According to a more thoroughly externalist picture, we should think of the representation of states of knowledge and belief, and the content of occurrent thoughts, this way: Thinkers are things with a capacity to make their actions depend on the way the world is, and with dispositions to make their actions depend on the way they take the world to be. Theorists and attributors of thought characterize these capacities and dispositions by locating the world as the thinker takes it to be in a space of relevant alternative possibilities. The theorist uses actual things and properties to describe these possibilities, and that is why content depends on facts about the actual world….When the way the world actually is diverges from the way the subject takes it to be with respect to the identity and nature of things, and in particular when he or she conflates distinct things, or thinks of one thing as two, or when the changes in the world as it is diverge from changes in the world as it is taken to be, we may find it difficult to characterize a world according to the thinker that is apt for describing that person's cognitive capacities and dispositions. But our descriptive resources are rich and flexible, and in context, we can usually find a way. What counts as a correct description of the world according to the thinker may depend on the

> attributor's context. A principle of epistemic transparency is satisfied, according to this picture, not because the thinker is directly acquainted with an inner object that has an inner content essentially, but because an apt description of a thinker's cognitive state, if it is to explain the rational capacities and dispositions it is intended to explain, must represent the way the world is according to the thinker in a way that satisfies it. (130-1)

Discussion of Stalnaker's Alternative Approach to Mental Content

Before looking at how Stalnaker's more thoroughly externalist alternative approach to mental content is supposed to help us with our puzzle, I want to comment briefly on his characterization of the difference between our respective approaches to mental content.

I think that Stalnaker is right that the target of my discussion is a conception of mental content that incorporates both internal and external components. But I would describe those components a little differently, and I would emphasize both how well motivated they are and how widely accepted they are by those working in the philosophy of mind.

About the internal component, I don't think it's right to say that it consists in the view that mental contents are carried by word-like vehicles that can be identified independently of their meanings. All that the internal component is committed to is that a thinker can be introspectively aware that he has an occurrent thought when he has one, something that seems so intuitive as to need no argument. This can be made vivid by talk of consciously accessible word-like vehicles for contents, but that is not essential for the intuitive claim that we have a capacity for introspective awareness of occurrent thought (without prejudging whether we also have such a capacity for thought's content).

Stalnaker also equates talk of word-like vehicles for contents with talk of concepts. And he blames concept talk for generating what he takes to be a fruitless puzzle about whether Peter has one 'water' concept or two.

> The debate about how to understand the slow switching scenario is often framed as the question whether our unfortunate character has (at the later time) one "water" concept or two. Does Peter lose one concept or gain another, at the time of the "switch", or does he keep the old concept when he gains the new one, while systematically confusing them with each other? This way of putting the issue suggest that the issue is about Peter's internal cognitive mechanism: does he have two different folders labeled "water" in his mental file cabinet, one old and one new, or did he throw out the old one, moving its contents into the new one, when the switch occurred? (122)

He complains, however, that the arguments about the issue do not consider evidence about the form that mental representation takes, a speculative matter at best, and that the forms of mental representation do not seem relevant.

Once again, I think two distinct views are being conflated here. I accept concept talk, but I think of concepts not as mental words, but as propositional constituents. This is not entirely non-committal, of course, it is a lot less committal than a language of thought picture with consciously accessible mental words.

I'm not sure whether these differences in characterization are more than merely terminological, but I find it more illuminating to describe the conception of mental content that I endorse, and that Stalnaker opposes, as consisting in the following two theses. First, that there are *attributor-independent* facts about what contents a thinker is thinking. Second, that such contents are broadly Fregean in the sense that they don't involve *actual* objects and properties, but rather modes of presentations thereof. Let's call the first the thesis of Intentional Realism and the second, Fregeanism. These two theses are independent of one another. I'll call their conjunction, the Common View.

Now, many philosophers hold something like the Common View. And one way of representing the point of my paper on transparency was to say that when the Common View is combined with Twin Earth style anti-individualism about mental content, we get the sorts of Peter-type puzzles outlined above. And one good way to represent Stalnaker's argument in the chapter under discussion is to say that he thinks that the culprit here is not the anti-individualism per se but rather the Common View itself, that once we get rid of the Common View, and replace it with his more contextual and extensional conception of mental content, the puzzles disappear.

Before we look at what life might look like without the Common View, I want to emphasize the extent to which theorists working within the philosophy of mind accept it. I believe that the Common View is presupposed by anyone who regards the Twin Earth thought experiments to have established an anti-individualism about mental content of a *surprising* kind.

That the Twin Earth experiments presuppose Intentional Realism will be readily granted. But it may not be as obvious that they presuppose Fregeanism. To see that they do, we need only bear in mind that it's obvious that the *reference* relation is widely individuated, that facts about what our thoughts refer to supervene on facts outside our heads. Even Frege, for whom reference is determined by fit with the conditions laid down by senses, would have to think this, for, even on his view, what our thoughts refer to will depend on what in the world satisfies those conditions.

So, it's obviously true of Russellian propositions that grasp of them is widely individuated. The surprising claim, the ground for which was prepared by Putnam but which was most fully developed by Burge, was that the Twin Earth experiments showed that even grasp of *sense* had to be thought of as widely individuated. That is why those experiments are best understood as involving Fregean propositions.

Be that as it may, let us look at whether Stalnaker's alternative picture is able to defuse the puzzle about Peter.

Surprisingly, Stalnaker doesn't offer a direct treatment of Peter. Instead, he tells four variants on the Peter story, ones that involve less science fiction, and he indicates for some of them how he would be inclined to treat them in such a way that transparency is satisfied. Two of these stories involve proper names ("Treasure Island" and "Aston-Martin") and two involve indexicals ("this," "that" and "tomorrow"). True to his more contextual approach, different cases receive different treatments depending upon the circumstances of the case and the attributor's needs and purposes.

I should note that these variant stories would *not* have sufficed for my purposes in the paper on transparency, because I was trying to show that one could get a failure of the transparency of difference for senses, concepts, or modes of presentation, of properties. I took it to be obvious, as I said in the paper, that one would get such failures in the case of terms, such as proper names or demonstratives, that are best regarded as directly referring expressions.

Nevertheless, we can still look at Stalnaker's treatment of the variant stories to see if we can put together a possible response to the Peter case on the basis of the resources he provides.

The variant scenario whose treatment would be most directly adaptable to the case of Peter is the one involving a thinker, John, who conflates two different ships that he is looking at in a harbor to reason in a way that would be expressed linguistically as follows:

> (Ship)
> (P1) *This* ship (pointing to the bow) is an aircraft carrier;
> (P2) *This* ship (pointing to the stern of the second ship) is British;
> Therefore, there is a British aircraft carrier in the harbor.

Stalnaker says:

> Here it seems clear that the two are premises are about different ships, and so both are true. But the conclusion is false, so isn't John making an unwitting logical error? Isn't he mistakenly confusing the proposition that *this* ship is an aircraft carrier with the distinct one that *that* one is? Here I think the best way for the theorist to represent the reasoning is to take it to involve a false tacit presupposition, a suppressed premise, that *this* ship is *that* one, rather than a false belief that the two thought have the same content. [footnote suppressed] This way of representing the reasoning does not assume that John has entertained the possibility that the two ships are different – the possibility that distinguishes the two propositions – or that the proposition that excludes this possibility is in any way encoded at some perhaps subpersonal level in John's cognitive apparatus. Most of what we presuppose is presupposed simply by not recognizing the possibilities in which the presuppositions are false. The explicit statement of the tacit presupposition is part of the theorist's representation of the situation.

We could imagine applying this treatment to the case of Peter. We could say that, in his (Water) reasoning, Peter is tacitly presupposing that twater is water, that there is a false suppressed premise in his argument. This would render the reasoning valid. To make this solution work, we would be forced to say, as Stalnaker does in the variant case that he's discussing, that attributing this presupposition to Peter does not assume that Peter has entertained the possibility that water and twater are distinct substances, or that the proposition that excludes this possibility is in any way encoded in Peter's cognitive apparatus.

I think there are various grounds for being dissatisfied with this sort of solution to our problem, but I will focus on one main line of thought.

Stalnaker is aware, of course, that attributions of validity-restoring presuppositions can't be totally unconstrained, since some people do reason invalidly and it wouldn't always be right to attribute to them presuppositions that would absolve them of logical error. He doesn't say much about what does constrain such attributions, but what he says suggests that the reason it would be appropriate to credit John with the presupposition has to do with the fact that John would *understand* the proposition that *this* ship is an aircraft carrier while *that* one isn't, and would not take the claim to be a simple contradiction.

Applied to the case of Peter, we could try saying that we could credit Peter with the 'twater is water' presupposition because he would understand the proposition that this stuff he is looking at is twater and not water.

But it's entirely unclear that Peter would understand this proposition in his current state or that he would not regard it as a contradiction. Of course, he doesn't have two terms for water and twater. So, without any further instruction, or the introduction of new terminology, all we would be able to do is say to him: "This stuff is water and not water," meaning *water* by the first occurrence and *twater* by the second. This would obviously do us no good.

To get Peter to understand that there are two substances where he thinks there is one, we would have to explain to him how these substances are individuated and how this is said, by philosophers of language and mind, to affect what concepts one possesses.

So the way the story would have to unfold in the case of Peter is that we could say that we can credit him with the relevant presupposition because he would, after sufficient instruction and explanation, rationally come to accept that there are two substances here, and two concepts, and not just one of each.

But there are two problems with this tack. First, it seems very implausible to claim that I can be said to presuppose P because, given sufficient instruction and explanation, I would rationally come to accept P. Suppose I were to give an incomplete proof of Fermat's Last Theorem, not realizing that it depended on the truth of the Taniyama-Shimura Conjecture and perhaps not ever having entertained this proposition. Could we really say that I was presupposing that the Taniyama-Shimura conjecture was true, because I could be brought to rationally accept it, given sufficient instruction and explanation?

A worse problem derives from the requirement that Peter's acceptance of the 'twater is not water 'proposition be *rational,* a condition that would appear to be required. For the main problem with violations of transparency, as I see it, is that they lead us to conflate someone's lacking empirical information with his failing to be rational. They do so because our notion of rationality constitutively involves the idea that a rational person will not make simple logical errors.

Now, though, if we need to say that someone counts as not violating transparency, and hence as not making a simple logical error, because he can be credited with tacitly making a certain presupposition; and he can be credited with making that presupposition because he would be willing to rationally accept a particular proposition given sufficient instruction and explanation; and he can be said to rationally accept the relevant proposition because he arrives at it without making simple logical errors; then it seems as though we will have turned around in a tight little circle. And we will therefore have failed to explain how it is possible for someone not to violate the transparency requirement, given anti-individualist views of content.

Up to this point, I have been willing to grant Stalnaker his rejection of Intentional Realism and to allow that we may have to find ways of saving transparency on an ad hoc basis, using different resources depending on the context. I now want to briefly raise a question about Stalnaker's rejection of Intentional Realism.

Stalnaker says, as we have seen:

> Thinkers are things with…dispositions to make their actions depend on the way they take the world to be. Theorists and attributors of thought characterize these…dispositions by locating the world as the thinker takes it to be in a space of relevant possibilities….What counts as a correct description of the world according to the thinker may depend on the attributor's context.

Now, there is a puzzle, it seems to me, seeing how facts about a thinker's mental content could depend, quite generally, on facts about the attributor's context, including his purposes and beliefs. For purposes and beliefs are, of course, states with intentional content. So if we say that A's thinking p depend on B's purposes and beliefs, we would appear to be assuming that there is some fact of the matter about what B's mental contents are that don't themselves depend on what some other attributor would say. So it looks as though we are committed to at least some mental contents getting fixed independently of an attributor's variable purposes and beliefs. But then it looks as though Intentional Realism must be true for at least some contents.[5]

[5] The issue here is reminiscent of the kind of issue one might raise for views, like Daniel Dennett's, to the effect that a thinker's having intentional states just amounts to his being treated from the intentional stance.

If this is right, Stalnaker's view faces a dilemma. He either has to show that certain contents may legitimately be exempted from his generalized contextual and externalist approach, or he must show that applying his approach across the board won't lead to the sort of instability just outlined.

As with all the many interesting claims that Stalnaker makes, this one could use a great deal more discussion than I am able to give it here. But perhaps it will suffice to have given him the opportunity to explain his views further.[6]

References

Boghossian, P. (1994) "The Transparency of Mental Content," *Philosophical Perspectives* 8: 33-50.

Brown, J. (2004) *Anti-Individualism and Knowledge* (Cambridge, Mass.: MIT Press).

Burge, T. (1998) "Memory and Self-Knowledge," in P. Ludlow, and N. Martin (eds.), *Externalism and Self-Knowledge* (Stanford: CSLI Publications), 351-71.

Sciffer, S. (1992) "Boghossian on Externalism and Inference," in *Philosophical Issues* 2, 29-38.

Tye, M. (1998) ""Externalism and Memory," *Proceedings of the Aristoyelian Society* 72: 77-94.

[6] Thanks to Stephen Schiffer for comments on this essay.