What is Wrong with Arguments from Reference?

Daniel Cohnitz

August 20, 2007

Abstract

Sometimes philosophers draw philosophically significant conclusions from theories of references. This practice has been attacked [Sti96, BS98, Bis03, MMNS] for two different reasons. One line of attack against arguments from reference tries to show that they are invalid, the other attempts to show that empirical results from social psychology undermine all such arguments. In this paper I show that this criticism of arguments from reference is misplaced. There is nothing wrong in principle with arguments from reference.

1 Introduction

Stephen Stich[1], Mike Bishop and others launched two attacks against so-called "arguments from reference". These are arguments that establish a conclusion about ontology or epistemology by departing from a substantive theory of reference as one of their major premises. In a first attack on such arguments [Sti96, BS98, Bis03], arguments from reference are criticised for being invalid. They, allegedly, always omit a crucial premise although it is hard to see how the proposers of arguments from reference could actually establish this premise. In a more recent attack, that Stich co-authored with Ron Mallon, Edouard Machery, and Shaun Nichols [MMNS], Stich criticises arguments from reference for being undermined by empirical results of social psychology.

I will argue that both criticisms of arguments from reference are misplaced. Following the reconstruction of arguments from reference in actual philosophical debates that is common in both attacks, it turns out that these arguments are valid. It also turns out that the empirical data does not undermine arguments from reference *per se*. However, there is indeed something wrong with some arguments from reference so reconstructed. Thus

[1] For brevity's sake I will most of the time refer to Stephen Stich as the author behind the arguments, although many of the papers I refer to he co-athored with others. I tried to make it clear always which argument was co-authored with whom, if it was co-authored with anyone.

these arguments from reference would be bad arguments if they existed. I will show, however, that Stich et al. have largely misreconstrued their sources — the bad arguments from reference apparently were never put forward by anyone.

I take the confusions that I try to clear up in the arguments of these authors to be deep rooted misconceptions of what a theory of reference is supposed to explain. Accordingly, this paper is more than just a discussion note on these specific arguments, but a systematic contribution to the philosophy of language. It might also be of interest to philosophers not particularly interested in these writings of those authors, but interested in substantive theories of reference in general.

2 Arguments from Reference

Stich's first attack on arguments from reference was an attack in part against his former self. In *Deconstructing the Mind* [Sti96] Stich declares that he no longer finds his own argument for eliminativism convincing. As he explains, his own argument for eliminativism was an argument from reference and he now believes that all such arguments are fatally flawed. In the subsequent papers [BS98, MMNS] Stich includes also arguments from reference in areas of philosophy other than the philosophy of mind in his criticism. In my discussion I will mainly use the example from the philosophy of mind though.

Arguments from reference are supposed to have the following general structure:

1. A certain substantive theory of reference is defended.
2. It is deduced from the theory of reference whether a given expression denotes or not.
3. A conclusion about ontology or epistemology is drawn.

Stich seems to concentrate mainly on conclusions about ontology. The examples that are given in the papers cited are arguments for or against the existence of propositional attitudes (like beliefs or desires), metaphysical realism, the existence of races, and the apriority of moral knowledge. To see how arguments from reference are supposed to work, I will explain the argument for eliminativism in some detail. In section 5 we will also discuss the arguments for and against metaphysical realism in more detail.

2.1 An Argument for Eliminativism

Apparently, only very few authors have *explicitly* put forward an argument from reference as characterised above. Classifying certain arguments as arguments from reference often results from *critical reconstructions* of actual arguments. It is claimed that from the explicit premises of the argument under scrutiny the intended conclusion could not follow, then a theory of reference is presented as a possible premise that would complete the argument, typically with the intention to attack this premise in the following critique of the argument. This was also the case, it seems, in the central example, *the argument for eliminativism.*

Eliminativism in the philosophy of mind is the thesis that folk psychological states, such as beliefs and desires, do not exist. In Paul M. Churchland's words:

> Eliminative materialism is the thesis that our common sense conception of psychological phenomena constitutes a radically false theory, a theory so fundamentally defective that both the principles and the ontology of that theory will eventually be displaced, rather than smoothly reduced, by completed neuroscience. [Chu81, 67]

As said above, Stich concentrates on the ontological part of this claim. In [Sti96] he starts out to inquire how this conclusion could follow from the usually provided premises. The premises that most arguments for eliminativism would share (and that are left undisputed here) are:

(P1) Beliefs, desires and various other mental states can be viewed as 'posits' of a widely shared commonsense psychological theory, often called 'folk psychology'.

(P2) Folk psychology is a seriously mistaken theory because some of the central claims it makes about the states and processes that give rise to behaviour, or some of the crucial presuppositions of those claims, are false or incoherent.

from these premises would then either first concluded

(C1) Beliefs and desires will not be part of the ontology of the science that ultimately gives us a correct account of the workings of the human mind/brain.

and then

(C2) Beliefs and desires do not exist.

or the other way around.[2]

[2] Stich explains:

> The simplest route goes directly from the premises to the conclusion that beliefs, desires and other posits of folk psychology do not exist. And of course, if that's right, it follows that no mature science which succeeds in explaining human behavior will invoke the posits of folk psychology. Beliefs, desires and the rest will not be part of the ontology of the science that ultimately gives us a correct account of the workings

As Stich explains, he himself thought that the conclusion would more or less automatically follow as soon as the premises (P1) and (P2) could be established, being confident that the suppressed premises in this argument could be spelled out and defended without running into any difficulties. As far as it concerns himself, he believed that the link between the premises and the conclusion would be provided by David Lewis' theory about the meaning and reference of theoretical terms. Indeed, in [Sti83] he already recapitulated Lewis' theory about the meaning of theoretical terms — and might thus had it in the back of his mind when arguing against eliminativism — but apparently didn't explicitly use it as a premise in an argument for eliminativism.[3]

As Stich reconstructs Lewis' theory, theoretical terms acquire their meaning from being implicitly defined by the theory in which they occur as the occupants of the causal roles specified by that theory, whatever they happen to be. A simplification of this account could be that a theory is true if and only if something instantiates its causal roles, if and only if its theoretical terms denote. However, as Stich notes, Lewis is ready to admit that a theory that is only a little wrong, could still be true enough for its theoretical terms to denote. But if a theory is very mistaken, such that nothing comes even close to instantiating it, the implicitly defined theoretical terms will be denotationless.[4]

With this account of the meaning of theoretical terms, the argument for eliminativism could perhaps be completed into a deductively valid argument (with slightly adjusted premises):

> of the human mind/brain. The second route that an eliminativist's argument can follow reverses the order of these two conclusions. From the Premises it initially concludes that folk psychology will not be part of any mature science. This, in turn, is taken to support the stronger conclusion that these folk psychological states do not exist. [Sti96, 4]

Stich then uses the stronger conclusion (C2) in his discussion, not coming back to the two different routes. As we will see in the final section of this paper, which route to take does matter for whether or not there is a good argument from reference that could be made for eliminativism.

[3] Stich obscures this in [Sti96]. After quoting a passage from [Lew72] in order to explain how arguments for eliminativism could be reconstructed as arguments from reference, taking Lewis' theory as a premise, Stich adds in a footnote:

> This is the second time I've quoted this passage from Lewis. The first time was in (Stich 1983, 15–16) [=[Sti83, 15–16]] where I was laying the groundwork for a series of arguments for eliminativism. [Sti96, 84, FN 24]

Besides that he actually quoted it on pp. 16–17, he also did not quote it in support of his argument for eliminativism, but in order to explicate the theory-theory of belief ascriptions, that he then goes on to criticise.

[4] Thus, the biconditionals in the formulation above must be replaced by conditionals: 'if and only if' by 'only if', or else 'true' and 'instantiates' be qualified by 'almost true' and 'almost instantiates'.

(P1*) 'Belief', 'desire' and other mental terms are theoretical terms of folk psychology.

(P2*) Folk psychology is a seriously mistaken theory (such that nothing comes even close to instantiating it).

(P3) (From Theory of Reference) Theoretical terms of a seriously mistaken theory are denotationless.

(C2) Beliefs and desires do not exist.

So far, so good. Since (P3) never explicitly appeared in arguments for eliminativism, but seemed to be tacitly assumed (at least by Stich himself), the discussion around that argument centered on whether (P1*) and (P2*) would be true.

2.2 The Counter-Argument Against Eliminativism

This changed when William Lycan, allegedly, countered this argument in a series of papers that "woke" Stich from his"dogmatic slumbers"[Sti96, 34]. Lycan noted (for example in [Lyc88]) that Lewis' account of the meaning of theoretical terms wasn't the only and perhaps not even the most popular game in town:

> Unlike [David Lewis], and unlike [Dennett] and [Stich], I am entirely willing to give up fairly large chunks of our commonsensical or platitudinous theory of belief or of desire (or of almost anything else) and decide that we were just wrong about a lot of things, without drawing the inference that we are no longer talking about belief or desire. To put the matter crudely, I incline away from Lewis's Carnapian and/or Rylean cluster theory of reference of theoretical terms, and toward [Putnam's] causal-historical theory. [...] I think the ordinary word "belief" (qua theoretical term of folk psychology) points dimly toward a natural kind that we have not fully grasped and that only mature psychology will reveal. I expect that "belief" will turn out to refer to some kind of information-bearing inner state of a sentient being [...], but the kind of state it refers to may have only a few of the properties usually attributed to beliefs by common sense. Thus I think our ordinary way of picking out beliefs and desires succeeds in picking out real entities in nature, but it may not succeed in picking out the entities that common sense suggests that it does. [Lyc88, 31–32]

As Stich presents the case in [BS98] and in [MMNS], Lycan

1. points out that by assuming a different theory of reference, namely a causal-historical theory rather than a decsrptivist theory of reference, “the eliminativist conclusion does not follow” [MMNS, ??], and
2. “is not content to stop here” [BS98, 38], but starts his own flight to reference strategy, when “by assuming a different theory of reference than the eliminativist, Lycan draws the opposite conclusion, viz. that beliefs and desires *do* exist.” [MMNS, ??].

Stich’s point is that while all other premises remain the same, Lycan can argue for the opposite conclusion by just plugging an alternative theory of reference into the argument.

In order to receive a *valid* counter-argument for anti-eliminativism from the argument above, we are led by Stich to reconstruct Lycan’s argument in the little quote given above in the following way:

(P1*) ‘Belief’, ‘desire’ and other mental terms are theoretical terms of folk psychology.

(P2*) Folk psychology is a seriously mistaken theory (such that nothing comes even close to instantiating it).

($P3^-$) (From Theory of Reference) Theoretical terms of a seriously mistaken theory do nevertheless denote.

($C2^-$) Beliefs and desires *do* exist.

Whether this reconstruction is fair to Lycan will be discussed in the final section. In any case, this seems to be how Stich and his co-authors understood Lycan’s argument and considered it an instance of the type “argument from reference”. We should now turn to Stich’s diagnosis of arguments of this type.

3 The First Attack: The Myth of the Missing Premise

Confronted with Lycan’s counter argument one might be tempted to move the discussion about eliminativism to the discussion over which theory of reference is the correct one. Since the other premises in the argument for eliminativism are left undisputed in Stich’s reconstruction of Lycan’s counter-argument, and since only one of the two arguments can be sound, the fate of eliminativism (or of desires and beliefs, for that matter) seems to depend on which theory is the true account of reference.

No doubt, there are many suggestions on the market. Lewis’ theory is certainly not *the* descriptivist theory, but one (though well elaborated) member of a whole family of theories. Similarly, Putnam’s version of the causal-historical theory is just one among many. It is also not obvious that all members of the descriptivist family of theories would support the elimnativist version of the argument, neither is it obvious that all versions of a causal-historical theory would support the anti-eliminativist conclusion, or even obvious that it must be a member of one of these two families that will eventually yield a true account of reference. Accordingly, one should first decide which theory is the true account of reference and then see whether this account can support one of the two arguments:

> So it looks like both eliminativists and their opponents would be well advised to turn their attention to the theory of reference. They have to determine which sort of account of reference is right and find some plausible way of explaining away the objections to that account. [...] In any event, the theory of reference has now moved to center stage. [Sti96, 37]

However, as Stich tells us, just when he was about to engage in the vibrant and exciting discussion over theories of reference, he stumbled over a rather odd problem: how to decide — from a given set of alternative accounts of reference — which of them is the true (or at least best) one?

3.1 Two Different Explananda

If we are to say what a theory of reference is actually a theory of, there are, according to Stich, two different possible answers, each with its own methodology how to decide — from a given set of alternative accounts of reference — which of them is the true (or at least best) one. While at the same time there doesn't seem to be any consensus (as far as philosophers seem to be at all aware that there are two alternatives) which of the two research programs is to be followed in the philosophy of language. Since this problem seems to be a minor aspect of Stich's overall argument, I will not go deeply into this. However, as we will see in what follows, Stich's opinion here reveals a fundamental misunderstanding of what a theory of reference is, which is why it is nevertheless worth being mentioned.

The first possible answer to the question raised above would be that a theory of reference is simply an account of "folk semantics" [Sti96, 41]. If confronted with actual and hypothetical cases and asked what we intuitively believe about the reference of a given term in such a case, we can normally come up with answers. These intuitive reactions might be interpreted as resulting from a tacit, internalized theory of reference that we are subconsciously applying when evaluating these cases. Theories of reference are reconstructions of these tacit theories we have. The better a so reconstructed theory agrees with our intuitive judgments about reference, the closer it is to the "true account" of reference.

The second possible answer that Stich can think of (what he calls the "proto-scientific" project [Sti96, 38]) is that our folk semantics is — just as our "folk physics" — not more than a reconstruction of our pre-scientific understanding of reference, that could, in principle, be massively mistaken about the true account of reference. Perhaps for heuristic purposes it might be useful to know what we intuitively believe about reference, but even if it is perhaps a good method to form initial hypotheses about reference, it should not be our primary method when selecting from a set of alternative proposals. In this case, other considerations might matter more, including how fruitful a theory of reference proves to be when put to use in empirical theories (history of science, linguistics, cognitive psychology and evolutionary biology are among the candidates that Stich can think of here).

Since Stich assumes that it is unclear which of these two proposals is asked for when it is asked for a true account of reference, he decides to use both accounts in order to assess the alternative theories of reference. However, since Stich believes that so far not much progress has been made towards the proto-scientific project, he choses to start with the folk semantic research program just in order to stumble over the next odd problem. Again, the point that Stich makes here is only a minor point in his overall argument, but we will come back to this in section 8.2.2, and it also leads to the first argument why arguments from reference are flawed in principle.

3.2 The Argument from Idiosyncracy

Starting the folk semantic research program one should begin with collecting some data. As a theory of reference seems to be just as good as it is able to predict our intuitive judgments, we might first try to get a sufficient stock of judgments in order to make a selection between alternative theories. Stich began to collect such data, in particular he tried to collect intuitive judgments about the reference of terms of seriously mistaken theories, using probes of the following kind:

> Here's what Democritus said about atoms (or what Mendel said about genes): . . .
>
> Here's what modern science says about atoms (or genes): . . .
>
> If we assume that modern science is correct, was Democritus actually referring to the same things that modern science is? (Or, . . . was Mendel actually referring to the same things that modern science is?) [Sti96, 46]

These cases were designed on actual cases from the history of science (as the examples above) or similar fictional cases were made up. As Stich reports, this data turned out to be useless for determining *the* true account of folk semantics:

> After doing some preliminary analysis on about a hundred responses, however, I decided to drop the project, for what I found was that [...] [t]here was little agreement on the cases [...]. Perhaps there were some systematic and widely shared principles about reference underlying the very messy data that I was collecting, but if there were, I couldn't find them. My conclusion from this little pilot study was that there probably are no systematic, widely shared principles that guide people's judgments when they are presented with questions about the reference of terms in mistaken theories. [Sti96, 47]

However, it gave rise to a new worry. Stich interprets this result such that folk semantics leaves it indeterminate whether claims like[5]

$$\text{`\ldots is a belief' applies to something} \qquad (\alpha)$$

are true or false. But since folk semantics is all there is to the truth or falsity of such claims (in the folk semantic research program) and since

$$\text{Beliefs exist if and only if `\ldots is a belief' applies to something} \qquad (\beta)$$

it seems that also the ontological question, namely whether

$$\text{Beliefs exist} \qquad (\gamma)$$

is indeterminate. Thus, the ontological dispute between eliminativists and anti-eliminativists is pointless if folk semantics does not settle whether or not (α) is true.

However, let's assume that folk semantics settles the issue in one way. Let's assume that of the many possible word-world[6] relations, such as

'. . . stands in an appropriate causal chain to this token of . . . '

or

'. . . is the unique satisfier of the cluster of descriptions associated by speakers of this language with . . . ',

[5] In what follows, I will — in accordance with common practice — not talk about "referring terms", but about "application" when talking about the relevant word-world relation of predicates and about "denotation" when talking about the relevant word-world relation of names.

[6] The relations below are, for the sake of elegance, rather world-word relations, but the order of the objects in the relations doesn't really matter. Saying that x denotes y iff the usage of x is appropriately linked to y is the same as saying that x denotes y iff y is appropriately linked to the usage of x.

one is indeed favored by our folk intuitions, then this, Stich assumes, must be so for contingent historical reasons. We just happened to settle on this word-world relation, rather than another.

But if the reference relation is selected for a contingent reason, how interesting can an ontological conclusion be that depends on it?

> If the difference between [our intuitively sanctioned word-world mapping] and [an alternative relation] is simply that one of these idiosyncratic mappings happens to have gotten itself embedded in our folk semantics, while the other has not, then it's hard to see why we should care whether or not the extension of '. . . is a belief' is empty. If the only thing that distinguishes [our intuitively sanctioned word-world mapping from any of the possible alternatives], then the fact that '. . . is a belief' refers to nothing just isn't very interesting and important. But if that isn't interesting, then neither is the fact that beliefs don't exist. [Sti96, 50]

Thus the upshot of the argument from idiosyncrasy is that eliminativism is neither troublesome nor particularly interesting.

Stich is silent on how these considerations connect with the ones summarized earlier that started from the messy empirical data to the conclusion that the truth of (γ) is indeterminate. Perhaps Stich's intention was that we put the two arguments together to a dilemma:

(D1) Only folk semantics could determine whether a term denotes something. [Assumption]

(D2) Folk semantics either does not determine the truth of (α) or it does determine it. [Assumption]

(D3) (β) [Assumption]

(D4) (First Horn): Folk semantics does not determine the truth of (α). [First disjunct of (D2)]

(D5) (Conclusion of First Horn): (γ) is indeterminate. [From (D4), (D1) and (D3)]

(D6) (Second Horn): Folk semantics does determine the truth of (α). [(Second disjunct of (D2)]

(D7) Every determination of a reference relation by a social/historical process is contingent on uninteresting facts. [Assumption]

(D8) A determination of the reference relation by folk semantics would constitute a determination by historical accident. [Assumption]

(D9) (α) is contingent on uninteresting facts. [From (D6), (D7), (D8)]

(D10) The truth of (γ) is contingent on uninteresting facts. [From D9 and D3]

(D11) Whatever is contingent on uninteresting facts is itself uninteresting. [Assumption]

(D12) (Conclusion of Second Horn): The truth of (γ) is uninteresting. [From D10 and D11]

Assumption (D1) is just a statement of the folk semantic research program: whatever the folk theory of semantics is, it can't be false, since it is all there is that could determine the denotation of terms. Assumption (D2) might qualify as a logical truth, (D3) Stich would presumably regard as an almost analytic truth.[7] The assumptions (D7) and (D11) might be a bit controversial, but one could perhaps argue that 'uninteresting fact' is meant as 'a lot less interesting fact than the fact that the ontological debate over eliminativism *prima facie* seemed to be about (and that philosophers find worth spending their time working on)', and with this many would presumably agree. (D8) is what Stich already argued for in chapter 5 of [Sti90]: there are many possible alternative world-word relations and folk semantics just happened to settle on one idiosyncratic possibility.

With that argument we would indeed arrive at the startling conclusion that the issue between eliminativists and anti-eliminativists is either indeterminate (and then their debate pointless) or uninteresting (and then their debate pointless). Apparently, it didn't take long for Stich to become convinced that this argument does not go through.

As Stich tells us [Sti96, 51–52], John Searle and Christopher Gauker convinced him that there must be something wrong with the second horn of the dilemma. It seems to me though that their point can actually be made for both horns. Apparently, Searle and Gauker noted a strange feature of the argument above: it doesn't seem to be restricted to the eliminativism issue, but — if it worked — would undermine *all* ontological disputes. To see this, consider a variant of the argument, with (α), (β), and (γ) replaced by:

$$\text{`\ldots is a black hole' applies to something} \qquad (\alpha*)$$

$$\begin{gathered}\text{Black holes exist if and only if}\\ \text{`\ldots is a black hole' applies to something}\end{gathered} \qquad (\beta*)$$

$$\text{Black holes exist} \qquad (\gamma*)$$

[7] Stich does not believe in the analytic/synthetic distinction, but I guess he would have considered this premise as rather unproblematic truth when making the argument. Why I believe that will become clear in section 3.3.

Thus by the same strategy one could also argue to the conclusion that any debate over the existence of black holes is pointless, regardless of which arguments are brought forward in that debate. But this conclusion is truly *mad*. The question of whether black holes exist is an important question in modern physics. That should suffice as a *reductio* of the argument. Although this refutes Stich's argument from idiosyncracy, it does not tell us *where* the argument goes wrong.

According to Stich, the absurdity does not disappear if we simply assume that it was a mistake to settle on the folk semantic research program. After all, assuming that the true account of the reference relation will be provided by the eventual outcome of the proto-scientific project would make all ontological issues (over black holes, god, phlogiston, witches, propositional attitudes, etc.) dependent on that outcome. But, so Stich argues, it doesn't seem right to assume that a result in linguistics should settle such important matters, moreover since it also might be the case that the proto-scientific project leads to different answers about what the true account of reference is (perhaps one in cognitive psychology, the other in linguistics) and thus the physicist patiently waiting to learn from a theory of reference whether black holes exist, would get conflicting advise. Something must be wrong here.

3.3 The Argument from Invalidity

Stich's diagnosis, that he first presented in [Sti96] and then elaborated with Michael Bishop in [BS98] constitutes his first argument against "the flight to reference strategy" or — simply put — against "arguments from reference". As we have seen in 2.1 and 2.2, Stich started with the observation that the eliminativist argument could be turned into an anti-eliminativist argument by just changing one premise, i.e. the one that assumes a particular theory of reference.

To remind you, the general[8] structure of arguments from reference is this:

1. A theory of reference is explained and defended.
2. It is argued that — according to the theory of reference adopted in (1.) — a certain term, $\ulcorner t \urcorner$[9] (or perhaps a family of terms) does (or does not) denote something.
3. The conclusion about reference arrived at in step (2.) is used as a premise in an argument to an ontological conclusion: ts do (or do not) exist.

[8] We will now concentrate on ontology to flesh things out a bit more.

[9] The use of these corner-quotes is explained in footnote 10.

Stich's diagnosis is that arguments of that sort all fail since they all lack a support for the move from (2.) to (3.). Why should the fact whether $\ulcorner t \urcorner$ denotes entail anything about whether t exists? The extra premise needed for that move would be an instance of the following schema[10]

$$t \text{ exists if and only if } \ulcorner t \urcorner \text{ denotes something.} \tag{δ}$$

Instances of this schema have also been used above in the "dilemma", namely (γ) and ($\gamma *$). This premise, as Stich observes, is never made explicit when an argument from reference is made. Accordingly, it is also never argued for. But does it really need an argument? Couldn't we just take (δ) for granted?

According to Stich there might be two reasons for believing that (δ) is true: we might believe that (δ) is constitutive for reference, such that nothing could be a reference relation unless it satisfies all instances of (δ). Or we might believe that, although (δ) is perhaps not constitutive for the reference relation, the reference relation alluded to in (1.) satisfies all instances of (δ) anyway. Which of these two lines we take, in either case we seem to have to give a reason for supposing that the word-world relation that we alluded to in (1.) and assumed as the true account of reference, satisfies all instances of (δ). But this has not been done for any theory of reference yet and it is not clear how that at all could be shown. From this, Stich concludes that all arguments from reference are invalid: they are all enthymematic and it is not clear how the missing premise could be established.

Prima facie, this argument against arguments from reference does not

[10] Note that this is a schema, in which 't' is supposed to be substituted with (in case of denotation) names of the language the schema is in (some sort of English). '$\ulcorner t \urcorner$' is to be substituted with (in case of denotation) the same names of the language, but put between single quotation marks. The corner-quotes are used to remind the reader that we are talking about a specific term of English, not about the 20th letter of the alphabet. We will assume throughout the paper, also when making distinctions between meta- and objectlanguage, that English is the meta- as well as the objectlanguage, or at least that the objectlanguage and the metalanguage have the same expressive resources.

But there are also three other things you should keep in mind about this schema. First of all, things can go into and out of existence. Thus, although Julius Caesar existed once, he doesn't anymore. But 'Julius Caesar' denotes him, of course, still. This isn't a case in which the biconditional is false, but a case in which you need to read the left hand side accordingly. Similarly, terms might denote something at some points in time, but not at others. 'The present King of France' once denoted someone, but doesn't anymore. Thirdly, one and the same speaker might use an expression in one utterance in a referring way, and in another utterance in a non-referring way. In these last two cases it is of importance to understand the substitution instance of 't' (also the one within corner-quotes) on both sides of the biconditional in the same way. These aren't cases that falsify the biconditional, but cases in which language needs to be properly disambiguated. Failing to do so leads to confusion, as we will see below. How to disambiguate these cases depends on your preferences.

seem to be very strong. Many good arguments are enthymematic[11], and although Stich might not *see* how it can be established that the word-world relations alluded to in (1.) satisfy all instances of (δ), they nevertheless might satisfy them, and perhaps it is after all a trivial task to prove it. In other words: Stich needs to give us a reason to suppose that it will really be difficult to establish that a word-world relation satisfies all instances of (δ).

Neither in [Sti96] nor in [BS98] is it made very clear what the reason is to suppose that it will be difficult to establish that a reference relation satisfies all instances of (δ).[12] John Hawthorne, in a review of [Sti96] summarizes Stich's argument thus:

> [Stich's] attack is to cast suspicion on the very project of engaging with existence questions via semantic ascent. One makes no progress on the issue of whether there are any F's by asking whether 'F' is true of anything, since there is no hope of arriving at a robust and sufficiently uncontroversial theory of reference that will enable us to settle questions in the material mode via semantic ascent to the formal mode. If we can't agree about whether there are F's we won't be able to agree about whether 'F' is true of anything either. As Stich is thinking about it, one won't be able to wield a theory of reference to convince someone who disbelieves in F's to change his mind, since his disbelief in F's will make him suspicious about the theory of reference that is wielded against him. [Haw00, 481–482]

[11] Note that an argument does not have to be *deductively* valid in order to be valid. Of course, some arguments might have premises that we believe to be true and a conclusion that we believe to be clearly supported by the premises, although we might not be certain how exactly the premises link to the conclusion. Perhaps that is because a premise (we tacitly endorse) is missing, perhaps this is because the support relation between premises and conclusion is not deductive.

[12] As we see in what follows, John Hawthorne has probably misunderstood Stich's point. So did Don Gustafson in his review, when he summarizes Stich's point thus:

> Reference and word-world relations are too idiosyncratic to make simple eliminativism an interesting doctrine. Stich concludes, skeptically, that appeal to reference and semantic facts is simply not relevant to decisions concerning contested ontological questions. [Gus98]

As we have seen, this summary confuses the (rejected) argument from idiosyncrasy with Stich's actual argument from invalidity.

Other reviewers, like [Kit98] and [Cra98] just praise Stich's argument without bothering to say what it is and why it is praiseworthy or convincing. [Sch02] devotes a whole chapter to Stich's discussion of arguments for eliminativism, but manages not to loose a single word about his argument from invalidity. I take all this (either wrong or no reconstructions of Stich's arguments in places where one would expect them) to be evidence that Stich's argumentation is a little unclear.

As Hawthorne notes, that would be a bad argument, since it seems to assume that our convictions about issues of ontology always override our convictions about reference. That, of course, need not be the case.[13]

But it is also clear that *this* cannot be Stich's argument. This point he could have made already in the beginning, sparing everyone the trouble of finding a missing premise in the argument from reference. He should simply have argued that arguments from reference are bad just for the fact that they argue from convictions about reference to convictions about ontology, and since the latter are stronger than the former they are always unsuccessful, period.

But there is also a different argument one might reconstruct from the Hawthorne quote: perhaps the trouble with establishing that a given theory of reference satisfies all instances of (δ) is that any attempt to do so would beg the question against the opponent in the ontological controversy.

As we have said, a theory of reference identifies the reference relation with a certain word-world relation. Let's assume that we try to establish a descriptivist theory of reference, and thus assume the following:

$$\begin{gathered}\forall x \forall y (x \text{ satisfies the cluster of descriptions} \\ \text{associated by speakers of this language with } y \leftrightarrow y \text{ applies to } x)\end{gathered} \tag{ϵ}$$

But now in order to show that our suggested reference relation also satisfies all instances of (δ), we have to show that it satisfies all instances of

$$\begin{gathered}\forall y \exists x \exists z (y = \ulcorner F \urcorner \rightarrow ((x \text{ satisfies the cluster of descriptions} \\ \text{associated by speakers of this language with } y) \leftrightarrow (Fz))\end{gathered} \tag{ζ}$$

Perhaps we could do this by checking for a given term whether the left hand side of the biconditional is satisfied (which is an empirical question) and ask experts (or people sufficiently informed about what modern science knows about Fs, or the theories which feature $\ulcorner F \urcorner$) for intuitive judgement about its right hand side. However, in going through all the Fs, defending descriptivism would at some point involve showing that if (as in the case

[13] As Bruno Mölder pointed out to me, this argument isn't perhaps as bad as Hawthorne makes it look like, if the discussion were restricted to the ontological argument in the philosophy of mind. To block the arguments from reference in the debate about eliminativism it would be enough to assume that ontological convictions override semantic convictions *here*. I guess this is right (although I am not sure that it is obvious that the ontological convictions are here indeed stronger than whatever theory of reference one feels inclined to). However, let's not forget that Stich is in the business to give a principled argument against arguments from reference. For this he needs to establish the stronger claim, namely that ontological convictions always override. And this, I think, is far from obvious.

of eliminativism) there is no such x satisfying the conditions laid down in the antecedent of the biconditional given $\ulcorner F \urcorner =$ '…is a belief', there also is no z, Fz. But this consequent the anti-eliminativist need not accept, since he assumes that there *are* beliefs. Thus at least the intuitive judgement of the anti-eliminativist would *not* agree with the biconditional in that case. Wouldn't thus any attempt to show that the suggested reference relation satisfies all instances of (δ) simply be begging the ontological question?

It is *very* unlikely that it would, since there will hardly be two equally systematic and otherwise convincing theories of reference that will *only* disagree as to whether '…is a belief' denotes. But then the overall track record on all other instances of (δ) could decide which of the competing theories of reference is the right one. To claim that, even in the light of such an independent success record, no one could ever be convinced to give up an ontological conviction in favor of what seems to be otherwise the best account of reference, is again question begging against arguments from reference. The prospect for a convincing begging-the-question-accusation is thus rather dim.[14]

However, I don't believe that this is the *only* argument that Stich wants

[14] As we will see below, since we can show that the relevant theories of reference simply imply every instance of (δ), we need only argue that the theory of reference is true, in order to show that every instance of (δ) is true, and this could be done on the basis of considerations that have nothing to do with the instances of (ζ).

Anyway, in reaction to the argument above, Mike Bishop has suggested that my argument here would leave

(i) Arguments from reference are ineffective in the following sense: In all possible worlds, arguments from reference can work only in a small percentage of cases.

intact, which would be bad enough for arguments from reference. His argument is that in order to show that a theory of reference is true, we'd need to show for the vast majority of "easy cases", that (ζ) is true. But then (a) we need to have an independent method for solving ontological issues that works already reliably well in so many cases, and (b) only a minority of hard cases can be left for a theory of reference to decide. But why should we think that the method that is inappropriate to decide the majority of cases is appropriate for the few hard cases? Hence:

(ii) It is not obvious that even in the small minority of cases in which arguments from reference might be feasible, that they are the most effective way to resolve ontological disputes.

(i) shouldn't worry us much, I think, as long as these few cases are the philosophically interesting cases, or the cases in which our intuitions do not guide us to a clear ontological judgment. Bishop argues here for (ii) on the assumption that an argument from reference would have been *inappropriate* in all other cases. But the fact that we didn't use that method for arriving at an ontological judgment, doesn't make it "inappropriate". In the other cases we just had *two* appropriate methods to arrive at an ontological judgment. The idea of the few hard and many easy cases would rather be that our alternative method fails us in the hard cases, thus an argument from reference would then be the *only* appropriate method to arrive at the judgment. But as we will see below in detail, we can leave this discussion aside, since the considerations from (ζ) do not in fact play a role in arguments from reference.

to make (but see [Sti96, 59] where he apparently makes this rather hopeless argument and attributes it to Hartry Field). At least his argument with Michael Bishop in [BS98] seems more motivated from deflationary considerations and less involved with the question-begging-considerations just discussed. To see what Stich and Bishop might have in mind, we need to make a little detour via Alfred Tarski's constraints on definitions of truth.

As is well known, Tarski formulated a *criterion of material adequacy* for any definition of 'true', the famous "Critertion T". Following Wolfgang Künne's formulation of Criterion T, it says:

Criterion T. *A formally correct definition of 'true' for a given object language L in the metalanguage English is materially adequate if and only if it implies all sentences which can be obtained from the schema*

(T) *s is true in L if and only if p*

by substituting for the place-holder 's' a revealing designator of a declarative sentence of L and for the place-holder 'p' the English translation of that sentence. [Kün03, 183]

A designator is, in Künne's sense, *revealing* if and only if somebody who understands it can read off from it which (orthographically individuated) expression it designates. Revealing designators of expressions are often just quotational designators, the expression put between quotation marks.

Establishing that a definition of truth is materially adequate in this sense is normally not an easy task. It has to be shown that the definition implies all substitution-instances of (T). Now, although substitution-instances of (T) are all likely to be indifferently accepted by advocates of pretty much all possible conceptions of truth (such as coherentists, pragmatists, correspondence theorists, etc.), that does not prove that their truth definitions *de facto* imply all these instances. This is a complicated task for substantive theories of truth.

However, for some "deflationary" accounts of truth, it is perhaps easier to show that they entail all these instances, since these accounts hold that there is not much more to truth than the substitution instances of (T). Truth is a devise for disquotation — as some deflationists claim —, and that is really what the substitution-instances of (T) are about, there just is nothing more substantial to learn about truth.

Such a definition then, in turn, implies all these instances (it can simply consist in an infinite list of axioms, that exhausts all substitution instances of (T), provided for some technical difficulties).

Parallel to Criterion T, one could formulate criteria for the material adequacy of other notions in the theory of reference, such as *denotation*, *application*, etc.

Criterion A. *A formally correct definition of 'applies to' for a given object language L in the metalanguage English is materially adequate if and only if it implies all sentences which can be obtained from the schema*

$$\text{(A) } \forall x (\textit{t applies to x in L if and only if x is r})$$

by substituting for the place-holder 't' a revealing designator of a (perhaps complex, but monadic) predicate of L and for the place-holder 'r' the English translation of that predicate.

And again a deflationist about denotation or application could simply hold that there is nothing more to learn about 'applies to' or 'denotes' than what is said in the substitution-instances of schema (A). As Bishop and Stich explain, a deflationist has no troubles to establish that his theory of reference implies all these instances, since his theory just *is* all these instances:

> Since [deflationists] do not think that there is any unified conceptual or naturalistic reduction of reference, [schema (A)] captures all there is to be said about the reference relation in general. Thus there is no need to argue that the deflationists' substantive account of reference makes [all substitution instances of (A)] come out true. They do not offer a substantive account. [BS98, 40, FN 6]

However, substantive theories of reference will have to show that all substitution-instances of (A) come out true, either because they are submitted as (in Tarski's sense) formally and materially adequate definitions of the key notions of the theory of reference, or because else they cannot support an argument from reference. As Bishop and Stich explain, substitution instances of schema (A) can play the role of the missing premise in arguments from reference. Let's look again at the argument for eliminativism: first we assume a descriptivist theory of reference according to which theoretical terms of a massively mistaken theory do not apply to anything. Then turning to folk psychology and finding it massively mistaken, we infer that its terms do not denote anything; in particular, we infer

1. $\neg\exists x$('. . . is a belief' applies to x)

Now we can bring in our substitution-instance of schema (A):

2. $\forall x$('. . . is a belief' applies to x if and only if x is a belief)

and conclude

3. $\neg\exists x$(x is a belief)

Thus schema (A) is the missing ingredient that we were looking for to complete our argument from reference. Of course, as Bishop and Stich continue, the deflationist cannot make an argument from reference either, since in order to show that a substitution-instance of (A) is true, he has to argue from the right hand (ontological) side of the biconditional, and that *he* can't do without begging the question against the ontological opponent. Thus in [BS98] the worry is obviously *not* that arguments from reference fail, because a crucial premise cannot be provided without begging the ontological question. This would constitute a principled argument against the possibility of establishing the missing premise. They explicitly admit that they lack any such principled argument:

> We do not claim to have an argument showing that it is impossible to defend [the assumption that the substantive account of reference endorsed makes all instances of schema (A) true]. But since we have no idea how one would even begin to construct such an argument, and as far as we know no one else has ever tried, we are inclined to be more than a bit skeptical. What we do claim is that without a defense of this essential assumption, attempts to invoke the flight to reference are fatally flawed. [BS98, 40]

As I reconstruct their argument, it is rather that arguments from reference rely on a premise, namely that the theory of reference assumed satisfies Criterion A, that is *de facto* not established for any of the substantive theories of reference, and establishing it is a non-trivial task.

4 The Myth of the Missing Premise Exposed

As interesting as these considerations might sound, I'm afraid that this argument is nothing but a red herring. Instead of discussing this issue in the framework of Tarskian criteria for a theory of reference, we should perhaps return to the original question that Stich had raised about the final step in arguments from reference. To remind you, the problem was how to get from

1. A theory of reference is explained and defended.
2. It is argued that — according to the theory of reference adopted in (1.) — a certain term, $\ulcorner t \urcorner$ (or perhaps a family of terms) does (or does not) denote something.

to

3. The conclusion about reference arrived at in step (2.) is used as a premise in an argument to an ontological conclusion: ts do (or do not) exist.

As it seems, we will need something like

$$t \text{ exists if and only if } \ulcorner t \urcorner \text{ denotes something.} \qquad (\delta)$$

to bridge the gap. The task seems to be to show that substantive theories of reference imply instances of schema (δ). Let's confront that task in a somewhat more naïve way.

Let's first consider whether a substantive theory of reference could make (δ) false for any term, or allow that it is false for some term. This might help us to understand better what a theory has to guarantee in order to guarantee the truth of all instances of (δ). It seems that this could be the case, if the theory would allow, for example, for referring expression to non-existent objects. In this case, a term $\ulcorner t \urcorner$ could denote although no t exists. For example, 'Pegasus' could be considered a denoting term in such a Meinongian theory, denoting a non-existent object. Then

'Pegasus' denotes something

and

Pegasus does not exist,

would both be true and hence (δ) false for some terms.

In this case, our Meinongian theory of reference violates the biconditional in (δ) in the right-to-left direction, the direction relied on in Stich's reconstruction of Lycan's argument from reference.

Violations in the left-to-right direction, the direction relied on in the argument of the eliminativist, are not so easy to imagine. Anyway, here is a far fetched possibility: let's assume that after struggling with Russell's paradox for a long time you finally draw the radical conclusion that all terms of the theory of reference must not be iterated, or else the resulting sentence is simply false. Further reflecting on Grelling's paradox, leads you to accept (the perhaps intuitively plausible):

There are some things that no description applies to.

as true, while you can safely reject

'things that no description applies to' applies to something.

as false (perhaps you were worried that accepting the latter as true might force you to say something incoherent when asked what it is that this complex description applies to). This theory of reference, if it could be spelt out coherently (which I doubt), would clearly violate the left-to-right direction of (δ), since although 'ts exist' could be true on such a theory while '$\ulcorner t \urcorner$ applies to something' false.

This shows already that it is a highly nontrivial task to come up with a theory that would not imply (δ), and allow for possible cases that violate at least one direction of its biconditional. The theories used in arguments from reference and discussed by Stich and Bishop do, however, not allow for the possibilities just described. But if — under the acceptance of any of the theories in question — it is not possible that any of the instances of (δ) are false, then it is obvious what we should conclude from this: These theories *imply* (δ) for any given term. Let us see that in more detail.

Let us concentrate first on the right-to-left direction of (δ). We might wonder why a (standard) descriptivist theory of reference should make all instances of

$$\ulcorner t \urcorner \text{ denotes something} \rightarrow t \text{ exists} \qquad (\delta^{\rightarrow})$$

true. But inspection of the theory reveals that whenever the antecedent is true (a term $\ulcorner t \urcorner$ succeeds in denoting) then there is a unique object that satisfies a Russell-type definite description (or a cluster of these) associated with $\ulcorner t \urcorner$. Why should that guarantee that also the consequent is true? Well, because 't exists' is — according to the theory — analyzed as 'there is an object such that it uniquely satisfies the Russell-type definite description (or a cluster of these) associated with $\ulcorner t \urcorner$', and that this condition is satisfied is guaranteed by the truth of the antecedent.

Similarly for the causal-historical theory. The antecedent is true (a term $\ulcorner t \urcorner$ succeeds in denoting) only if there is an object standing in the right kind of causal chain with the present usage of $\ulcorner t \urcorner$. Why should that guarantee that also the consequent is true? Well, because 't exists' is — according to the theory — analyzed as 'there is an object such that it is standing in the right kind of causal chain with the present usage of $\ulcorner t \urcorner$', and that this condition is satisfied is guaranteed by the truth of the antecedent.

Let us now look at the other direction, which is actually the one needed for the argument for eliminativism:

$$t \text{ exists} \rightarrow \ulcorner t \urcorner \text{ denotes something} \qquad (\delta^{\leftarrow})$$

No surprises here. The antecedent is true, according to the descriptivist theory, if and only if there is a unique object that satisfies a Russell-type definite description (or a cluster of these) associated with $\ulcorner t \urcorner$, but that is a sufficient condition for successful reference of $\ulcorner t \urcorner$. Similarly for the causal-historical theory. The antecedent is true, according to *it* if and only if there is a unique object standing in the right kind of causal chain with the present usage of $\ulcorner t \urcorner$, but that is a sufficient condition for successful reference of $\ulcorner t \urcorner$, according to the same theory.

Of course, if we evaluate the antecedents and consequents of ($\delta^{\rightarrow}$) and ($\delta^{\leftarrow}$) each with a different theory respectively, we will get violations of (δ).

There might be an object standing in the right causal chain with the present usage of $\ulcorner t \urcorner$, but in fact no object satisfies a Russell-type definite description (or a cluster of these) associated with $\ulcorner t \urcorner$, or vice versa. But this can hardly constitute an objection to an argument from reference. Stich's and Bishop's objection was that we have not been given a reason to believe that the theories in question make all instances of (δ) true, but now we have seen that they actually do: *both* theories imply all instances of (δ).

Of course, there might be reasons to doubt the first premise of an argument from reference (namely if one believes that a different theory of reference than the one assumed is true), but Bishop and Stich can hardly claim that we have never been given reasons for believing that an assumed theory of reference is true, or reasons for believing that a rejected theory of reference is false. They were claiming that even after a theory of reference is explained and defended, we still need an additional reason to believe that that theory makes all instances of (δ) true, and this reason we have just provided. It does not require an additional premise, since it is implied in the first step, when the theory of reference is stated, for the theories considered.

Note that the reasons given for believing that a theory of reference is true *might* include intuitive judgments like the ones considered with the argument that Stich attributed to Field. Such judgments result from (actual or fictional) cases in which the conditions for reference are fulfilled (or not). Then we are asked whether we would intuitively think (on the basis of our background knowledge) that things by that name exist (or not). A theory might well come to counterintuitive results here. For example, if the theory is a holistic descriptivist (conceptual role) theory and proponents of that theory would be forced to conclude that bumble bees and jumbo jets don't exist, because our best theory of thermodynamics is incompatible with the otherwise also well entrenched belief that they fly, etc. This would certainly count as a counterintuitive implication of this theory of reference. But such reasons are given in the debate in the philosophy of language, of course, and as I argued above, there is nothing wrong in principle with arguing that way. Although we are using (intuitive) ontological judgments here to argue for or against a theory of reference, this need not be question begging if the reasons we are providing for a theory of reference comprise more than just the ontological questions at issue in particular arguments from reference.[15] Note that the latter is usually provided for by the simple

[15] This is, of course, familiar from all areas of philosophy. However, here is one example from Stich's own writings: In [Sti96, 245-246] Daniel Dennett is criticized for his instrumentalistic rebuttal of folk psychology. Stich argues that by the same strategy one could also save the theory of "immaterial spirits" of the alchemists, or Copernican astronomy from falsification. Stich's argument is that accepting a specific strategy would elsewhere lead to conclusions (here about the falsification of a theory by certain empirical results) that both participants in the debate would not be ready to accept (namely that also the intuitively quite clearly falsified

fact that philosophy enjoys a division of labour between the philosophy of language and ontology. Although arguments from reference show us in what way the two topics interrelate, the theories of reference alluded to in arguments from reference were not just cooked up in order to support a specific ontological argument. They were established and discussed in the philosophy of language as solutions to the problem how reference works, not as solutions to *specific* problems of ontology.

It is perhaps even more important to point out that, although reasons for believing that a theory of reference is true might include intuitive judgments, they need not. Since we have seen now that the theories of reference imply every instance of (δ), every reason for believing the theory is a reason for believing the truth of the instances of (δ). No ontological considerations need to enter here. We might be totally in the dark about what exists and what doesn't — if we have a reason to believe that one of our theories of reference is true, we will have a reason to believe that every instance of (δ) is true.

Above we discussed Field's worry that in order to establish that a (in this case, descriptivist) theory of reference is true, we'd need to show that the theory is true for many or all instances of

$$\begin{gathered}\forall y \exists x \exists z (y = \ulcorner F \urcorner \rightarrow ((x \text{ satisfies the cluster of descriptions} \\ \text{associated by speakers of this language with } y) \leftrightarrow (Fz)),\end{gathered} \tag{ζ}$$

which would involve an alternative method to arrive at a judgment about the right hand side of the biconditional, for many instances of (ζ). As we see now, this is not needed. Since the theories of reference imply (δ), *every* argument for the truth of a descriptivist theory of reference is an argument that every instance of (ζ) is true. Although we *can* go through (ζ) in order to argue for a descriptivist theory of reference, we need not.[16]

Accordingly, arguments from reference are *not* invalid. They aren't even enthymematic. Everything that is needed to draw the ontological conclusion is provided for in the first two steps. But then again, theories of reference

theories of the alchemists and Copernicus could be saved that way).

Similarly, in the debate here, intuitive ontological judgments that both participants in a debate could agree upon could well be used to discuss which theory of reference should be the more acceptable, without anyone begging any question.

[16] Why does this now seem all so trivial, when in the preceding paragraph we were convinced that it is such a highly non-trivial task to show that a definition of 'applies to' satisfies all instances of Criterion A? I think the answer is simply that the business considered above, namely that of giving a non-circular definition of 'applies to' is a totally different task from showing that a theory implies all instances of (δ). The latter allows us to analyze also the meta-language according to our theory of reference, while the former does not. We just had to show that the theories of reference alluded to in arguments from reference imply every instance of (δ), and this we have done.

are center stage, we just need to find out which is the true account to settle the dispute over eliminativism.

5 Bishop's Zoo

In a follow-up paper [Bis03], Michael Bishop has presented an argument, which, if sound, would establish that theories of reference, as they are used in arguments from reference, cannot imply all instances of (δ). Bishop suggested (in personal communication) that his argument from [Bis03] could accordingly be used to block my conclusion above with the following reasoning: Some arguments from reference are proposed by scientific realists to argue against the so-called "pessimistic induction". If the theories of reference used in these cases would imply all instances of (δ), realists would be committed to assume the existence of many entities, of which science tells us that they do not exist. To interpret the argument of realists in this way would be very uncharitable. Therefore we should rather interpret them to hold theories of reference which do not imply every instance of (δ). But then my argument above must break down: there are arguments from reference in which theories of reference play a role that do not imply the instances of (δ). Before entering into the details, three quick remarks:

1. In [Bis03] it is argued that the argument of the realist is invalid due to an equivocation if the realist's theory of reference is understood the way that Bishop suggests. I don't know in what sense it should be more charitable to interpret someone as making a mistake of equivocation than to interpret him as being committed to too many entities.

2. If we follow Bishop's suggestion, we should consider *all* arguments from reference as invalid because otherwise *some* of the arguments from reference in the realism/anti-realism debate would commit their proponents to too many entities.

   But to do so would certainly not be charitable. At best we should conclude from this (if we think that we can resolve (1.) in Bishop's favor) that the realists in the realism/anti-realism debate are holding theories of reference that do not imply all instances of (δ), while the arguments of reference in all other debates work in the way that I described it (and that it was just a mistake by Stich and Bishop to lump all of these together as "arguments from reference").

3. Bishop's argument in [Bis03] is obviously different from the argument in [BS98], thus it constitutes an attack against theories of reference in its own right. Therefore, even if remark (2.) establishes that Bishop's argument is irrelevant to what we have discussed so far (concentrating on the ontological conclusion), we should consider it in order to see

> what, if anything, is wrong with arguments from reference, even if what we will learn is only about some arguments from reference (perhaps those with non-ontological conclusions).

So let's see what Bishop's argument is all about.

5.1 The Pessimistic Induction

According to Bishop's reconstruction, the "Pessimistic Induction" is an inductive argument from an evidential base, taken from the history of science:

EB1 The important posits of many past successful theories did not exist.

EB2 Most of the central claims of many past successful theories were not true.

PI1 Most of the important posits of our best scientific theories do not exist.

PI2 Most of the central claims of our best scientific theories are not true.

The conclusions of this argument (PI1) and (PI2), directly contradict the central claims of many scientific realists, who believe that the success of our present scientific theory is evidence for their truth, and for the existence of their posits. Bishop believes that (EB1) is established by the anti-realist under the hypothetical assumption that our present theories are true, because only that way does the anti-realist have a basis to claim that the posits of the theories of the past do not exist. Assuming that our present theories are true, we can just list what they quantify over, and check whether the posits of the theories of the past are on that list. By this procedure, so the anti-realist argues, we will find many posits of the past not on the list and should then conclude that they do not exist [Bis03, 166–167].

(EB2) which is intended to express that the central claims of many past successful theories weren't even approximately true, should, according to Bishop, follow directly from (EB1): If the posits of a theory do not exist, all claims of the theory must be false, since the claims aren't about the world. Bishop invokes schema (B) to make that point:

> B If the expression "a" in a sentence "Fa" designates a posit that does not exist, then "Fa" is not approximately true.
>
> Schema [(B)] has a lot of intuitive appeal. To say that a sentence is approximately true is to say that it just missed something about the world. But if the subject of a sentence doesn't exist, then the sentence isn't about the world (in any sense that could give comfort to the realist). So it can't have just missed having been true about the world. [Bis03, 167][17]

[17] Apparently 100 years of 'On Denoting' weren't sufficient to rub this in, but $\ulcorner Fa \urcorner$ is not unproblematic as the logical form of a sentence of science, especially not of a sentence that says

But if many theories of past weren't even approximately true, although appeared to be very successful, the apparent success of our current theories should not delude us. It is likely that also our current theories do not refer to anything in the world and that therefore their claims are not (even approximately) true.

5.2 The Realist's Counter-Argument

In countering this argument, the realist now (to my mind correctly[18]) points out that such an argument presupposes a descriptivist theory of reference. Kitcher, for example, argues in [Kit93] that assuming a hybrid theory of reference would allow the realist to say that the scientists of the past on occasion said true things when using certain theoretical notions that were introduced in theories we now believe to be false. On these occasions, the scientists referred successfully to entities in the world when using (the now obsolete) theoretical terms. Kitcher's hybrid theory of reference would analyze the use of terms such as 'dephlogisticated air' on these occasions according to a causal-historical theory, such that the term on these occasions referred, in this example, to oxygen. But then, on these occasions, statements like

> Dephlogisticated air supports combustion better than ordinary air.

would have been about something that exists (namely oxygen) and also true of it. Thus, also past false theories wouldn't have kept the scientists from discovering and stating many true things also in the past when using theoretical terms, and (EB2) is undermined.

something "about" a theoretical posit. The present King of France rears his ugly bald head here. What follows from $\ulcorner Fa \urcorner$ being false? $\ulcorner \neg Fa \urcorner$ being true? Then dephlogisticated air does not support combustion better than ordinary air? Interesting. We can follow Bishop here and treat theoretical terms as names, but then we need to amend the term-introducing theory such that it says of each theoretical term that it names something. Lewis discusses this option in [Lew70].

[18] And also to the mind of Bishop when writing the paper with Stich, but Bishop apparently changed his. He now believes that the pessimistic induction is itself not an argument from reference, but an argument that could also succeed with a deflationist conception of reference. But, as we have seen, Bishop thinks that (EB1) is established by the anti-realist by checking for a given posit of a past theory whether it is among the things that our present theories posit. As Bishop emphasizes (the quote is provided below in 5.4) this is not done by checking whether a certain expression that was used in the past is also to be found in the vocabulary of present science, but whether the same things are posited by present theories that were posited in the past. I don't see how to do this without a substantive theory of reference. If you can't do this by checking whether the same expression is still there, it seems to me that you'd need to check whether what the expression would have denoted in the past is something that our present theories would also denote with their theoretical terms. To get his argument off the ground, the anti-realist should probably use a descriptivist theory here.

5.3 Bishop's Dilemma for the Realist

It seems that in order to establish that the referent of these statements exists, the realist has to assume that his theory of reference supports (δ). Otherwise, the claim that 'dephlogisticated air' referred on occasion could not support that 'Dephlogisticated air supports combustion better than ordinary air.' said something true about the world. But this, Bishop argues, the realist *should* not assume. The reason is that by (δ) it would follow that dephlogisticated air exists from the fact that 'dephlogisticated air' applies to something. But this would commit the realist in turn to the existence of dephlogisticated air. But the realist surely doesn't want to embrace such a "metaphysical zoo" in which every odd thing exists, of which current science tells us that it doesn't, including witches, races, phlogiston, etc.

> [R]ealists typically offer a substantive account of reference that yields the conclusion that the central expressions of suitably successful obsolete theories did, as a matter of fact, refer. [...] Suppose the realist is not willing to apply schema [(δ)]. Then the [anti-realist] wins the day. The realist's claims about [reference] do not touch on the real ontological issue separating realism and anti-realism. If the realist is willing to apply schema [(δ)], then the realist must embrace the metaphysical zoo; the realist must admit, not just that the *expressions* for the obsolete posits of successful theories refer, but that those posits actually *exist*. [Bis03, 196–170]

Hence the realist is in a dilemma. He either has to endorse a theory of reference that is stripped off (δ), which then loses its metaphysical bite, or he endorses a theory that includes (δ), but is then committed to the metaphysical zoo. Bishop's suggestion is to interpret the realist as committing the fallacy of equivocation in his counterargument against the anti-realist's pessimistic induction: The realist is in fact using a theory of reference that does not imply (δ) when claiming that the theoretical terms of past theories on occasion referred, but then uses a stronger notion of references when moving on to the conclusion that therefore the claims of scientists in the past were on occasion true. If that is true, then there is something wrong with arguments from reference, and my claim in the last section, that theories of reference as used in arguments from reference imply (δ), is false.

5.4 Bishop's Dilemma Resolved

Let's note first that, as I said above in my remarks already, the dialectical situation of the realist, as reconstructed by Bishop, is obviously different from the dialectical situation of, for example, Lycan in his counter-argument to eliminativism. Lycan is happy to be commited to the existence of beliefs, in fact that is what he wants to say. So at least for arguments from reference with ontological conclusions there is no reason to assume that the theories of reference that play a role there do (or should) not imply (δ), even if Bishop were right about the realism/anti-realism debate.

Let's note then that the (classical) causal-historical theory says that natural-kind terms are exactly these terms that refer (via a causal-historical chain) to a natural kind. In case of 'dephlogisticated air' the term does refer to one, namely oxygen. This would even be found if we'd use Bishop's recommended method to see whether a theoretical posit in the past referred to something we now believe exists. As Bishop remarks about the process by which we check this (for the example of 'ether'):

> [T]his procedure is not equivalent to one that simply checks whether our current theories retain the *word* "ether". The procedure aims to discover whether our current theories posit the ether, regardless what it might be called. [Bis03, 165]

Thus being committed to the existence of dephlogisticated air is no embarrassment for the realist, it commits him to the existence of oxygen. Terms that potentially cause an embarrassment, like 'race', 'witch', 'phlogiston', on the other hand, aren't natural kind terms, according to the standard causal-historical theory (plus our current scientific theories), and do not refer. One might worry whether the inherent essentialism of the causal-historical theory is very attractive, and there are externalist theories of reference that do not assume the existence of natural kinds. But in the context of any externalist theory, the challenge[19] will be to provide some external fact that determines what referent is picked out by the use of a term. If a term refers, then it does so in virtue of that external fact providing something in the world for it to denote. Therefore, a theory of reference that implies all instances of (δ), will not lead to ontological embarrassment. It only does so if we change theories of reference between evaluating the left and the right side of (δ). If you take the right hand side of (δ) to be saying that something is standing in a causal-historical chain to the (relevant historical) use of 'dephlogisticated air', and the left hand side to be saying that there is something that satisfies the cluster of descriptions that speakers of the language at that time associated with the term 'dephlogisticated air', endorsing the right hand side brings an unwelcome ontological commitment on the left hand side. But current theories of reference aren't schizophrenic like that.[20] That is still

[19] I'm not claiming that every argument from reference in the history of philosophy is a good one. Perhaps some are bad, and some are perhaps bad, because the theory of reference used in them is crap. Perhaps so, because it doesn't properly meet this challenge.

[20] There are, as we said, hybrid theories that allow for expressions to be used in some utterances in accordance with a causal-historical theory and in other utterances in accordance with a descriptivist theory. As we noted when introducing schema (δ) for the very first time, both occurrences of 't' in the schema need to be substituted with expressions suitably disambiguated

the same old confusion we encountered before in Stich's arguments.

Bishop has two objections to this proposal. The first is that it leads to an "Orwellian view" about scientific language. Scientists seem to be engaging in disputes about the existence of dephlogisticated air, light waves, and light particles, while the clever philosohper knows it all better and can tell them that all their debates don't really make sense, and that they do not really know what they are talking about when using terms like 'dephlogisticated air', 'light wave', 'light particle', etc.:

> This view of scientific language seems to violate the weakest principle of interpretive charity because it makes scientists so deeply ignorant about the content of their own theories. [...] [T]he possible world they would confidently identify as the one that contained their posits is not the possible world that contains their posits. Further, this view forces us to take some of the most interesting disputes in the history of science about what exists and radically interpret those disputes so that the participants are (unwittingly) positing the same entities but just disagreeing about what those entities are like. [...] [T]his account of scientific language leads to bad history. [Bis03, 170]

The second objection is that this Orwellian semantics saves a version of realism that isn't worth saving. It might perhaps establish that our best scientific theories are true, but to the prize that we don't know anymore what our theories say. We might have a reason to believe that our theories are true, but we might not understand them any better than Maxwell and Priestley understood theirs. Bishop calls this the "We Don't Know What We're Talking About Realism" [Bis03, 171].

Let's answer the second objection first. If our theories are true, then not only do our theoretical terms refer to something in the world, but also everything that we believe about these things would be true. And this would be so on what Bishop calls "Orwellian semantics". Thus if the realist could save that much realism, he should be very happy. The situation would be way different from the situation of Maxwell and Priestley who only on occasion said something true about things in the world they were otherwise pretty mistaken about. If our theories are true then, on this picture of scientific language, there isn't much left that we could be wrong or ignorant about. The second objection is just a red herring.

The first objection seems to overlook two things. The first is that externalist theories of reference, as the causal-historical part of Kitcher's hybrid

to avoid confusion. Thus, if we use 'dephlogisticated air' in the right hand side in a way that it denotes oxygen, such that it is true of 'dephlogisticated air' that it denotes something, then we will also use it on the left hand side in this way, and vice versa. Bishop, I'm afraid, overlooked this.

theory, just have it that it is an *external* fact what our terms refer to and that this is not determined by what we believe about what the referents of our terms are like. That is why they are called “externalist”. One can find a proposal like this outrageous, but in this case one should engage in philosophy of language and the discussions about semantic externalism. Externalism simply implies that (under certain circumstances) we might not know what we are talking about in the sense that, for example, were it the case that our rivers and lakes are actually (and as yet unbeknownst to us) filled with XYZ, we would be referring to XYZ with ‘water’ although we firmly believe that we refer to H_2O. In this way we would take an H_2O-world to be the world we are talking about when we say that lakes and rivers are filled with water, although, in fact, we are then talking about another world.

The second thing that this objection overlooks is that none of the disputes in the history of science would turn out to be pointless when we’d analyze them even in the most straightforward causal-historical fashion. Take Bishop’s example of the dispute between wave-theorists and particle-theorists. Both theorists on occasion said true things about electromagnetic radiation when making claims in their theoretical vocabulary. Bishop thinks that we could not describe this dispute as a dispute about what entity makes up light, since the Orwellian semanticist would have to describe the situation such that both in fact posited the same, viz. electromagnetic radiation.

But that is clearly not so. It is rather that if we’d assume in a non-Orwellian, descriptivist semantics that the referent of a theoretical term must be fully determined by what the speaker using that term believes about it, that wave and particle theorists were both theorizing about some non-existent stuff, but not disagreeing about light (since light is electromagnetic radiation). I think that *that* would be bad history. Of course, both theories make different theoretical assumptions and genuinely disagree about the nature of light. Thus taking ‘light wave’ and ‘light particle’ both to be referring to what actually is electromagnetic radiation does not make postulating waves the same as postulating particles. Moreover, on Kitcher’s hybrid theory it is totally unproblematic to make sense of both, ontological disagreement between both parties, and occasional co-reference of the theoretical terms of the disparate theories. But again, this just turns on questions about the true theory of reference, and has nothing to do with arguments from reference or the realism/anti-realism debate. Hence also Bishop’s first objection is a red herring.

6 First Summary

Let me briefly summarize the observations Stich made with respect to arguments from reference and what we have said about these. We will take the conclusions we have reached about Stich's worries as a starting point in the discussion of his second attack on arguments from reference.

6.1 Observation I: Two Research Programs

According to Stich, the attempt to decide the soundness of arguments from reference by investigating which theory of reference is true, faces the problem that there is no clear conception within philosophy what the true account is supposed to be true of. One possible conception is that a theory of reference is simply an account of "folk semantics". Theories of reference are reconstructions of tacit theories we have and that we subconsciously apply when making intuitive judgments about the denotation of terms in actual and hypothetical cases. The better a so reconstructed theory agrees with our intuitive judgments about denotation, the closer it is to the "true account" of reference.

The second possible conception, which Stich calls "the proto-scientific project", is that our folk semantics is — just as our "folk physics" — not more than a reconstruction of our pre-scientific understanding of reference, that could, in principle, be massively mistaken about the true account of reference. In this case, other considerations might matter more than just agreement with out pre-theoretic intuitions, including, for example, considerations about how fruitful a theory of reference proves to be when combined with other empirical theories (in, for example, the history of science, linguistics, cognitive psychology and evolutionary biology).

6.2 Observation II: The Shaky Empirical Basis

Trying to establish a result in "folk semantics" Stich collected a sample of intuitive judgments on probes. Stich reports that there was a large amount of variation in his pilot study. This led Stich to the conclusion that either folk intuitions do not determine a folk theory of reference, in which case the denotation of a term would be indeterminate and so would be the truth of sentences in which the term is used to make claims about existence (or non-existence). Or, if folk intuitions determine after all one folk theory of reference, then this theory will be rather idiosyncratic. Since *that* the folk semantic theory is the idiosyncratic way it is will be due to contingent historical and social facts, the corresponding ontological questions will have uninteresting answers.

Stich takes the "implausible consequences" of this result (which contain a general scepticism about ontology) to constitute a *reductio* of the premises that led him to this conclusion.

6.3 Observation III: The Alleged Invalidity of Arguments from Reference

As a common premise of both, the reasoning that led Stich to the "implausible consequences" and the ontological conclusions in arguments from reference, Stich identifies instances of the following schema

$$t \text{ exists if and only if } \ulcorner t \urcorner \text{ denotes something.} \qquad (\delta)$$

Stich and Bishop argue that although this premise is a necessary ingredient in all arguments from reference — since otherwise ontological conclusions would not follow from observations about the denotation or denotationlessness of terms — it is never argued for when an argument from reference is made. Stich and Bishop also believe that it would be difficult to argue for that premise from the point of view of a substantive theory of reference.[21]

As I have shown now, it is not difficult to argue for that premise. At least for the substantive theories of reference that feature in arguments from reference (descriptive theory and causal-historical theory), all instances of (δ) are implied by the theory. Thus, if one has any reason to believe that such a theory of reference is true, one has reason to believe that all instances of (δ) are true, and reason for the former is provided in the very first step of an argument from reference, according to Stich's and Bishop's own reconstruction. Thus so far we have not been given a reason to distrust arguments from reference *per se.* Let us move on to the second attack against arguments from reference.

7 The Second Attack: The Argument from Social Psychology

Based on the empirical results they reported and interpreted in [MMNS04], Edouard Machery, Ron Mallon, Shaun Nichols and Stephen Stich start another attack on arguments from reference in [MMNS]. This time they argue from an empirical result, which suggests strong variation in the intuitions used to find the correct theory of reference, to the unfeasibility of the project of finding a correct (substantive) theory of reference in general. If, however, all substantive accounts of reference are doomed, so are all arguments from reference.

[21] Stich does not doubt that all instances of (δ) *are* true. Since he is a deflationist about reference, he believes that the instances of (δ) are all there is to be known about reference. What he doubts is that someone who holds a substantive (non-deflationist) theory of reference can show that his theory makes all instances of (δ) true.

Although they keep the original characterization of arguments from reference explained above (cf. [MMNS, ??]), the new argument does not depend on the former argument against theories from reference (cf. [MMNS, ??, FN 2]) developed in [BS98]. As we will see, the new argument is even incompatible with the old argument.[22]

7.1 The Empirical Basis

In the study which is the empirical basis for the Argument from Social Psychology, Edouard Machery, Ron Mallon, Shaun Nichols and Stephen Stich (in what follows I refer to this group with 'Stich et al.') intended to test whether there is a significant difference between members of Asian and Western cultures in these intuitions that guide the academic discussion between proponents of a descriptivist theory of the reference of proper names and a causal-historical theory. How the arrived at the initial hypothesis that there should be such a difference needs a little explaining.

7.1.1 The Hypothesis

As Stich et al. reconstruct it, analytic philosophy of language is dominated by two basic views on the reference of proper names:

> Two theses are common to all descriptivist accounts of the reference of proper names:
>
> D1. Competent speakers associate a *description* with every proper name. This description specifies a set of properties.
>
> D2. An object is the referent of a proper name if and only if it *uniquely or best satisfies* the description associated with it. An object uniquely satisfies a description when the description is true of it and only it. If no object entirely satisfies the description, [...][23] the proper name refers to the unique individual that satisfies most of the description [...]. If the description is not satisfied at all or if many individuals satisfy it, the name does not refer.
>
> The causal-historical view offers a strikingly different picture [...]:

[22] This, in itself, would not be a problem. After all, Stich could pose a dilemma for all arguments from reference by stitching the two arguments together. As we will see, what is more problematic is that the ease with which they assume in the new argument that all instances of (δ) are implied by the theories discussed, suggests that the first argument is just hopeless.

[23] The deleted part reads "many philosophers claim that". But I assume that it was a mistake of Stich et al. to present the "theses common to all descriptivist accounts" as a claim about what many philosophers claim. Accounts of the reference of proper names should rather claim something about reference, which is why I guess that this deletion is in line with the intentions of Stich et al. Cf. [MMNS04].

C1. A name is introduced into a linguistic community for the purpose of referring to an individual. It continues to refer to that individual as long as its uses are linked to the individual *via a causal chain* of successive users: every user of the name acquired it from another user, who acquired it in turn from someone else, and so on, up to the first user who introduced the name to refer to a specific individual.

C2. Speakers may associate descriptions with names. After a name is introduced, the associated description *does not play any role* in the fixation of the referent. The referent may entirely fail to satisfy the description.

[MMNS04, B2–B3]

Stich et al. note that the causal-historical view became the dominant view among philosophers after Saul Kripke presented a number of thought experiments, the intuitive response to which was (among philosophers) predominantly in favor of a causal-historical description of the cases presented in these thought experiments. The thought experiments they are having in mind include "the Gödel case" and the "Jonah case".

The Gödel case is the fictional story of a person who associates a description (viz. "the man who discovered the incompleteness of arithmetic") with the name Gödel that encompasses everything he believes about what he takes to be the original bearer of that name. In fact, however, this description is, for historically contingent circumstances, not true of person a who is the original bearer of the name, but in fact true of a person b whose original name is 'Schmidt'. On the descriptivist theory the person now using the name 'Gödel' is referring to b, since b is the unique satisfier of the description that is now associated with the name. On the causal-historical account the person is instead referring to a, since a is the original bearer of the name and the present usage of the name 'Gödel' is historically and causally linked to this original bearer a.

The Jonah case is the fictional story in which it turns out that no actual person satisfies the description presently associated with the biblical name 'Jonah' (which would presumably include that he was the prophet swallowed by a whale). On the descriptivist account it *follows* that Jonah did not exist. On the causal-historical account this does not follow. It *could* be that the biblical description that is presently associated with the name is just a legendary story about a real person and that we are referring to this person when using the name 'Jonah'. Thus the causal-historical account does not exclude the latter as a logical possibility.

In both cases most philosophers seem to have the intuition that the causal-historical account provides the correct description. Stich et al. suspect that philosophers assume thereby that their intuitions are universally shared.[24] However, as social psychology has revealed in some recent studies, such universality assumptions should not be made a priori. It might be that the intuitive responses to probes modeled on philosophical thought experiments vary between different social (ethnic, cultural, or socioeconomic) groups (cf. [WNS01]).

In particular, studies in social psychology seem to suggest that people with an East Asian cultural background differ from people with a Western cultural background in the way they recognize causal relations as relevant or salient in the cognitive tasks posed by thought experiments, such as classification of cases as cases of a certain kind. It seems that people with a Western background are more responsive to the causal relations that obtain in and between such cases than people with an East Asian background. But since the relevant difference between a descriptivist and a *causal*-historical account involves the role of causal relations, one might suspect that such differences also show up with respect to the Kripkean thought experiments:

> The cross-cultural work indicates that [East Asians] are more inclined than [Westerners] to make categorical judgments on the basis of similarity; [Westerners], on the other hand, are more diposed to focus on causation in describing the world and classifying things [...]. This differential focus led us to hypothesize that there might be a related cross-cultural difference in semantic intuitions. On a description theory, the referent has to satisfy the description, but it need not be causally related to the use of the term. In contrast, on Kripke's causal-historical theory, the referent need not satisfy the associated description. Rather, it need only figure in the causal history (and in the causal explanation) of the speaker's current use of the word. [MMNS04, B5]

With that reasoning, Stich et al. arrived at the following hypothesis:

Hypothesis. *When presented with Kripke-style thought experiments, Westerners are more likely to respond in accordance with causal-historical ac-*

[24] Some philosophers indeed do so. Frank Jackson is a famous example:

> I am sometimes asked in a tone that suggests that the question is a major objection why, if conceptual analysis is concerned to elucidate what governs our classificatory practice, don't I advocate doing serious opinion polls on people's responses to various cases? My answer is that I do when it is necessary. Everyone who presents the Gettier case to a class of students is doing their own bit of fieldwork, and we all know the answer they get in the vast majority of cases. But it is also true that often we know that our own case is typical and so can generalize from it to others. It was surely not a surprise to Gettier that so many people agreed about his cases. [Jac98, 36–37]

counts of reference, while East Asians are more likely to respond in accordance with descriptivist accounts of reference.

7.1.2 The Experiment

The hypothesis was tested by conducting an experiment with Western undergraduate students from Rutgers University (31 participants) and English speaking Chinese undergraduates from the University of Hong Kong (41 participants). In a classroom setting, both groups were presented with the same four probes. Two of these were modeled on the Gödel case, two on the Jonah case. In both stories a person was using a proper name with an associated description which, in the Gödel case, was satisfied by a person that was not the original bearer of the name, and in the Jonah case not satisfied by any person including the original bearer of the name. Participants were then asked what the persons in the story using these names were "talking about", the original bearer of the name, or rather the actual or (non-existing) fictional person that satisfied the description.

Each question was then scored binomially, such than an answer that would accord with the Kripkean, causal-historical description of the case was given a score of 1 and the answer more in accord with a descriptivist description of the case was given a 0. Then the scores of each participants were summed. Thus for each case, the cummulative score coul range from 0 to 2.

In Table 1 the means and the standard deviation (SD) for summary scores are listed. When testing the significance of the results, the differences between Westerners and East Asians for the Gödel test proved significant[25], while the result in the Jonah case was reversed (!) but not significantly so.[26]

	Score (SD)
Gödel cases	
Western participants	1.13 (0.88)
Chinese participants	0.63 (0.84)
Jonah cases	
Western participants	1.23 (0.96)
Chinese participants	1.32 (0.76)

Table 1: Mean scores for experiment, standard deviation in parentheses

Thus the hypothesis was confirmed for the probes modeled on the Gödel case: Westerners were more likely then East Asians to give responses in

[25] On an independent sample t-test, $t(70) = -2.55, P < 0.05$.

[26] On an independent sample t-test, $t(69) = 0,486$, n.s.

accordance with the causal-historiucal account of reference. The responses to the Jonah case, however, were not in accordance with the prediction.

In [MMNS] they also report a "significant" intra-cultural difference:

> While for each vignette a majority of Americans gave causal-historical responses, in each case a sizable minority of the population (as high as 45% in one case) gave descriptivist responses. Similarly for the Chinese population, for each vignette, a majority of Chinese participants gave descriptivist responses, but in each case a sizable minority (in some cases[27] over 30%) gave causal historical-responses. [MMNS, ??]

Unfortunately, Stich et al. didn't report the actual results (which wouldn't have been a very problematic thing to do with four cases and two groups). Nor do they tell us in what sense the result was "significant" here. Thus the "significance" of the intra-cultural difference and how one should interpret it, is not possible to assess for us.

7.1.3 The First Interpretation of the Result

In their first interpretation of these results, Stich et al. draw the obvious conclusion: if these result obtained for the Gödel case should prove to be stable in subsequent tests, philosophers should stop to assume a priori that their intuitions are universally shared.

When considering the possible reply that philosophers might not be interested in lay intuitions (and thus the findings of no real significance for philosophical methodology), since philosophers are following (what [Sti96] called) the proto-scientific project, for which only (or predominantly) expert opinion counts, Stich et al. respond:

> We find it *wildly* implausible that the semantic intuitions of the narrow cross-section of humanity who are Western academic philosophers are a more reliable indicator of the correct theory

[27] "Some cases"? Since there were only two Gödel cases tested (and this apparently is only about the Gödel cases, since there was no majority of Chinese participants giving descriptivist responses to the Jonah case, right?), shouldn't it really be "all cases", or is the plural just misleading the reader here? I suspect it is the latter, since the paragraph begins already with "In two separate studies using four different vignettes, we found that Americans were more likely than Chinese to give causal-historical responses." — as we know from [MMNS04], they used only *two* vignettes on the Gödel case, and they don't mention (!) the Jonah case in [MMNS] (let alone, of course, that it disconfirmed the hypothesis).

In other words: from [MMNS] the reader gets the impression that they tested exactly four probes, all of which confirmed the hypothesis, while in fact two of the four *disconfirmed* it. This is not merely a somewhat tendentious report of the actual result, but very misleading. As we will see, for the argument in [MMNS] it is somewhat crucial to downplay the negative result in the Jonah case.

> of reference [...] than the differing semantic intuitions of other cultural or linguistic groups. [...] In the absence of a principled argument why philosophers' intuitions are superior, this project smacks of narcissism in the extreme. [MMNS04, B9]

But in [MMNS04] the suggestion is that the proto-scientic project, according to which there is an intuition-independent true account of reference, should, anyway, rather *not* be the project to be followed in philosophy. What philosophers *should* follow is (what [Sti96] called) the folk-semantic program:

> A more charitable interpretation of the work of philosophers of language is that it is a proto-scientific project modeled on the Chomskyan tradition in linguistics. Such a project would employ intuitions about reference to develop an empirically adequate account of the implicit theory that underlies ordinary uses of names. If this is the correct interpretation of the philosophical interest in the theory of reference, then our data are especially surprising, for there is little hint in philosophical discussion that names might work in differen ways in different dialects of the same language or in different cultural groups who speak the same language. [MMNS04, B9]

Besides the unfortunate use of "proto-scientific" when characterizing the project which was in the earlier publications the official opposition to the "proto-scientific" program, Stich et al. suggest here that philosophers take semantic intuitions as immediate data, to be explained by the true account of reference, the reconstruction of the implicit theory that underlies the ordinary use of names. Now, if that data varies with different groups, the suggestion should be that there must be more than just one true account of reference, right? Accordingly, there should be an account of reference for intuition group A, one for group B, etc. At least that seems to be the conclusion of Stich et al. As we will see, they recently changed their minds about this.

7.2 The Argument from Social Psychology

In [MMNS] the same group of authors takes the empirical result just reported as a "strong prima facie case that intuitions about reference used to construct theories of reference might vary from culture to culture" [MMNS, ??]. Taking this as a starting premise they then consider what consequences follow for arguments from reference, as characterized above. In particular, is there a way that a theorist of reference can accommodate this empirical result, and does this way to accommodate the result still support an argument from reference? Their argument seems to be that there are only three possible strategies to accommodate these empirical results, the first of which leads directly to abandoning arguments from reference while the other two are just hopeless.

7.2.1 Strategy I: Deflationism about Reference

The first strategy, which Stich recommends already since [Sti83] is simply to abandon substantive theories of reference and settle for a deflationist theory. Stich assumes that also Field's deflationist theory of reference is motivated by similar considerations:

> This might be an appropriate place to say a bit more on how I interpret Field's current view about reference. [T]here are passages in [[Fie94]] which suggest that Field, like Rorty, denies that ordinary folk have any intuitions about reference. However, Field tells me that this is not what he intended in those passages. Rather, he is inclined to think that people generally do have a tacit internalized theory of reference, but that this theory is really quite minimal. It contains little more than the semantic ascent (or disquotation) schema and also, perhaps, some information on how to use it. This internally represented schema underlies lots of intuitions about reference. But the intuitions in question are the "trivial" ones — i.e. the ones that are just instances of the schema [(δ)]. [Sti96, 87, FN 47]

The cultural variation when it comes to the Kripkean thought experiments would then presumably be caused by considerations that concern something else, other than reference. Instead of analyzing the reference relation in terms of any complex word-world relation, the theory of reference is thus nothing but a collection of platitudes, like all instances of

$$t \text{ exists if and only if } \ulcorner t \urcorner \text{ denotes something.} \tag{δ}$$

which are, perhaps, universally accepted intuitive truths.[28] Thus a deflationist theory could accommodate the variation in intuitions, simply because it would not consider these intuitions as intuitions about reference proper. But, on the other hand, since the right hand side of these biconditionals is left unanalyzed in Stich's deflationism, these biconditionals can not serve to support an argument from reference.

In a paradigmatic argument from reference we first need to establish *independently* some claim about the denotation of a given term. With a substantive theory of reference, like a descriptive theory, this is possible, because the theory tells us that theoretical terms of massively mistaken

[28] This claim would, of course, also be in need of empirical support. The empirical results reported above do not support it.

theories do not denote anything. That a term is a theoretical term of a certain theory as well as that that theory is massively mistaken can both be established independently of the ontological question at issue. True, a theory would also be massively mistaken if it gets the whole ontology wrong, but an independent way to establish that a theory is massively mistaken could simply consist in referring to its impressively bad empirical track record.[29] After that is established, the descriptivist theory of reference allows us to conclude that the theoretical terms of that theory do not denote anything, which in turn allows us to conclude with (δ) that things by that name do not exist.

If, however, our theory of reference is nothing but a list of all instances of (δ), the only way to establish that a term does (or does not) denote, is via the left hand side of the biconditional of the relevant instance of (δ). But then there is no independent way to show that a term does (or does not) denote that would not beg the ontological question at issue. Thus, a deflationist theory of reference could perhaps accommodate the empirical result, but could not support arguments from reference.

7.2.2 Strategy II: Changing the Methodology

The second possibility to accommodate these results would also keep to the idea that there is only one true account of reference, but that this account is perhaps underdetermined by intuitions alone or even largely independent of intuitions or at least independent of unschooled folk intuitions. In order to pick the true account of reference, other considerations need to come into play then. One such consideration could be how well a given account of reference supports the ontological convictions one has. For example, if one is an anti-eliminativist, one could have a reason to believe that the causal-historical account is right, because it is compatible with the falsity of folk psychology and the existence of beliefs and desires. Considerations like these would however not give *independent* support for theories of reference in order to use them for arguments from reference. Thus, if your only reason to believe that the causal-historical account is right is that it supports anti-eliminativism, you can't, or so Stich et al. seem to think, use that account in turn to argue for anti-eliminativism. But even if you could get away with it, arguments in which the theory of reference is not independently motivated, Stich et al. decide not to consider as arguments from reference proper.

But what could such an independent reason be like? Stich et al. apparently cannot think of anything here. Accordingly, they challenge anyone who wants to argue that there are other considerations that might serve as

[29] Or, as [Chu81] argued in case of folk psychology, in referring to the fact that the theory didn't make any progress during the last twothousand years, that it isn't compatible with and clearly not reducible to our best theories of the ontologically (more) fundamental level, etc.

reasons for or against theories of reference, to tell what these are and why they should be considered reasons. Whatever story one will come up with here, does seem to be prima facie problematic. Since Stich et al. don't know of any considerations other than intuitions or philosophical convictions that might speak for or against a theory of reference and which could play a role in philosophy, every such story must be at odds with the philosophical practice, "the dominant tradition of employing the method of cases in the philosophy of language" [MMNS, ??].

7.2.3 Strategy III: Reference Relativism

The third possibility to accommodate the empirical results discusses what seemed to be the recommended accommodation of the results in the conclusion of [MMNS04]. As we will see, Stich et al. do not think (anymore?) that this accommodation can work either.

As we remember, the "folk semantic" research program assumed that folk intuitions determine the correct account of reference, in the way that intuitions about grammaticality determine the right account of grammar. The theory that is compatible with all these intuitions, is just the true theory of reference.[30] In other words: folk intuitions about reference cannot be mistaken about reference, in the way that folk intuitions about physical facts can be mistaken about the physical facts.

The folk semantic research program was seen to assume that all human beings (or perhaps all language users, including extraterrestrials) would share the same intuitions about actual and possible cases when it comes to reference. Learning now, that this assumption is perhaps false, need not directly lead to abandoning the whole research program. One might hold on to the idea that the folk intuitions determine reference, but accommodate the empirical result simply by assuming that the totality of language users can be divided into groups that have different implicit theories of reference and consequently different intuitions about reference. A philosopher of language who allows for such a possibility is a "referential pluralist" and committed to the following:[31]

The Pluralist Method of Cases. *The correct theory of reference for a class of terms* T *employed by members of intuition group* G *is the theory which is best supported by the intuitions that competent members of* G *have about the reference of members of* T *across actual and possible cases.*

Stich et al. now believe that this view is highly implausible. The first problem that they note is that the analogy with grammaticality intuitions

[30] Notwithstanding the typical problems of underdetermination, not at all peculiar to theories of reference.

[31] Cf. [MMNS, ??]

is not perfect. In case of grammaticality intuitions, an empirical result that would show that speakers of what seems to be the same dialect have different intuitions about the grammaticality of sentences would undermine the assumption that "intuitions about grammaticality provide reliable evidence grammatical properties of the dialect they speak" [MMNS, ??]. But that is exactly what seems to be established by the empirical results here. Thus a similar conclusion should follow about the reliability of intuitions about reference:

> Faced with this variation, it is very tempting to abandon the assumption that intuitions about reference provide evidence about reference alltogether. Instead, one might, for example, propose that a speaker's intuitions about reference are caused by a variety of factors that turn out to have nothing to do with reference, including her culture and perhaps her philosophical commitments [...]. But referential pluralism is committed to the method of cases, and so must make this assumption, that is in dire need of justification. [MMNS, ??]

The second problem that they spot with this strategy concerns the consequences such a view would have for arguments from reference. As we had said, arguments from reference assume a substantive theory of reference as the true theory of reference and argue then for ontological conclusions, like

$$\text{Beliefs exist.} \tag{γ}$$

But assume that we have two distinct intuition groups A and B and, as it happens, for group A descriptivism is the true theory and for group B the causal-historical theory is the true theory. Then the argument for eliminativism could be sound for members of group A, while the argument for anti-eliminativism could be sound for members of group B. Thus members of B could happiliy infer (γ) and members of group A happily infer

$$\text{Beliefs do not exist.} \tag{γ^-}$$

In order to avoid contradiction the referential pluralist is now forced to relativize the truth of the utterances of (γ) and (γ^-) to the intuition groups of the respective speakers. Thus an utterance of (γ^-) is true if uttered by a member of A but false if uttered by a member of B. If there is inter-group conflict about the truth of (γ), the conflict is only apparent. In fact they do not disagree but just talk past each other.

Now, since Stich et al. assume also intra-cultural variation including Western philosophers, the relativization of ontological claims must also cover these:

> In philosophy, this means that arguments over the existence of beliefs (or the existence of races, the progress of science, the nature of our epistemic access to moral properties and so on for the conclusions for every other argument from reference) have to be relativized. [MMNS, ...]

If all is correct what Stich et al. are inferring so far, then the conclusion here is not confined to conclusions of arguments from reference, but at least true for every utterance of an existence claim. Because of (δ) being true for all theories of reference considered here[32], these utterances will differ in meaning if the speakers belongs to different intuition groups, regardless of whether they utter these sentences as conclusions of an argument from reference or under other circumstances. And philosophers would thus agree or disagree about such existence claims only if they belonged to the same intuition group.

> We take it that these conclusions are very surprising, and would involve a very substantial revision of philosophical methodology. For, they suggest that philosophical disagreement and agreement among even speakers of the same language, who belong to the same culture, have the same socio-economic status, and even attended the same graduate program in philosophy, may be illusory if the speakers have different intuitions about how terms refer in actual and fictional cases. [MMNS, ??]

But that is not what Stich et al. consider the most absurd consequence of referential pluralism. What they think is more absurd is that unless we have tested the referential intuitions of everyone and classified him or her into the right intuition group with its peculiar theory of reference, we will not even know whether or not two philosophers are disagreeing over, for example, the truth of (γ^-), or just talk past each other (because they belong to different groups). And it also seems that nobody even knows him- or herself to which group he or she belongs, since nobody has yet confronted him- or herself with all actual and possible cases. Isn't that truly absurd? That we might be talking past each other and not know it?

> Together, these considerations strongly support the view that we simply do not know to which intuition group any of us belongs. And that completes our reductio, for since it is unclear which intuition group each of us belongs to, and because we may well belong to different groups [...], referential pluralism leads to the absurd conclusion that no one knows when proponents of arguments from reference agree, when they disagree, and when they speak at cross-purposes. [MMNS, ??]

[32] Cf. my discussion of this in section 4

This conclusion Stich et al. finally find so bizarre that they reject referential pluralism and thus the third way to accommodate the empirical results.

The upshot of their argument is that philosophers must abandon arguments from reference in order not to be in conflict what they believe is likely to be an empirical truth, namely that there is variation in intuitions about reference.

8 The Argument from Social Psychology Refuted

I think there is an impressive amount of confusion underlying this whole argumentation, so let's uncover this step by step. Perhaps we should begin with a few remarks about the empirical basis.

8.1 The Irrelevance of the Empirical Data

First of all, the result that they obtained in [MMNS04] is strictly irrelevant for arguments from reference. The empirical research was limited to the denotation of *proper names*, whereas the denotation of proper names plays no role in any argument from reference. There we are typically dealing with theoretical terms, kind terms in particular, and this simply wasn't tested. There are, on the other hand, good reasons to assume that this is of relevance. Although some theories of reference treat proper names and predicates in similar ways, there might be culture-dependent reasons for differences. For example, it might well be that in one culture names have a descriptive significance that names in other cultures lack (the proper name 'Gottlob' might not carry to the Western ear any more descriptive significance than the proper name 'Heinz-Günter' does, but the proper name 'Clever warrior who fights the enemy with perfidy' might well carry descriptive significance to members of other cultures), while for kind terms there is perhaps no such cultural difference.

Second of all, the experiment at best established that there is intercultural variation when it comes to the Gödel case, but the cases of relevance for the arguments from reference are clearly cases of the Jonah-type, for which the experiment did not show any significant variation between the two groups. Remember that in arguments from reference (of either sort), ontological conclusions are to be inferred. As in the argument for and against eliminativism, the relevant question was whether a term could refer at all, although the theory that the term was introduced in, was massively mistaken. This has some resemblance to the Jonah case, where *no* actual being satisfies the description commonly associated with a name, but no resemblance to the Gödel case, where there actually is another candidate that satisfies the associated description. Thus, at best, the variation in the Gödel case can be taken as opening the logical possibility that there could be a significant variation in semantic intuitions of the relevant kind, but it clearly didn't establish any such variation, nor did it make such a variation likely.

Thirdly, the results in no way suggest that variation in semantic intuition is such a fine-grained and wide-spread phenomenon as Stich et al. make it sound like, when saying that speakers of the same language, who belong to the same culture, have the same socio-economic status, and even attended the same graduate program in philosophy might have different intuitions about reference and then go on to argue what absurd consequences would follow for certain views from this. Well, members of these groups "might" have different intuitions about reference, and this "might" then be very surprising to learn, but that has not been shown yet. In fact, several experimental philosophers, including the Stich group (also in [MMNS04]) report and emphasize that the intuitions of, for example, philosophers who attended the same graduate program in philosophy *converge* significantly. But actually for neither of these claims has experimental philosophy delivered any empirical support.

Summarizing this, there is so far absolutely no reason to believe that there is a significant empirical variation in the intuitions relevant for arguments from reference. If we want to take this whole argument seriously at all, we thus should discuss it as departing from the logical possibility of relevant variation in semantic intuitions. Could such a variation, if it was found at all, undermine arguments from reference?

8.2 The Fundamental Confusion Cleared Up

In order to assess whether deflationism would be the only working accommodation of the results, if we assume for the sake of the argument that they found significant and relevant variation in intuition, we need to reconsider one of the older observations by Stich, the question which research program is actually carried out by philosophers of language when they try to find the true account of reference.

We remember that Stich distinguished two possible research programs, the "folk semantic program" and the "proto-scientific program". About the former we have just spoken. The latter program assumed that there either is an objective truth about reference, independent of the intuitions of the folk, or at least one conception of reference that proves best for empirical science, again largely independent of intuitions of the folk. Stich does not distinguish clearly between the last two options. But it seems possible to do so.

The first option in the proto-scientific program would assume that how words relate to the world is just another fact of nature, which is the way it is, independent of what we (intuitively or explicitly) believe about it. We could call this the "external fact account of reference". In this case our folk intuitions about reference can be true or false, and a reliable or unreliable guide to reference, in relation to how well they accord with the true word-world relation, which is, as we said, independent of these. So far, as Stich correctly points out, the natural sciences do not seem to have a branch that would investigate what that relation is. Accordingly, we so far don't have any good clue how reliable our intuitions are.

The second option in the proto-scientific program would assume that how words relate to the world is a matter that is for us under intentional control. We can, especially in using technical languages, to a large degree influence in what way *our* words relate to the world, and can to a large degree chose how we want to understand terms such as 'reference' and 'exists', etc. when reporting or stating facts about the denotation of terms used by others. For different scientific purposes (linguistics, psychology, history of science), different choices might seem to be pragmatically right. Choosing a reference relation would thus be like choosing a language or linguistic framework for scientific purposes in Rudolf Carnap's sense. The intuitions of the folk would be largely uninteresting for this choice. Perhaps it would be good not to choose a relation too alien to the folk intuitions, since that would make life much easier for the scientists, but that is already all that can be said about the folk intuitions. However, *this* proto-scientific project does not assume that there is any external fact about reference that should be captured. Perhaps there is one reference relation uniquely determined by the folk intuitions and at the same time pragmatically best for all sciences, but that would be sheer coincidence. Let us call this the "pragmatic account of reference".

8.2.1 The External Account of Reference

Let us begin with the first option in the proto-scientific project. Although in his former writing Stich seemed to have suggested that it is not charitable to assume that philosophers follow such a project, and also somewhat unclear what sort of research such a project would require, his argument above, that folk intuitions might be just massively mistaken about reference, seems to suggest that the truth about reference must be an intuition-independent fact, and thus perhaps an external fact. At least other authors (for example [Sch02]) have argued that any proponent of an argument from reference will have to assume that folk intuitions are fallible evidence for an external fact.

Let us assume that. As we said, since we seem to have no science yet studying this relation, we don't know much about it. Is it possible that reference is an external fact, independent of what we intuitively judge about reference in actual and possible cases? I have strong doubts that this is at all possible. We can be wrong about what the correct reconstruction of our implicit theory is. And we might on the basis of a mistaken reconstruction come to believe that someone, when using certain terms, did not refer to anything, while he in fact referred to something. Moreover, we might believe that we refer to something with a term, while in fact we don't. And we might also have all sorts of other false beliefs about the actual referents of our terms. But, assuming that language is created by us as a means for communication, could the way that language relates to the world be determined by something independent of how we intend to and actually use this instrument in communication? Perhaps there is a principled argument that shows that this would be absurd, but I think we do not even have to establish that in order to see that the external account of reference is just irrelevant for our concerns here.

Let's assume, for the sake of argument, that the external account of reference is true and that by bad cosmological luck every language user has a descriptivist theory internalized, while the external facts make it the case that a causal-historical theory is true. Thus, when somebody in our language community presents the argument for eliminativism, we all understand the premises in the same way (and believe they are true[33], after reflection on our intuitions), we also understand the conclusion in the same way (in the sense of 'understanding' that we are ready to make the same inferences from it, etc.) and believe that the argument is sound. We would then intuitively understand 'Beliefs do not exist' as expressing the claim that there is nothing that satisfies the description-cluster associated with '...is a belief' (while it in fact expresses the claim that nothing stands in the appropriate causal chain of our present usage of '...is a belief'). What we take the claim to express would be true, but we would be wrong in believing that it is this what 'Beliefs do not exist' actually expresses. Besides that we might think that it is absurd to assume that the facts about reference could be external facts in that way, it also doesn't seem to matter for the validity and soundness of the arguments from reference *as understood by us.* In the way that we are understanding 'Beliefs do not exist' we would be getting the ontological facts *right*, regardless of the fact that we would be getting the referential facts wrong, it seems.

Thus, if we are to inquire the referential facts in order to clarify matters of ontology, the external account of reference does not seem to recommend itself. If there is any reason to assume that such external facts about reference nevertheless exist, it might be an interesting project in itself to find out what they are, but it doesn't seem to be a project that could help us in

[33] We are still assuming that the other premises involved in the argument for eliminativism ('belief', 'desire', etc. are theoretical terms of folk psychology, folk psychology is a massively mistaken theory) are true.

any significant way with our ontological worries.

8.2.2 The Folk Semantic Program

The argument above gives strong support to the folk semantic program, as the program philosophers should engage in, in order to clarify matters of ontology, because it seems that it is *this* project that will help us to see whether ‘Beliefs do not exist’ is true, as we understand this sentence. But the folk semantic project has been challenged in at least two ways and we should address these challenges.

The most serious attack was made by Sebastian Schulz in [Sch02] in reaction to Stich’s initial suggestion that philosophers would by and large follow the folk semantic project. Schulz objects that the folk semantic project would just be either of irrelevance for matters of ontology or a highly implausible project, and that it is thus uncharitable to assume that philosophers (and those who put forward arguments from reference in particular) are following this project. Hence if Schulz’ objections could be substantiated, also the folk semantic project could not help us in matters of ontology. Schulz begins with observing that if we understand the theory of reference in analogy with a Chomskian theory of grammar (this is, as we remember, how Stich characterized the folk semantic program), our theory of reference could not be mistaken in the way folk physics can be mistaken. It does not make sense to say of our grammatical rules that they are true or false, accordingly it should not make sense to say of our theory of reference that it is true or false:

> [I]f we analyze our theory of reference on a par with such a grammatical theory, then there just seems to be no fact of the matter for a theory of reference as well. Statements like “The theoretical notion ‘electron’ of modern physics refers to an entity” would be analyzed like ‘X is a grammatical sentence in German”. [Sch02, 178]

Schulz goes on to argue that a “Chomskian grammatical sentence” can be correct though and that all there is to determine its correctness is to be found in the intuitions of the speakers of the language. Thus if we take the analogy seriously, we should also assume that common-sense intuitions are constitutive for the correctness of claims about reference. This could be either because there is no independent reality that makes these statements true, or because we have “full first-personal grasp” of this independent reality which is mirrored in our intuitions. As Schulz concludes, the claims of the eliminativist, like ‘Beliefs do not exist’ would thus either derive from claims that are not about an independent reality (and then ‘Beliefs do not exist’ itself not about an independent reality), or the eliminativist must assume that our intuitions about reference are infallible about an independent reality. Both possibilities do not seem to be open for the eliminativist. He seems to intend to make a claim about an independent reality when claiming 'Beliefs do not exist'. On the other hand, assuming that our intuitions about reference are infallible about an independent reality would be "terribly *ad hoc* and implausible" [Sch02, 179]. Thus the eliminativist should not understand himself as engaging in the folk semantic program. Schulz notes that this argument cannot be found in Stich's writings. But I believe the confusion underlying this argument is also one of Stich's basic confusions, which is why it seems worth analyzing.

First of all, we should make a distinction between our folk theory of reference as it is, according to the folk semantic program, internalized by the speakers of a language community and underlying their intuitive judgments about reference in possible and actual cases on the one hand, and our explicit reconstruction of that theory on the other. Let us call the former the 'internalized theory' and the latter the 'reconstructed theory'. While the internalized theory cannot be wrong, and while it perhaps doesn't make sense to say that it is true or false (as with the internalized grammatical rules[34]), our reconstruction of it (the reconstructed theory) *can* be true or false, depending on whether it is in accordance with the internalized theory. In a similar way can a reconstruction of the grammar of a language as used by a language community be true or false, depending on whether it conforms to the rules actually underlying the grammaticality intuitions of the speakers of that language. Thus, contra Schulz, it *does* make sense to say of a theory of reference that it is true or false, viz. if we are dealing with the reconstructed theory. This is important to emphasize, since arguments from reference have theories of reference as premises, and we want these to be truth-evaluable. The true nucleus in Schulz' remark is that the internalized theory of a language community is not in any sense true or false about anything, and neither are the intuitive judgments of speakers that are (without performance errors) results of the (subconscious) application of the internalized theory of reference.

Second of all, the statement

The theoretical notion 'electron' of modern physics refers to an entity.

is neither a statement of either, the internalized theory nor the reconstructed theory, nor is it derivable from any of these in a way that

X is a grammatical sentence in German.

[34] Note that this still allows for a performance/competence distinction, and that it also allows for speakers not (yet) fully competent. Thus the best reconstruction of the intuitive responses of an individual speaker at a time might not correspond to the internalized theory of that speaker (performance/competence), nor does the internalized theory of that speaker automatically coincide with the internalized theory of the language community (for the speaker might just not be fully competent with it).

might be derivable from a recursive reconstruction of German grammar, including the German lexicon.

The reason is that a theory of reference (of either kind) only specifies what conditions must be fulfilled for a term of a type (proper name, definite description, natural kind term, etc.) to denote or apply to anything. But, of course, the theory of reference (of either kind) is silent about whether these conditions *are* satisfied for a given term. Therefore, in order to arrive at

The theoretical notion ‘electron’ of modern physics refers to an entity.

we could not just rely on, for example,

$$\begin{gathered}\forall x \forall y (x \text{ satisfies the cluster of descriptions} \\ \text{associated by speakers of this language with } y \leftrightarrow y \text{ applies to } x)\end{gathered} \tag{ϵ}$$

even if the truth of (ϵ) were constituted by the intuitions of speakers of our language, but would need the additional substantial information that

ϕ is the cluster of descriptions associated by speakers competent in modern physics with ‘electron’,

and

$$\exists x(\phi(x)).$$

Clearly, at least the last claim is about an independent reality. Thus it is the “ontological input”, if you like, necessary to arrive at

The theoretical notion ‘electron’ of modern physics refers to an entity.

and eventually at an ontological claim. While the truth conditions of ‘Electrons exist.’ might well be constituted by the referential intuitions of speakers and by what they take their physical theory to be saying, whether these truth conditions are *satisfied* is not anymore a matter of intuition, but a matter of what the world (or independent Reality) is like. Accordingly, we will need to look elsewhere to find reasons to believe that these conditions are satisfied. Speaker intuitions play, typically[35], no role here.

Thus Schulz is just confused about the role intuitions play in the folk semantic program. The eliminativist, who follows the folk semantic program, does not need to assume that speaker intuitions are infallible guides to ontological truths, nor does he need to assume that statements like ‘Beliefs do not exist.’ are actually not about an independent reality. The folk theory of reference, as understood in the folk semantic program, merely entails claims like (ϵ) and (δ), but it does not entail claims about the actual reference of

[35] Except in cases like ‘Speaker intuitions exist.’, etc. where speaker intuitions would also be a relevant fact of independent reality.

specific terms, nor, for that matter, about the existence or non-existence of anything. Whether the latter claims are true might however follow from (ϵ) and (δ) together with claims about independent reality whose truth is *not* constituted by speaker intuition.

Although Stich does not himself endorse Schulz' argument, he nevertheless seems to share the confusion. A first sign of this is already the Argument from Idiosyncracy, discussed in 3.2, Stich's worry was that if the reference relation were depending on highly contingent and idiosyncratic facts, also our ontological claims should be then highly contingent and depending on idiosyncratic facts:

> [T]he word-world mapping that will be captured by the correct theory of reference will be a highly idiosyncratic one. It will be one member of a large family of word-world mappings, a member that stands out from the rest only because it happens to be favored by intuition. [...]
>
> [T]he fact that our intuitions pick out the particular word-world relation that we call *reference* rather than one of the many others in the envelope of genetic possibility is largely the result of the historical accident, in very much the same way that details of the grammar of our language or elements of our principles of politeness are in large measure the result of historical accidents. [...]
>
> If the only thing that distinguishes reference from [alternative possible word-world relations] is an historical accident, then the fact that '. . . is a belief' refers to nothing just isn't very interesting or important. But if that isn't interesting, then neither is the fact that beliefs don't exist. [Sti96, 49-50]

But that just seems to be the same confusion we have just discussed. Although it might be a contingent fact that the *truth conditions* of, for example, 'Beliefs exist' are the way they are, because it is an idiosyncratic, contingent fact which of the many logically possible reference relations a language community happens to adopt, this does not entail that therefore the *truth* of 'Beliefs exist' is an idiosyncratic, contingent fact. The latter is a question of whether the truth conditions of 'Beliefs exist' are *satisfied*, and this need not depend on the historical circumstances that settled the reference relation, but on very robust facts about the way the world is. *This* is the actual flaw in the Argument from Idiosyncracy.

Having this confusion out of the way, we should turn to the second objection against the folk semantic program. With this we now turn to one of the basic mistakes in the Argument from Social Psychology. As we have seen, Stich et al. argue that the folk semantic project is undermined by the (alleged) empirical variation in intuition about reference. The empirical variation would force the folk semanticist into referential pluralism, which suffers from two problems. One is that it assumes that intuitions about reference provide evidence about reference, which is again undermined by the empirical variation; the other is that referential pluralism entails that we are, unknowingly, talking past each other very often, which is absurd.

Let's first discuss whether an empirical result, establishing strong variation in intuition about reference in actual and possible cases could undermine the assumption that intuition is evidence for reference. As we have said, the folk semantic program is seen as relevantly similar with the Chomskian reconstruction of grammar. As Stich et al. argue, if we were to find such variation in intuition about grammaticality between speakers of what appears to be the same dialect, we would give up the assumption that speaker intuitions are evidence about the grammatical properties of the dialect these speakers speak. Similarly, we should give up this assumption in case of reference, when confronted with the variation results.

This is nonsense. As we said (and as Stich sees it), according to the Chomskian reconstruction of grammar, speaker intuitions are constitutive for grammar, and according to the folk semantic program, constitutive for reference. If we were to find heavy variation, we would give up the assumption that the speakers speak a common dialect, or share a language with a common reference relation. Unless we have started a project according to the external facts account of reference, which we have dismissed above, it does not make sense to say that speaker intuitions might not be good evidence for reference, just as it simply doesn't make sense to say in the Chomskian project that speaker intuitions aren't evidence about grammaticality.

Note what would be involved in case of grammaticality: speakers would be uttering sentences which they — also on reflection — find themselves grammatical, while at the same time finding the sentences of the other speakers, intuitively and even on reflection, ungrammatical. Why on earth would we think that speakers, who disagree on grammaticality in such a way, producing sentences in actual and regarding possible cases that do not cohere with the language use of the other members in the group, are speaking the same dialect? If we would fail to sort them into subgroups and would find that the grammaticality judgments of the speakers are also intra-personally unstable, we would at best conclude that there isn't much of a grammar in this dialect. But it wouldn't lead us to the conclusion that speaker intuition is no evidence about grammaticality, and I don't even see why it should provide us with a reason to doubt the feasibility of the project to take speaker intuitions as constitutive for grammaticality.

Perhaps Stich et al. think that the feasibility of such a project is put into doubt in light of the other "absurd" consequences that follow if we combine referential pluralism with the assumption that there is strong variation in intuition about reference. The absurd consequences were (a) that philosophical agreement or disagreement would depend on whether the philosophers would belong to the same intuition group, and (b) that since we do not know to which intuition group we (or any of our colleagues) belongs, we could be talking past each other without even knowing it.

Perhaps we should take the sting out of (b) first. So there is, under the as yet totally unsupported empirical assumption that we differ wildly in intuition about actual and possible cases, the consequence that we *could* be talking past each other, in philosophical disputes, without knowing it. Note first that we couldn't be talking at cross-purposes *and* know it. Talking past each other presupposes that we are unaware that we use certain terms with different meanings. Thus, *that we don't know it*, when we are talking past each other, doesn't add any "bizarreness" to the possibility that we might be talking past each other. Stich's emphasis of this is just a red herring.

Note furthermore, that Stich et al. do not claim that it would be impossible for us to find out about whether we are talking past each other or genuinely disagree. By reflecting on actual and possible cases, finding our intuitions about reference constantly in disagreement with those of our colleagues, we could come to the conclusion that our concept of reference is simply different from theirs and that we thus understand, for example, 'Beliefs do not exist.' just differently. Consequently, it is also not, in any sense, *inscrutable* whether we are actually disagreeing or merely talking at cross-purposes. So, in order to distinguish actual disagreement over, for example, 'Beliefs do not exist.' from mere talking past each other, we would need to sort philosophers into intuition groups, with respect to their referential intuitions about actual and possible cases. Apparently, we just reduced the "absurdity" to (a).

What consequences would that have for the profession and its methodology? Segregation? Should, for example, ontology conferences have different sections for philosophers of different intuition groups? Would we need journals for each intuition group with intuition-group specific selected reviewers to make sure that the papers are really reviewed by our (intuition-) peers? Should we, in order to reduce talking past each other in class, have intuition-tests before accepting students to graduate schools? Wouldn't all of this be absurd? Perhaps we should in the light of absurd consequences like these give up the idea that there are substantive reference relations, shouldn't we?

The mere impracticality of these practical consequences should certainly not lead us to give up the idea that there are substantive reference relations, or the idea that at least one of them must be true. Although science is usually helpful in making our lives a whole lot easier, we can't reject the consequences it arrives at by pointing out that they would make our lives too complicated. How should that argument work? I guess that the best version would be this: in order to argue from theories of reference to conclusions about ontology, it is likely (given the assumed variation in intuitions) that there will be a lot of work to be done before we will be able to agree on the premise stating the relevant reference relation. This makes arguments from reference inefficient, if there is a quicker way to arrive at the ontological consequences. But we have seen above, that this reasoning would be confused: because of (δ), any ontological claim expressed by some existential statement is in danger of being ambiguous if there is the assumed variation in intuition. So there can't be any other method that would give us an ontological conclusion in any faster way. We still would have to settle what reference relation underlies all relevant intuitions, and that would be just the same work as with the flight to reference strategy. Thus, if we believe that reconstructing all relevant reference relations in any case is too much work, we should not just give up arguments from reference, but ontology altogether. But this would surely not convince philosophers: they are interested to find out the truths about ontology, and they know since over 2000 years that that's not easy. Why should they give up now, being informed that it is more work that they might have thought so far.

But, anyway, the consequences envisaged wouldn't follow even if empirical research established that there is massive variation in intuitions about reference among philosophers. To some degree, we could avoid the segregation by relativization and disambiguation. Once we knew that, for example, 'Beliefes do not exist.' is ambiguous, in the sense that philosophers from different intuition groups might attach a different meaning to it due to different implicit theories of reference, we could replace the ambiguous claims with disambiguated ones, for example 'There is nothing that satisfies the set of causal (and, perhaps, other) roles that folk psychology specifies and labels 'beliefs'.' Note that also the latter claim is an existence claim, although now with a definite description in place of, what might perhaps be, a natural kind term. The ambiguity between a causal-historical and a descriptivist implicit theory could be resolved this way, since they (according to their common reconstruction) determine the reference of definite descriptions in the same way.

Of course, while we are exploring logical possibilities, it could turn out that the implicit theories of reference do not agree in the determination of the reference of any (type of) term. In that case a disambiguation that is interpreted according to both theories in the same way could be difficult to achieve. But even in that case, there is a strategy by which we could avoid segregation and even relativization completely — a strategy which, I believe, is already in use.

8.2.3 The Pragmatic Account of Reference

A few paragraphs above, we have distinguished between two possible accounts of reference that Stich seems to conflate in his conception of the "proto-scientific" program. The *external account* of reference that assumes the reference relation to be determined by facts external to the intentions and beliefs of speakers, and the *pragmatic account* that assumes the reference relation to be a possible subject of explicit convention, and suggests to fix a word-world relation that suits the sciences best. In light of massive variation between speakers in intuition about reference, and the constant danger to speak at cross-purposes, we should perhaps just fix this relation once and for all by explicit convention.

What can be fixed by explicit convention has its limits though. Especially when it comes to the meaning of fundamental concepts such as 'reference' and 'existence', we are pretty much in the situation described by Otto Neurath's famous analogy:

> We are like sailors who are forced to reconstruct totally their boat on the open sea with beams they carry along, by replacing beam for beam and thus changing the form of the whole. Since they cannot land they are never able to pull apart the ship entirely in order to build it anew. The new ship emerges from the old through continuous transformation. [Neu98, 216]

But this isn't bad news. It isn't news, because the same holds for concepts like 'logical consequence'. It isn't bad, because, like the sailors on Neurath's boat, we can arrive at the intended result by continuous transformation of what we started with. Through our competence with language we have intuitions about reference, and we developed theoretical approaches to systematize these. But not only do our systematizations depend on our intuitions, the latter are also shaped by our theorizing. As Stich et al. have noted themselves, "given the intense training and selection that undergraduate and graduate students in philosophy have to go through, there is good reason to suspect that the [intuitions of professional philosophers] may be reinforced intuitions" [MMNS04, B9]. In other words, by practicing the method of possible cases, philosophers not only (!) collect the data that a true account of reference has to agree with, they also train their intuitions and bring their internalized theory of reference in line with the internalized theories of their peers. This process has been described with the metaphor of finding a *reflective equilibrium*.[36]

[36] It would be unfair not to note that Stich is very critical of the idea that fundamental notions of philosophy could be founded in a reflective equilibrium process. In his criticism (Cf. [Sti98], [SN97]), Stich refers to the fact that folk intuitions regarding basic concepts of philosophy sometimes do not agree with our theoretical reconstructions, and that we — especially when it comes

According to this view, our reconstructed theory refines the internalized theory we started with, and both theories change in the process that leads, eventually, to their agreement. All kinds of consideration can enter into this deliberation process (often referred to as 'wide reflective equilibrium'), including, of course, our intuitions about ontology and other philosophical convictions we might have. The degree to which they might influence the end result, will (in part) be a question of their relative strength. This clears up another confusion: contra to what Stich et al. suggested in reaction to Strategy II (see section 7.2.2), it is *not* uncommon in philosophy to consider more than just prima facie intuitions when reflecting over the right account of reference. Stich et al. worried that other reasons will fail to be "independent", and then ultimately question begging. But this, first of all, need not be the case. Note that the ontological convictions we might rely on when selecting a theory of reference need not be the very same ontological convictions that we argue for in an argument from reference. I might fancy a theory of reference, because it (amongst other things) accords with my intuitions when it comes to the existence of phlogiston, and I might then use this theory of reference in an argument from reference. Although one of the reasons to chose the theory is an ontological reason, the reason is, nevertheless, *independent*. Second of all, however conventional I might choose the reference relation, the ontological claims that follow from (or rather *with*) it will ultimately depend on the way the world is. A reference relation fixes the truth conditions of existence claims, but it doesn't fix their truth (as we have seen above). Thus, by this procedure we can hardly beg the question anyway.

To sum this up, reference relativism is a way to accommodate the hypothetical result that there is strong variation in intuitions about reference. It would neither lead to absurd consequences about the actual communication situation, nor would it lead to bizarre methodological consequences: the methodology already in place can perfectly deal with it in theory and the claimed results of experimental philosophers about the convergence of intuitions among professional philosophers suggest that the methodology perfectly deals with the situation in practice. Accordingly, the hypothetical result — as interesting as it would otherwise be — would certainly not undermine arguments from reference, or substantive theories of reference in

to the central notions of logic or statistics — would rather consider these deviant intuitions mistakes than expressions of reflective equilibria in their own right and hence not regard them as just as valid as the theoretical reconstructions they disagree with. Thus, Stich argues, intuitions alone cannot serve as a foundation of basic concepts, consequently the metaphor of reflective equilibrium misdescribes the actual foundation of these concepts. Stich's criticism overlooks, of course, that these folk intuitions simply aren't in reflective equilibrium with general principles these folk endorse, and hence can not constitute interesting counterexamples. It overlooks furthermore that the deviant judgments are considered mistakes from the point of view of *another* system and not mistakes *simpliciter*. Stich's view is discussed and rejected in [CR06].

general.

9 What *is* Wrong with Arguments from Reference

What utterances like 'Beliefs exist.' and 'Dephlogisticated air supports combustion better than ordinary air.' say, depends (in part) on what theory of reference is true. Accordingly, the truth conditions and, given the way the world is, the truth-value of these claims depends on which theory of reference is true. The part of the theory of reference this depends on is not to be found in (δ), and other such consequences of our theories of reference that deflationists about reference are content to settle for. It is rather found in how the theory of reference explicates the reference-relation, and such an explication is only to be provided by substantive theories of reference. (δ) being a consequence of our theories of reference is the *reason* why substantive theories of reference are part of what determines the truth-value of such claims.

Insofar as we are interested in assessing the truth of such claims, we need to know which theory of reference is true. As it concerns our own ways of speaking, we have tacit knowledge of which theory of reference is true for these ways. This is why we understand what these sentences say, in the relevant sense, already without having reconstructed a theory of reference. But if we are able to make our substantive theory of reference explicit, we might thereby get a better understanding of how claims, as the ones mentioned above, follow from other claims. Knowledge like this enables us then to draw conclusions about the truth-value of such claims from other parts of our knowledge. This is why theories of reference are important, especially for philosophers whose business it is to examine whether claims made on the basis of other claims, are properly supported by these.

9.1 What *is* Wrong with the Argument for Eliminativism

The argument for eliminativism exemplifies this. If 'belief', 'desire', etc. are theoretical notions of folk psychology, and if folk psychology is massively mistaken, then we can conclude on the basis of a descriptivist theory of reference that beliefs and desires do not exist. What could be wrong with such an argument, if we assume that a descriptivist theory of reference is true?

Imagine a mad scientist in some obscure lab, working on a theory of quarks. Since he is a little eccentric, his theory, let's call it Θ, is somewhat unconnected to the standard theory, and, in addition to that, hopelessly mistaken. Now, since ‘quark’ is a theoretical notion of Θ, shall we conclude from this that quarks do not exist? Perhaps not, if we belief that our actual theory of quarks is not mistaken. ‘Quark’ could also be a theoretical notion of our (presumably true) theory Θ*, and the falsity of Θ thus irrelevant for whether quarks exist.

One might object that the situation described in this thought experiment is different from the situation in the philosophy of mind case. Here we assume (with the eliminativist) that there is no such theory as Θ*, which has ‘belief’ and ‘desire’ also among its theoretical notions, and is, furthermore, true. But what does it mean that there is no other theory that has ‘belief’ and ‘desire’ among its theoretical notions? That these expressions do not occur in our fancier psychological theories? That would be a rather uninteresting claim, even if were on Carnap’s side when it comes to the distinction between internal and external questions of ontology. That is, even if we assumed that ontological question can, without loss, be translated into the formal mode; two theories that are merely distinct with respect to the expressions they use, should not be considered *ontologically* distinct. If that would be what the eliminativist wants to claim, when claiming that ‘Beliefs do not exist.’ he wouldn’t be making an ontological claim at all.

The eliminativist has to claim a bit more; he wants to say that ‘belief’ and ‘desire’ are theoretical notions of a massively mistaken theory, and it is unlikely that *what these expressions would denote if the theory were true* could ever be among the things that a true theory of that domain quantifies over. One way to make that point, past a mere comparison of the two vocabularies, could be to analyse the theoretical terms of both theories in the Ramsey/Carnap/Lewis-way. In this case we assume a shared non-theoretical vocabulary of “O-terms” for both theories, T and T^*, and define the “T-terms”, $\tau_1, \ldots, \tau_n$ and $\tau_1^*, \ldots, \tau_n^*$, by means of definite descriptions:

$$\begin{gathered}\tau_1 = \imath y_1 \exists y_2 \ldots y_n \forall x_1 \ldots x_n (T[x_1 \ldots x_n] \equiv .y_1 = x_1 \& \ldots \& y_n = x_n)\\ \ldots \\ \tau_n = \imath y_n \exists y_{n-1} \ldots y_n \forall x_1 \ldots x_n (T[x_1 \ldots x_n] \equiv .y_1 = x_1 \& \ldots \& y_n = x_n)\end{gathered}$$

and

$$\begin{gathered}\tau_1^* = \imath y_1 \exists y_2 \ldots y_n \forall x_1 \ldots x_n (T^*[x_1 \ldots x_n] \equiv .y_1 = x_1 \& \ldots \& y_n = x_n)\\ \ldots \\ \tau_n^* = \imath y_n \exists y_{n-1} \ldots y_n \forall x_1 \ldots x_n (T^*[x_1 \ldots x_n] \equiv .y_1 = x_1 \& \ldots \& y_n = x_n).\end{gathered}$$

As [Lew70] suggests, by replacing each T-term by its definiens throughout the postulates of T and T^*, we obtain two O-sentences (in the O-vocabulary) which say, respectively, that T is realized by the n-tupel consisting of the first, second, …, nth component of the unique realization of T, and that T^* is realized by the n-tupel consisting of the first, second, …, nth component of the unique realization of T^*. Perhaps this way it might be possible to see past the theoretical vocabulary whether what the one theory would be about, is among the things the other theory is about.

Now the eliminativist's claim is a claim about translatability, or reducibility. And, indeed, the eliminativist is making that claim:

> [T]he greatest theoretical synthesis in the history of the human race is currently in our hands, and parts of it already provide searching descriptions and explanations of human sensory input, neural activity, and motor control.
>
> But [folk psychology] is no part of this growing synthesis. Its intentional categories stand magnificently alone, without visible prospect of reduction to that larger corpus. [Chu81, 75]

But, of course, this sort of theory-comparison will yield a positive result only if both theories have the same truth value. Since we assume that folk psychology is massively mistaken, the ontological commitments of the theory, so reconstrued, will, quite trivially, not be reducible to the ontological commitments of a true theory. But this holistic form of descriptivism is not plausible as a theory of reference. The reason that it is not plausible is *not* that if it were true, there would be cases of talking past each other between scientists and incommensurability between succeeding theories, which we took, pre-theoretically, to be cases of disagreement and progress. It is implausible, because disagreement and progress would become *impossible*, and these notions *meaningless*. Genuine disagreement would be logically impossible, since at least one party is not talking about anything; genuine progress, in the normal, gradual sense, would be logically impossible since all false theories would not be about anything and no two logically contrary theories comparable with respect to what they got right.[37] Any argument relying on such a holistic theory would seem problematic; not because it has a theory of reference as a premise, but because it has a theory of reference as a premise that seems implausible.

As we have said already in the beginning of this paper, the descriptivist is free to be a bit more liberal here, and the argument for eliminativism would still be valid. Lewis, for example, allows for the theoretical terms of a corrected theory T to refer to the "nearest near-realization" of T, even if that is not a realization of T itself, but a realization of components of the unique realization of the corrected version of T.

Other descriptivist proposals, as [Pap96] and [KN01], suggest that the assumptions of a given theory, involving a given term τ_i, could be distinguished into those that contribute to τ_i's definition, $T_y(\tau_i)$, those that perhaps contribute to it, $T_p(\tau_i)$, and those that do not contribute to it, $T_n(\tau_i)$. The

[37] There would at best be a notion of "progress" in a non-standard sense, in which the sudden event of a true theory would be progress with respect to all activities before, that had the intention to result in a true theory but failed.

admitted vagueness of the borders between the these three sets of assumptions, still allows for determinate reference iff the assumptions in $T_y(\tau_i)$ and the assumptions in $T_y(\tau_i) \bigcup T_p(\tau_i)$ have a unique satisfier. Imprecise definitions of this sort, however, might become problematic if the assumptions in $T_y(\tau_i) \bigcup T_p(\tau_i)$ lack a satisfier. In this case, given the vague status of the assumptions of a theory that "perhaps" contribute to the definition of a theoretical term, $T_p(\tau_i)$, it might seem indeterminate whether the term in question refers or not (similarly, on Lewis' account, this might be so if it is indeterminate whether there is a nearest near-satisfier of T). This could be the situation with the argument for eliminativism; cf. [Pap96, 18].

Claiming on that basis that the entities denoted by τ_i do not exist, resolves the vagueness in one way. It assumes that it is illegitimate to drop any of the assumptions in $T_p(\tau_i)$ from the definition of τ_i. An argument for eliminativism, in this case, would not seem problematic because it has a theory of reference as a premise that seems implausible, but because it attempts to introduce a precisification of the definition of τ_i in a certain direction, by disguising it as straightforward deduction from the theory of reference. As a deductive argument, this seems invalid unless reasons are provided why the assumptions from $T_p(\tau_i)$ should not be dropped from the definition of τ_i. To counter such an argument for eliminativism one might thus propose a precisification of the definition of τ_i that has a unique satisfier, or is likely to have such, and, clearly, other than semantic considerations must be put forward on both sides to settle this dispute. In this case, arguing from reference would simply be inconclusive, even though there would not be any doubts about which substantive theory of reference is the true one.

9.2 What *is* Wrong with the Argument for Anti-Eliminativism

Of course, one might also find the argument for eliminativism problematic, because it relies on a descriptive theory of reference. If one holds a causal-historical theory of reference for theoretical terms, one might think that no set of assumptions of T *need* to be satisfied in order for a theoretical term τ_i to denote. This would stop the argument from reference for eliminativism, but could it give rise to a sound argument for anti-eliminativism? The situation might seem very similar here. If we take the argument for anti-eliminativism as it is reconstructed here (following Stich), the causal-historical theory claims that *all* theoretical terms refer, even if the theory in question is massively mistaken. But, again, this seems just highly implausible to assume for a theory of reference.

Some theoretical posits just don't refer. To have a clear case, let's consider a scientists who posits an as yet undetected star in a certain region of space to account for the fact that the astronomic data he collected is not in accordance with the predictions of the established theory. Assume furthermore that no such star exists in this region of space, the established theory (minus the *ad hoc* assumption of the existence of a further star) is true, but his data was corrupted by a systematic measuring mistake. Let's assume that the theoretical term introduced was τ_i. Did τ_i refer? If your theory of reference says that it does, then this theory will just seem to many very implausible. To what should it refer? The measuring mistake? Again, arguments from reference involving such a theory might thus seem problematic, because they have an implausible theory of reference as a premise.

Typically, causal-historical theories do not claim that every theoretical term refers, but claim, as we have said before, that theoretical terms refer to the entity at the other end of the causal chain that leads to the present usage of that term, if there is one. In case there is no entity at the other end, the term does not refer to anything. But from the mere knowledge that a term is a theoretical notion of a massively mistaken theory, we do not *thereby* know that there is anything at the other end of the causal chain that leads to our present usage of the term. There might be, and in that case the term in question might have a denotation, but that is the best we can say. Again, as in the case of the argument for eliminativism with the weakened descriptivist theory, the argument seems invalid, unless it provides an additional reason why we should believe that there is something at the other end of the causal chain that leads to our present usage of the central terms of folk psychology. Arguing from reference is inconclusive, even if we have no doubts about which theory of reference is true.

To sum this up: if we follow Stich's reconstruction of the argument for eliminativism and the argument for anti-eliminativism, both arguments seem to be problematic, because they either assume (among their premises) a theory of reference that seems just wildly implausible, or, if the relevant premise is weakened to account for the implausibility, become invalid. In the first case, however, the arguments do *not* fail, because they establish an ontological conclusion with a semantic premise. The theory of reference is not implausible in comparison with the ontological conclusion it leads to — it is implausible as a theory of reference. In the second case, arguing from reference isn't inconclusive because the theories of reference are questionable or because they'd lack a premise that connects a claim about reference with a claim about ontology, but simply because more factual information is needed to arrive at a conclusion about whether the terms in question refer, and this additional information is not provided in Stich's reconstruction of the respective arguments.

9.3 Are there any Bad Arguments from Reference?

As I indicated already in the introduction of this paper, I have my doubts that many arguments from reference were ever put forward. But since at least Stich claims of himself that his (early) argument for eliminativism was an argument from reference, I shall not attempt to try to understand the author better than he understands himself. For many of the other examples, it though seems to me that the actual arguments were a bit better than how Stich reconstructs them. Let's look at Lycan's argument again:

> Unlike [David Lewis], and unlike [Dennett] and [Stich], I am entirely willing to give up fairly large chunks of our commonsensical or platitudinous theory of belief or of desire (or of almost anything else) and decide that we were just wrong about a lot of things, without drawing the inference that we are no longer talking about belief or desire. To put the matter crudely, I incline away from Lewis's Carnapian and/or Rylean cluster theory of reference of theoretical terms, and toward [Putnam's] causal-historical theory. [...] I think the ordinary word "belief" (qua theoretical term of folk psychology) points dimly toward a natural kind that we have not fully grasped and that only mature psychology will reveal. I expect that "belief" will turn out to refer to some kind of information-bearing inner state of a sentient being [...], but the kind of state it refers to may have only a few of the properties usually attributed to beliefs by common sense. Thus I think our ordinary way of picking out beliefs and desires succeeds in picking out real entities in nature, but it may not succeed in picking out the entities that common sense suggests that it does. [Lyc88, 31–32]

I do not read this passage such that Lycan is deriving his confidence that 'belief' will "turn out to refer to some kind of information-bearing inner state of a sentient being" from the causal-historical theory of reference he endorses. As I understand this passage, he believes (on other grounds) that some kind of information-bearing inner state of a sentient being is what stands at the other end of the causal chain that leads to our present usage of the term 'belief', and this — in turn — is his reason to believe that 'belief' applies to something. Similarly, Kitcher doesn't believe that 'dephlogisticated air' refers to something, because his hybrid theory of reference tells him so, but because his theory of combustion tells him that oxygen was at the other end of the causal chain of occasional usages of 'dephlogisticated air', and since he takes his theory of combustion to be *true*, he can be confident that 'dephlogisticated air' referred to something on these past occasions.

Whether Lycan's argument is, after all, a very good one, is then, of course, still debatable. One worry could be that causal-historical theories of reference do not really recommend themselves for analyzing the reference of theoretical terms. One of the biggest problems for a causal-historical theory in this realm is that terms like 'neutrino', 'positron', and 'quark' were explicitly introduced to refer to hypothetical entities which were conjectured to play certain theoretically specified roles, before any direct experimental manifestation of these entities were available for any dubbing ceremony (see [Pap96, 4]. Thus something like Kitcher's hybrid theory would be more convincing in such an argument for the reference of a theoretical term. But I guess Lycan could have made his response also within the framework of a descriptivist theory, by answering to the eliminativist that he is ready to reduce the set of defining assumptions for 'belief', $T_y(\text{'belief'}) \bigcup T_p(\text{'belief'})$, to the assumption

$$\text{belief} = \imath y \forall x(x \text{ is an information-bearing inner state of a sentient being} \equiv (y = x)).$$

Whether such a radical reduction in the defining assumptions for 'belief' makes sense, is then, of course, for psychologists to decide.

9.4 Conclusion

In this paper I analyzed several attacks against so-called "arguments from reference". These are arguments that establish a conclusion about ontology or epistemology by departing from a substantive theory of reference as one of their major premises. I argued in detail that most criticisms of arguments from reference are misplaced. Following the reconstruction of arguments from reference in actual philosophical debates that is common in these attacks, it turns out that these arguments are valid. It also turns out that the empirical data does not undermine arguments from reference *per se*. However, as we have seen, there is indeed something wrong with some arguments from reference so reconstructed. They rely on a too strong (and thus implausible) interpretation of the respective theory of reference, but would become invalid, if a more adequate interpretation were used. Thus these arguments from reference would be bad arguments if they existed. I suggested, however, that this seems in the prominent cases to be a misreconstruction of the sources. Neither Lycan, nor Kitcher seem to have presented the argument ascribed to them.

One might find arguments from reference problematic for a variety of reasons still. For example, one might not be persuaded that folk psychology is a theory, or a theory in the standard sense; or one might not be persuaded that it is massively mistaken, etc. It might also turn out that all theories of reference considered in this paper are not the best account for the reference of theoretical terms, while the best account turns out to be really irrelevant for matters of ontology and truth. My sole aim in this paper was to show that the criticisms of arguments from reference, prominently put forward by Stich and his co-authors, are, in any case, totally misplaced and involve an impressive amount of confusions, that I hope to have cleared up somewhat.

References

[Bis03] Michael Bishop. The pessimistic induction, the flight to reference and the metaphysical zoo. *International Studies in the Philosophy of Science*, 17:161–178, 2003.

[BS98] Michael Bishop and Stephen Stich. The flight to reference, or how not to make progress in the philosophy of science. *Philosophy of Science*, 65:33–49, 1998.

[Chu81] Paul M. Churchland. Eliminative materialism and the propositional attitudes. *The Journal of Philosophy*, 78(2):67–90, 1981.

[CR06] Daniel Cohnitz and Marcus Rossberg. *Nelson Goodman.* Acumen, Chesham, 2006.

[Cra98] Tim Crane. How to define your (mental) terms. *Inquiry*, 41:341–54, 1998.

[Fie94] Hartry Field. Deflationist views of meaning and content. *Mind*, 103 (New Series)(411):249–285, 1994.

[Gus98] Don Gustafson. Deconstructing the mind by stephen p. stich. *Philosophical Psychology*, 11(4):542–546, 1998.

[Haw00] John Hawthorne. Deconstructing the mind by stephen p. stich. *Philosophy and Phenomenological Research*, 60(2):479–483, 2000.

[Jac98] Frank Jackson. *From Metaphysics to Ethics:.* Oxford University Press, Oxford, 1998.

[Kit93] P. Kitcher. *The Advancement of Science.* Oxford University Press, New York, 1993.

[Kit98] Patricia Kitcher. Deconstructing the mind by stephen p. stich. *The Journal of Philosophy*, 95(12):641–644, 1998.

[KN01] Fred Kroon and Robert Nola. Ramsification, reference fixing and incommensurability. In P Hoyningen-Huene and H. Sankey, editors, *Incommensurability and Related Matters*, pages 91–121. Kluwer, Amsterdam, 2001.

[Kün03] Wolfgang Künne. *Conceptions of Truth.* Oxford University Press, Oxford, 2003.

[Lew70] David Lewis. How to define theoretical terms. *Journal of Philosophy*, 67:427–446, 1970.

[Lew72] David Lewis. Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50:249–258, 1972.

[Lyc88] William G. Lycan. *Judgement and Justification.* Cambridge University Press, Cambridge, 1988.

[MMNS] Ron Mallon, Edouard Machery, Shaun Nichols, and Stephen Stich. Against arguments from reference. In David Chalmers and Ryan Wasserman, editors, *Metametaphysics*, pages ??–?? Oxford University Press, Oxford, ?

[MMNS04] Edouard Machery, Ron Mallon, Shaun Nichols, and Stephen Stich. Semantics, cross-cultural style. *Cognition*, 92(3):B1–B12, 2004.

[Neu98] Otto Neurath. Probleme der kriegswirschaftslehre. In Rudolf Haller and Ulf Höfer, editors, *Otto Neurath: Gesammelte ökonomische, soziologische und sozialpolitische Schriften, vol 5*, volume 5, pages 201–249. Hölder-Pichler-Tempsky, 1998.

[Pap96] David Papineau. Theory-dependent terms. *Philosophy of Science*, 63(1):1–20, 1996.

[Sch02] Sebastian Schulz. *Alien Minds: Investigating Eliminative Materialism.* Mentis, Paderborn, 2002.

[SN97] Stephen Stich and R. E. Nisbett. Justification and the psychology of human reasoning. In C. Z. Elgin, editor, *The Philosophy of Nelson Goodman Vol. 2: Nelson Goodman's New Riddle of Induction*, volume 2, chapter 247–288. Garland, 1997.

[Sti83] Stephen Stich. *From Folk Psychology to Cognitive Science: The Case Against Belief.* MIT Press, Cambridge, Massachusetts, 1983.

[Sti90] Stephen Stich. *The Fragmentation of Reason.* MIT Press, Cambridge, Massachusetts, 1990.

[Sti96] Stephen Stich. *Deconstructing the Mind.* Oxford University Press, Oxford, 1996.

[Sti98] Stephen Stich. Reflective equilibrium, analytic epistemology and the problem of cognitive diversity. In M. R. DePaul and W. Ramsey, editors, *Rethinking Intuition: The Psychology of Intuition and Its Role in Philosophical Inquiry*, pages 95–112. Rowman and Littlefield, Lanham, MD, 1998.

[WNS01] J. Weinberg, S. Nichols, and S. Stich. Normativity and epistemic intuitions. *Philosophical Topics*, 29(1 and 2):429–459, 2001.