JOURNAL OF PHILOSOPHICAL RESEARCH
VOLUME 28, 2003

CHARITY, SELF INTERPRETATION, AND BELIEF

HENRY JACKMAN
YORK UNIVERSITY

ABSTRACT: The purpose of this paper is to motivate and defend a recognizable version of N. L. Wilson's "Principle of Charity." Doing so will involve: (1) distinguishing it from the significantly different versions of the Principle familiar through the work of Quine and Davidson; (2) showing that it is compatible with, among other things, both semantic externalism and "simulation" accounts of interpretation; and (3) explaining how it follows from plausible constraints relating to the connection between interpretation and self-interpretation. Finally, it will be argued that Charity represents a type of "minimal individualism" that is closely tied to first person authority, and that endorsing Charity in our interpretations of others reflects a commitment to capturing, from the third-person starting point, their first-personal point of view.

I. INTRODUCTION

The purpose of this paper is to motivate and defend a recognizable version of what is commonly known as "The Principle of Charity." The version of the Principle familiar through the work of Quine and Davidson has struck many philosophers as implausible, but it will be argued here that this is because those authors misrepresent the Principle's character and the possible motivations for it. Charity is not subject to most of the objections commonly launched against it, and, properly understood, it can be motivated in terms of plausible constraints internal to the interpretation process.

II. THE PRINCIPLE AND ITS GENERAL FORM

The Principle of Charity was originally formulated by N. L. Wilson as the following semantic rule used to determine the referents of the names in a speaker's language:

> We select as designatum [of a name] that individual which will make the largest possible number of [the speaker's] statements true.[1]

This may initially seem like a mere description of how *we* go about guessing what the independently determined referents of a speaker's words are. Nevertheless, Wilson clearly intends Charity to be part of a more general account of what determines the referents of the speaker's words. As Wilson puts it, the Principle is part of an answer to the question, "How do words hook up to things?"[2] For instance, immediately after the characterization of Charity quoted above, Wilson rephrases the Principle, claiming that "the designatum is that individual which satisfies more of the [assertions] containing [the name] than does any other individual." Charity is thus a principle that should guide how the interpreter assigns a referent to a name because it captures factors that are constitutive of the semantic relations that the interpreter is investigating. Indeed, the Principle bears more than a passing resemblance to the sorts of "cluster" theories of proper name reference advanced around the same time. On such cluster accounts, the object that satisfies most of the beliefs associated with a name is what the name refers to. The cluster account is clearly a thesis about proper name reference, not just a description of our interpretive practices, and the Principle of Charity should be understood in the same way.

Wilson's principle is tailored to proper name reference, but his account generalizes quite naturally along three dimensions. Doing so replaces statements with the more general notion of commitments, referents with the more general notion of semantic values, and strict numerical maximization with a more general "weight-sensitive" notion.

Wilson formulates Charity in terms of the statements made by the speaker, but maximizing the truth of all of the speaker's *beliefs* is clearly his target.[3] It is, from here, not much of stretch to treat the Principle as enjoining us to select the objects that make true the most of what will hereafter be called the "commitments" of the speaker. The commitments include not only explicitly held beliefs, but also the interpretee's implicit presuppositions and assumptions. In Quinean terms, one might say that a subject's commitments include all of the sentences that she would be disposed to assent to were she to be queried about them. Many of these implicit commitments need not be understood as explicitly represented in any way. Rather, we need only to be disposed to act in a way that will manifest a commitment to the truth of what is implicitly understood. "Belief" has, by contrast, a comparatively "occurrent" feel, and many people might be inclined to think that one could be disposed to assent to statements such as "My bed will not turn to oatmeal when I sleep" without being said to actually *believe* them. Furthermore, one's commitments extend beyond those sentences that one would assent to. One is, for instance, committed to the logical consequences of one's commitments, whether one would assent to them or not.[4] Some are happy to use the term "belief" for this more general sense of commitment, but since others are not, the subject's "commitments" will be the focus of this more general formulation of Charity.

Furthermore, the notion of maximization must be generalized as well. All of the speaker's commitments need not be treated equally. Some will be more centrally held than others, and thus carry more weight when it comes to determining which object or objects "maximizes the truth" of the set.[5] There is nothing charitable about an assignment of referents that makes a centrally held commitment false in order to preserve the truth of two marginal commitments.

How much weight a commitment has can be understood in terms of the comparative entrenchment of our commitments. In particular, a commitment (or set of commitments) *A* can be understood as *more entrenched* for an interpretee than commitment (or set of commitments) *B* if and only if the interpretee would hold on to *A* and give up *B* after learning that one of the two is (or may be) false.[6] Of course, there will be no guarantee that a subject's dispositions to revise their commitments will be well ordered. That is to say, a subject may be disposed to give up *A* to preserve *B*, *B* to preserve *C*, and *C* to preserve *A*. In such cases, there may be some indeterminacy as to just how much weight any particular commitment should be understood as having. Furthermore, the comparative entrenchment of a subject's commitments may itself turn out to vary from context to context. However, indeterminacy and context sensitivity in the weight of our commitments is not a problem if it corresponds to an indeterminacy and context sensitivity in what we are intuitively taken to mean, and it can be argued that this is precisely what we find.[7]

Indeed, the idea of a weighted total is essential once we have moved from statements to commitments. Since the number of statements we make is finite, the idea of counting statements is a coherent (if impractical) one. On the other hand, since we are committed to the consequences of our commitments, and there are countless sentences that we would be disposed to assent to, there is no way to count all our commitments.[8] Each commitment will have countless consequences, and so the idea of maximizing the *number* of true commitments seems unworkable. We must rely on the fact that my commitment to "John has exactly three dogs" has as much weight as the potentially infinite set of commitments: "John has more than one dog," "John has more than two dogs," "John has three dogs," "John has less than four dogs," etc. Given how we understand the comparative weight/entrenchment of our commitments in terms of our inclinations to revise our beliefs, we can make sense of one such set of commitments as having more weight than another in terms of the speaker's willingness to give the other set of commitments up were she to discover that the two sets conflicted. Consequently, "maximizing the truth of a speaker's commitments" should be understood as treating as true the maximally entrenched consistent subset of the speaker's commitments.

In addition to these other generalizations, "semantic values" will here be substituted for objects or designata. The semantic value of a word is whatever determines the contribution the word makes to the truth values of the sentences in which it occurs, and Charity need not, in itself, settle the question of what we should understand the semantic values of our terms to be. In many cases, the semantic value of the term may be best understood as an object, but this more general way of putting the issue will allow Charity to apply less controversially to an entire language, especially to predicates, logical connectives, etc.[9]

Finally, Charity can't be applied to single words, since an assignment that maximizes the truth of the commitments associated with a single word might bring a general decline in true commitments when sentences containing other words are considered. Consequently, the Principle should be formulated as one that applies holistically to all of the words in the language at once. With these generalizations in place, the Principle of Charity becomes:

> The semantic values of the words in a speaker's language are the values in the set that maximizes the truth of the speaker's commitments.

This generalization preserves the spirit of Wilson's principle, and allows his own formulation to be an application of the Principle to the special case of proper names. As we shall see, this cannot be said of many other presentations of Charity.

III. THE RECEPTION AND DISTORTION OF THE PRINCIPLE

Wilson's principle became the focus of a certain amount of philosophical attention through Quine's endorsement of it in his *Word and Object*. Nevertheless, the relation between Wilson's principle and Quine's views is far from clear. Quine approvingly cites Wilson's principle when he advances the following maxim for the translator: "assertions startlingly false on the face of them are likely to turn on hidden difference of language. . . . one's interlocutor's silliness, beyond a certain point, is less likely than bad translation."[10] However, Quine's discussion represents a major reworking of the Principle, and it has had serious consequences for how Charity has subsequently been perceived. Rather than being part of a philosophical account of what determines the semantic values of our terms, Charity becomes more of a common sense heuristic maxim that guides the interpreter generally. Quine's maxim is a useful guide for investigating independent facts about meaning, not a characterization of a principle that is partially constitutive of it. Heuristic maxims of the sort Quine has in mind may exist, but Wilson's original principle was not intended to be one of them.

Quine defends something more like Charity ten years later in his *Philosophy of Logic*, in which he seems to treat our interpreting others as endorsing the same logical laws as we do as more than just a mere heuristic. As he puts it:

> If a native is prepared to assent to some compound sentence but not to a constituent, this is a reason not to construe the construction as conjunction. . . . We impute our orthodox logic to him, or impose it on him, by translating his language to suit. We build the logic into our manual of translation. Nor is there cause here for apology. We have to base translation on some kind of evidence, and what better?[11]

While Quine applies to this principle primarily to the interpretation of the logical constants, it should be noted that he endorses the same reasoning for the other terms in the language. As he writes immediately after the passage quoted above:

> Being thus built into translation is not an exclusive trait of logic. If the natives are not prepared to assent to a certain sentence in the rain, then equally we have reason not to translate the sentence as "It is raining." Naturally the native's unreadiness to assent to a certain sentence gives us reason not to construe the sentence as saying something whose truth would be obvious to the native at the time. Data of this sort are all we have to go on when we try to decipher a language on the basis of verbal behavior in observable circumstances.

However, what Quine is defending here is still something different from Charity. In particular, the maxim of translation he has in mind here is better characterized, and characterized by Quine on the same page, as "Save the obvious."

> It behooves us, in construing a strange language, to make the obvious sentences go over to English sentences that are true and, preferably, also obvious. . . . Now this cannon—"Save the obvious"—is sufficient to settle, in point of truth value anyway, our translations of *some* of the sentences in just about every little branch of knowledge or discourse; for some of them are pretty sure to qualify as obvious outright (like, "1+1=2") or obvious in particular circumstances (like, "It is raining").

Saving the obvious would maximize the truth of a particular subclass of the agent's beliefs, namely those that "everyone, nearly enough, will unhesitatingly assent to," since this is what Quine means by "obvious" in the "ordinary behavioral sense."[12] Nevertheless, while Quine's new maxim will produce results that will often agree with those recommended by the Principle of Charity (since the truths that are obvious in Quine's sense will usually be heavily entrenched), the two principles are still distinct. Quine's principle seems to have little to say about the interpretation of the terms found in our non-obvious beliefs, and Charity may require giving up an "obvious" truth if it conflicts with a sufficient number of non-obvious truths. Quine's maxim has a foundationalist character, while Charity is more explicitly holistic.

Further tensions in how one should conceive of Charity are manifested in Davidson's work, since Davidson often fails to clearly distinguish the Principle from his views about the constitutive role of rationality in belief ascriptions. Davidson alternately suggests that the interpreter must maximize true beliefs and maximize rationality, claiming that we must interpret speakers as believing both the truth and what they rationally ought to believe given their other beliefs.[13] Davidson claims that "if we are intelligibly to attribute attitudes and beliefs, or usefully describe motions as behavior, then we are committed to finding, in the pattern of behavior, belief, and desire, a large degree of rationality and consistency."[14] Wilson's principle, however, puts no such requirement on interpretation, and Charity should not be confused with such rationality constraints.[15]

Davidson's running of Charity and rationality constraints together is exacerbated by his tendency to appeal to both in similar contexts, and provide each with the same sort of justification. For instance, Davidson combines both requirements when he argues that: "If we cannot find a way to interpret the utterances and other behavior of a creature as revealing a set of beliefs largely consistent and true by our own standards, we have no reason to count the creature as rational, as having beliefs, or as saying anything."[16] Davidson's commitment to finding a rational pattern in the interpretee's behavior may be defensible,[17] but it is clearly distinct from Wilson's Principle. One major difference between Charity and the rationality principle is that the latter is not only meant to help us decide which semantic values we should assign to the words in the sentences held true, but also meant to play a role in determining which sentences the interpretee should be viewed as holding true in the first place.

Rationality principles guide interpretations with maxims such as "*X* could never believe *P*," "In situation *S*, *X* would believe *P*," or "If *X* believes *P* and *Q*, then *X* would not believe *R*."[18] Charity, by contrast, has more to do with the move from uninterpreted sentences held true to the interpreted sentences. While Davidson treats Charity as more than a mere heuristic, his tacit equation of it with his own views about the constitutive role of rationality further blurs the nature of the Principle and makes it seem more like a general maxim guiding interpretation. Indeed, given that treating the speaker as rational may seem like a more plausible heuristic than treating him as a believer of truths, it can seem as if moving from Truth to Rationality is an attempt to make Charity more psychologically realistic.

Quine and Davidson have thus left many with the impression that Charity is a general principle guiding our interpretive practices that continuously enjoins us to assign true and/or rational beliefs to the interpretee. The Principle has thus come to be seen as a description of our actual interpretive practice: an account of how we generally go about ascribing beliefs to other people. Consequently, it is not surprising that, for example, Alvin Goldman, treats Charity as on par with other candidates for an explanation of what guides our interpretive procedures. Potential rivals to Charity would thus include the "Theory" theory (which claims that we interpret others using an implicit theory of human psychological processes) and the "Simulation" theory (which claims that we interpret others imagining what we would believe were we in their place).[19] Viewed in such a fashion, it is not surprising that Goldman finds the Principle implausible. After all, we frequently have no trouble ascribing false or even inconsistent sets of beliefs to those we interpret. People frequently fail to draw the consequences of their beliefs and often fail to perceive inconsistencies among them, and it may be psychologically, if not physically, impossible for them to do otherwise.[20]

Goldman is part of a long line of critics that have considered the Principle of Charity to be refuted by the fact that we frequently, and justifiably, attribute beliefs to speakers that stray considerably from any ideal of rationality or veridicality. In response to such critics, many have tried to salvage something from the Principle by:

(1) Weakening the constraint to something like Grandy's "Principle of Humanity."

(2) Requiring only a minimal amount of truth or rationality that can be uncontroversially found among most speakers.

(3) Treating the Principle as just a useful heuristic that may not work in all cases.

(4) Some combination of the above.[21]

However, such modifications, like the criticisms they respond to, leave Wilson's original formulation of Charity looking like a misguided and psychologically unrealistic attempt to describe our interpretive practices.

Fortunately, the Principle of Charity need not be modified in any of these ways. The facts about our interpretive practices frequently taken to suggest otherwise do not automatically tell against the general version of the Principle formulated above. Since Charity is not a heuristic meant to tell us whether or not someone holds a particular sentence true, considerations about how we actually go about interpreting others are unlikely to confirm or falsify it.[22] The simulation theory, for instance, suggests (roughly) that we interpret others by putting ourselves in their place, but it says nothing about the content of the attitudes we would have in their place, and this is precisely where the Principle of Charity is relevant. The simulation theory has a lot to say about what beliefs and desires we should attribute to a person were they to suddenly encounter a bear while hiking in the wilderness, but it is not concerned with the meaning of, say, "bear" in that person's language. Charity, on the other hand, is concerned with this later question, and has nothing to say about what attitudes we should ascribe to the startled hiker. The truth of the Principle of Charity is compatible with both "Simulation" and "Theory" theories of how we interpret others, and so the sorts of considerations brought to bear in the debate between them are not relevant to the Principle's evaluation.

IV. ELABORATION AND DEFENSE OF THE PRINCIPLE[23]

Even as an account of the semantic values of our words, however, the Principle of Charity has been criticized for systematically misidentifying the truth conditions of our utterances. This criticism, if true, would be damning, since helping to determine the truth conditions of our utterances is precisely the point of the Principle. If Charity is meant to be part of an account of what determines the semantic values of our words (rather than a mere interpretive heuristic), cases where it misidentifies the truth conditions of our utterances would be *counterexamples* rather than simply *exceptions*. Consequently, these purported counterexamples will be the focus of this section.

One group of such purported counterexamples suggests that a large number of common sense errors associated with misidentification are incompatible with the Principle. For instance, if, when walking down the street, I typically identify the condensation that falls from air conditioners onto my head as "rain," Charity would seem to dictate that the truth conditions for my sentence "It is raining" should include such condensation. After all, such an interpretation would seem to maximize the truth of my commitments by making an otherwise false statement true. However, we typically (and with some justification) take such claims to be mistaken, and so Charity seems to get the truth conditions of my utterance wrong.

Another, possibly more important, range of counterexamples relate to the Principle's compatibility with what is typically referred to as "semantic externalism." Since Charity maximizes the truth of our beliefs in a way that seems reminiscent of theories that tied the referents of names to clusters of descriptions, the cases that were counterexamples to such cluster theories would also seem to be counterexamples to Charity. Kripke, for instance, argues that we could still refer to Gödel by "Gödel" even if all our Gödel-beliefs were false of Gödel and true of someone else, and many take such cases to suggest that (roughly) our names refer to whoever their usage can ultimately be traced back to, whether or not the beliefs we associate with the name are true of that person.

Similar problems arise from the "social externalism" associated with the work of Tyler Burge. According to social externalists, what our terms mean is a function of how that term is used in our community, whether or not we have fully mastered this communal usage. However, there seems no reason to think that such *socially* determined referents need maximize the truth of any particular *individual's* belief set. Burge argues, for instance, that we would typically (and correctly) treat a speaker as referring to *arthritis* by “arthritis” even if he had the idiosyncratic belief that one could get arthritis in one’s thigh. However, Charity would seem to suggest that, since the speaker has a set of beliefs that were mostly true of arthritis, but entirely true of “tharthritis” (a set of diseases including both arthritis and various pains that affect the joints), we should treat him as referring to *tharthritis* by “arthritis.” Charity thus seems to be in tension with social externalism as well.[24]

However, the Principle is comparatively resistant to such purported counterexamples provided that we remember that:

(1) Charity is meant to “maximize” the truth of the speaker’s commitments in a way that is not strictly numerical.

(2) Charity is a global (or holistic) principle whose local application is often misleading.

(3) Many of the commitments that Charity deals with, indeed some of the most deeply held ones, are rarely made explicit.

Understood in the light of these reminders, Charity can accommodate purported counterexamples of the sorts mentioned above.

For instance, while putting the condensation from various air conditioners into the extension of “rain” would make a few more of my “It is raining” utterances true, it would be uncharitable because it would falsify much more deeply entrenched commitments of mine. Such commitments would include “Rain doesn’t come from air conditioners,” and “That’s not rain” (said of the condensation that I actually see falling off the conditioner). The impression that Charity is incompatible with the existence of commonplace misidentification results from focusing on the Principle’s improper local use rather than its proper global application. Once the rest of one’s (often implicit) commitments are brought into play, Charity can accommodate attributions of commonplace errors.[25] Much the same can be said of the larger issue of how Charity is compatible with the sort of causal/historical externalism focused on by Kripke, Putnam, and Donnellan. With such cases we must remember both that the Principle’s application is holistic and that many of our most central commitments are implicitly held.

For instance, we plausibly have an implicit commitment to perception, memory, and testimony having a causal structure.[26] While we may not *explicitly* think that, say, our memories are about those events that are causally responsible for them, such an understanding is manifested in our practices, and thus forms part of the larger set of commitments against which our thoughts acquire their contents. If we discover that the person we took a memory of ours to be a memory of could not have been causally responsible for that memory, we typically conclude that we must have been thinking of someone else. We know that our memories must be causally connected to what they are memories of, even if this knowledge is not explicitly represented. We would typically assent to things like, "If you remember meeting Peter, you must have met Peter in the past," because we have an implicit understanding of our thoughts and utterances being part of a causally structured informational system.

The thought experiments that writers such as Kripke, Putnam, and Donnellan muster in support of their accounts of intentionality can be understood as illustrating our implicit commitment to this informational system. If we did not already have such implicit commitments, we would not find it obvious that we could not be referring to, say, some causally isolated hermit by "Thales," no matter how many of our Thales-beliefs would have been true of him. Indeed, such reactions are partially *constitutive* of our commitment to the informational system. As a result, one could argue that everybody already knows some version of the causal/historical picture of intentionality, and that they just need prompting to recollect it.[27] Since the commitments to these aspects of the informational system are heavily entrenched and tied up with most of our beliefs about the world, it would not be charitable to give them up just to preserve the truth of a small set of Thales-beliefs. Cases like that of Thales or Gödel are thus not incompatible with the Principle of Charity, they only highlight how the commitments that must be taken into account extend beyond those that the speaker might explicitly manifest in his utterances.

The same sort of explanation can be given for the compatibility of Charity with social externalism. Not only do we have an implicit understanding of how our perceptions and memories are causally related to the world around us, we have an implicit understanding of how language and linguistic communication function, an understanding that involves our sharing a language with our peers. Our implicit commitment to this model of language is manifested not only in our content attributions, but also in our deference behavior. For example, when our usage is corrected, our assumption that we have made a certain sort of mistake—misusing a word, and saying something false (as opposed to using a nonstandard word and saying something misleading)—is explained by, and is partially constitutive of, our having an implicit commitment to such a picture of language.[28] On the account suggested above, the idea is *not* that the content of, say, my "arthritis" belief makes explicit reference to the usage of the community causally responsible for my use of "arthritis." That is to say, it is not a simple metalinguistic account in which by "arthritis" I just mean whatever my community means by "arthritis." Rather, the idea is that the belief's content simply involves *arthritis*, but that it does so partially in virtue of my own implicit commitments tying my usage to that of my community. Such an account, while accepting that Bert means *arthritis* by "arthritis," could still consider itself *methodologically* individualistic because the importance of the social environment stems from the speaker's own implicit commitments. There is no *essential* (as opposed to mediated) reference to the social context.

Not only do we make ascriptions and defer to correction in a way that manifests a commitment to shared meanings, but our practice of forming beliefs based on the testimony of others also manifests such commitments. For instance, my forming the belief that Calcium supplements promote arthritis simply because I hear a doctor say “Calcium supplements promote arthritis” manifests my implicit commitment to my meaning the same thing by “Calcium” and “arthritis” as doctors do. Much the same could be said of object-level beliefs of mine such as “Many doctors are trying to find a new treatment for arthritis.” As long as similar such commitments are collectively more entrenched for Bert than his simple belief that he has arthritis in his thigh, Charity will recommend treating him as meaning *arthritis* rather than *tharthritis* by “arthritis.”

These arguments that Charity is compatible with semantic externalism have relied heavily on the idea that many of our most important commitments are implicit. In this respect, the view developed here bears some resemblance to John Searle’s account of how our thoughts acquire their contents against a nonrepresentational background of implicit commitments. Nevertheless, the version of Charity defended here should not be equated with Searle’s “holistic internalism” according to which what is “in the head” is “entirely sufficient to determine the identity of each of our intentional states.”[29] In particular, while the view defended here is methodologically individualistic, it is not internalistic in Searle’s sense, and this gives it a number of advantages over Searle’s view.

Searle’s commitment to internalism forces him to try to specify the contents of our thoughts in a perfectly *general* fashion, with the only *particulars* referred to being our mental states themselves. While Searle’s attempt to do this is more sophisticated than earlier internalist theories, it is ultimately no more successful. For example, Searle attempts to deal with the Putnam’s Twin-Earth case by arguing that “water” is defined indexically, as “whatever is identical in structure with the stuff causing *this* visual experience [said while looking at a glass of water], whatever that structure is.”[30] This analysis not only builds too much of the Background into the explicit content,[31] but also is (in virtue of doing so) subject to obvious counterexamples. For instance, Searle’s analysis suggests that what we mean by “water” would change whenever we found ourselves looking at a glass of something we took to be water but which was not, in fact, H_2O. The indexical description that Searle focuses on is only one of *many* commitments that we take on with respect to a term like “water.” Furthermore, it may be one that we would be willing to give up if it turned out to be incompatible with other more deeply entrenched commitments (such as that we have had contact with many glasses of water prior to our present visual experience of the glass and its contents). Consequently, the “indexical definition” should not be *identified* with the content (indeed, no single commitment, or subset of them, should).[32]

In much the same way, Searle attempts to build background commitments directly into explicitly metalinguistic intentional contents when he tries to account for the social aspect of language with observations such as, “Often the only identifying description one associates with a name *N* is simply the object called *N* in my community or by my interlocutors.”[33] Such formulations, in addition to being implausibly metalinguistic, are susceptible to obvious counterexamples (the generally accepted usage has changed since I learned the name, I make an unnoticed slip of the tongue when pronouncing the word, etc.). Just as his account of content in the Twin-Earth cases did, Searle’s formulation neglects the point that our background commitments should be in the *background.* They are just one of the many commitments that determine what a term means, and so should not be “forgrounded” in a way that identifies them with the term’s content.

Awareness of implicit commitments can help us understand both how the contents of our thoughts are a function of our commitments, and why they should not simply be identified with them. The semantic values of one’s terms will depend not only on what is in one’s head, but also on one’s *actual context,* because one’s context often determines which commitments are, and are not, compatible with each other. Which commitments will be part of such a maximal and consistent set (and thus what one should be understood as referring to) can vary from context to context, and so what is in the head is not “entirely sufficient to determine the identity of each of our intentional states.” Identical sets of commitments can pick out different semantic values in different contexts. This is more than the claim that our words may refer to different things in different contexts (which internalists can easily admit), but rather that one’s *actual* context determines what sorts of thing our terms will pick out in all *possible* contexts. One can thus recognize the importance of these Background commitments and still insist that while they help determine externalistic contents, they do not show up in the contents themselves.

Searle is certainly right to think that our implicit commitments can help explain why our thoughts frequently have the world-involving contents that they do. Unfortunately, he is unable to fully exploit this insight because the need for generality stemming from his internalism leads him to build these implicit commitments explicitly into the contents of our thoughts.

Finally, a component essential to this defense of Charity, holism, needs some more clarification. Charity suggests that one take the semantic values of a speaker’s terms to be those that maximize the (weighted) amount of truth in the sentences the speaker is committed to. That is to say, Charity characterizes a function that takes the total set of sentences held true as its input and gives beliefs and semantic vales as its output. Such an account of content will undoubtedly be holistic: a term has the semantic value it does because of the role that value plays in contributing, either directly or indirectly, to the truth of countless beliefs. Holistic theories of meaning and content have come under a good deal criticism lately, and if Charity commits one to holism, then any problems for holism will be problems for Charity as well.

One of the most frequently cited problems with holism is that, since no two people have precisely the same beliefs, and holism entails that what one's terms mean is a function of all of one's beliefs, then holism would seem to suggest that no two people mean quite the same thing by any of their terms. If what one meant by, say, "elephant" were determined by all of one's elephant-beliefs, then someone who believed "all elephants are gray" would seem to mean something different by "elephant" than someone who believed that some of them were brown. Consequently, a translation from one of their languages into the other's couldn't be completely accurate, and their ascriptions of belief to each other would be, strictly speaking, false. Complete communication would thus be effectively impossible and we would never fully grasp the content of anyone else's words or thoughts.

Since most people take it to be simply *obvious* that there can be differences of belief without differences of meaning, such purported consequences of holism are extremely worrying. However, these criticisms assume that the holist is committed to what will here be called "semantic instability," namely, that any change in one's attitude towards a sentence will change the meanings of the terms contained in it and the contents of the associated beliefs. The prospect of such semantic instability lies at the heart of people's intuitive difficulties with holism.[34]

Fortunately, it isn't at all obvious that every holist must embrace instability. After all, holism only requires that the content of one's beliefs *depend upon* or be a *function of* one's other beliefs, and *this* claim need not commit one to semantic instability. Holism would only entail instability if it required that the function from beliefs to contents be one-to-one, and there is nothing in the general characterization of holism that requires that the type of function involved be restricted in this way. If the function in question were many-to-one, then content could be comparatively stable through changes in belief.[35] Consequently, there is no reason why *simply* being a holist about meaning or content need commit one to semantic instability or any of the problems associated with it.

Even if, say, countless beliefs played some role in determining a term's referent, there is little reason to think that a change in one (or even a considerable number) of these beliefs will change what is referred to. The truth of two different sets of elephant-beliefs may, in spite of their differences, be maximized by precisely the same set of objects. The defender of Charity can thus allow there to be changes in belief without changes in meaning, and hence accept holism while avoiding a commitment to semantic instability.[36] Charity's appeal to holism does not, then, present any serious problems for it.

V. MOTIVATING THE PRINCIPLE

Still, even if it is comparatively resistant to counterexamples, why should one accept the Principle? What reason do we have for thinking that it is more than a generalization that seems plausible, but which (like many plausible generalizations) turns out to be false when applied across the board? A defense of the Principle should also explain *why* it is true, but such explanations are surprisingly hard to find in the work of those who appeal to it.

Wilson doesn't give much of an argument for Charity other than appealing to the sort of intuitive considerations generally marshaled in favor of descriptive accounts of reference. That is to say, if a speaker makes five statements involving the name "Peter" and four turn out to be true if "Peter" refers to *A*, and just two of them are true if "Peter" refers to *B*, then it seems plausible to think that he is talking about *A* rather than *B*.[37] The lack of any serious attempt to justify the Principle is not surprising, since Charity is not the primary topic of Wilson's paper. Quine also initially treats Charity in an aside, and his more extended argument that the interpreter should save the obvious amounts to more of a defense of simulation accounts of interpretation than Charity itself.

Davidson, on the other hand, frequently tries to justify his appeals to Charity. Davidson's defense of the Principle amounts to a series of arguments to the effect that "If we cannot find a way to interpret the utterances and other behavior of a creature as revealing a set of beliefs largely consistent and true by our own standards, we have no reason to count that creature as rational, as having beliefs, or as saying anything."[38] Charity is, then, presented as a *precondition* of interpretation and meaning. As Davidson famously puts it, "Charity is forced on us; whether we like it or not, if we want to understand others, we must count them right in most matters."[39] If our beliefs weren't mostly true, we wouldn't have beliefs at all.

Davidson's claim that interpretation presupposes the attribution of predominantly true beliefs is highly controversial. Nevertheless, I will not discuss Davidson's defense of this claim in much detail because, while Davidson often presents his arguments for the general veridicality of our beliefs as if they were a defense of the Principle of Charity, they do not directly support it. Davidson's arguments all focus on the purported unintelligibility of massive error and consequent claim that most of our beliefs must be true, but even *if* such an anti-skeptical claim were a *consequence* of the Principle of Charity, it should not be *identified* with it. Charity requires that we pick the assignment of semantic values that gives the interpretee *the most* true beliefs, and this is significantly different from the requirement that we assign semantic values that give the interpretee *mostly* true beliefs. There could be *many* assignments of semantic values that could leave the interpretee with mostly true beliefs, but Charity requires picking out only those assignments that maximize the amount of true ones. Davidson's defense of Charity, however, gives little (if any) justification for picking out an assignment that maximizes the amount of true beliefs over a non-maximal one that still produces a set of beliefs that are mostly true.[40] Even if we could not understand someone unless they had beliefs that were mostly true, Davidson gives no reason for maximizing truth beyond this lower limit.

Furthermore, just as an assignment that gives mostly true beliefs need not be the one that gives the most true beliefs, an assignment that maximizes the true beliefs held by a speaker need not leave him with mostly true beliefs. Wilson's principle suggests that, all else being equal, if we have fifteen Aristotle-beliefs, seven of which are true of a certain man, and no more than six of which are true of anyone else, then "Aristotle" will refer to that man. It may thus be that one's Aristotle-beliefs are not *mostly* true of anyone, and the referent that maximizes the truth of the set still makes the majority of the beliefs false. This is one way in which Charity differs from the sorts of cluster theory of proper names criticized by Kripke. Such cluster theories would seem to suggest that, if a term has a referent at all, a majority of the descriptions associated with it must be true.[41]

This gap between maximizing truth and having mostly true beliefs becomes even wider once we realize that the type of maximization involved is sensitive to how central the various commitments are. A heavily entrenched minority might be preserved at the expense of a more peripheral majority. For instance, we might give up almost all of our Thales beliefs to preserve our commitment to the informational system. Consequently, an assignment of semantic values that maximizes the weighted set may be even less likely to make the majority of one's commitments true. It thus seems that Davidson's anti-skeptical arguments, even if they are sound, do not provide any justification for the Principle of Charity. They just support the independent position that most of our beliefs must be true.[42]

David Lewis provides a defense of Charity that takes a slightly different approach than Davidson's. Rather than arguing that Charity is a precondition of interpretation, Lewis claims that Charity is one of "the fundamental principles of our common-sense theory of persons" that "implicitly define such concepts as belief, desire and meaning."[43] That is to say, Charity is justified in terms of its being partially constitutive of the very concepts it deals with. However, Lewis makes little effort to justify his claim that Charity is part of such an implicit definition, so the question of justification is just pushed back a level. Furthermore, while Lewis claims that Charity requires that the interpretee "be represented as believing what he ought to believe, and desiring what he ought to desire," for Lewis, this amounts to his believing and desiring "what *we* believe, or perhaps what we would a have believed in his place, and…what we desire, or perhaps what we would have desired in his place."[44] The Principle of Charity is thus, for Lewis, very much like Grandy's Principle of

Humanity in that it suggests that we put ourselves in the place of the interpretee. Furthermore, Lewis only requires that the interpretee more or less conform to the constraint. Charity would, on such an account, be approached, but it need not be satisfied. Indeed, this "more or less" qualification might work for a heuristic principle such as Grandy's, but it is hard to see how Charity could "more or less" determine semantic value.

Since Wilson, Quine, Davidson, and Lewis do not seem to provide any compelling reason to accept it, the Principle of Charity still stands in need of justification. Fortunately, Charity can be justified, and it will be argued here that it follows fairly directly from constraints internal to the practice of interpretation.

First of all, one should note that acting in accordance with the Principle seems unavoidable in the case of *self*-interpretation. From the first-person perspective, the questions of what we believe and how we take the world to be are indistinguishable.[45] As a result, we cannot help but view each of our current commitments as true. Our self-interpretations may thus seem to endorse Charity in a very strong sense. However, even if we treat each of our commitments as true individually, we realize that our doing so is largely the result of our own epistemic limitations.

Nevertheless, even when such epistemic limitations are removed, Charity is reflected in how we determine the semantic values of our own terms. There may be no assignment of semantic values to our own terms that will make all of our commitments true, and two (or more) of our commitments associated with a word conflict when no available semantic value for that term will make them both true. Our epistemic situation can be understood as idealized to the extent that we become aware of such conflicts and go on to settle them. Resolving such conflicts among our commitments typically involves rejecting some commitments as mistaken, and holding on to others. The commitments we hold on to are, by definition, more entrenched for us than those that we give up. However, this means that even under ideal conditions we cannot help but interpret ourselves in a way such that our most heavily weighed commitments are taken to be true. We will always settle such conflicts in a fashion that preserves the truth of the most entrenched commitments. However, if we interpret ourselves, even under conditions of complete idealization, in a way such that our most deeply held commitments are true, then we cannot help but pick out as the semantic values of our terms those values that maximize the truth of our commitments. That is to say, even under epistemically ideal conditions, we cannot help but apply Charity to ourselves.[46] The general structure of our epistemic investigations into the world will lead us to treat our most entrenched commitments as true through the course of inquiry.

Charity is, then, an essential characteristic of self-interpretation. Do we have, however, any reason to think that it need be adopted in our interpretation of others, or in their interpretation of ourselves? That is to say, even if Charity is unavoidable from the first-person perspective, it need not be from the third. So why should it be endorsed from the third-person perspective that we have when we interpret others? Why should we assume that our idealized self-interpretation must be correct?[47]

Fortunately, Charity can be seen as a consequence of the fact that the primary goal of interpretation is *understanding*. Understanding someone requires discovering how they see the world that we share with them, and a good interpretation will thus allow us to see the world through *the interpretee's* eyes. However, we don't want to lose ourselves in the interpretee's perspective either. We want to capture her perspective as a perspective on *the world*, not just become immersed in her notional world. Understanding involves mapping the interpretee's notional world onto the real one, and such a mapping will involve characterizing some of her commitments as true and others as false.[48] Doing so in a way that still captures her perspective will, however, involve trying to make our interpretation of her match what her self-interpretation would be if she were aware of all that we were.[49] The idealization ensures that it is a perspective on *the world* that we end up with, and it is the fact that it is the agent's own commitments that are idealized that ensures that it is *her perspective* that we are capturing. Capturing the interpretee's perspective on the world thus involves trying to understand the interpretee as she would, ideally, understand herself.

Consequently, when trying to decide what the interpretee's terms refer to, the interpreter should pick out the referents that the interpretee would pick out if she were better informed. That is to say, if the interpretee has a set of commitments associated with a name, not all of which could be true of any one object, the interpreter should treat her as referring to whichever object the interpretee would decide that the term referred to if she had all the relevant information. In short, when deciding which of two competing commitments to characterize as mistaken, the interpreter should try to discern which of the two the *interpretee* would give up were she to become aware of the conflict.[50]

This procedure respects the commitment to capture the interpretee's point of view to a maximal extent compatible with recognizing that it is a view of an independent world and not just a creature of the interpretee's imagination. We should interpret the speaker as meaning what she intends to mean, and we can take her intentions to be clarified by the revisions she would make if the various tensions among her commitments were made manifest to her (leaving, of course, some room for dishonesty, bad faith, misinformation, etc.).

This respect for the interpretee's idealized self-interpretation quickly leads, of course, to the endorsement of a form of the Principle of Charity. Given a set of commitments associated with a word, not all of which can be true, this account suggests that the ones we should characterize as mistaken are the ones that the interpretee would give up were she aware of the tensions among them. As a result, the members of the set that turn out to be true are those that the speaker would not give up even were she better informed. However, how willing the interpretee is to give up a given commitment is a reflection of how entrenched that commitment is. Consequently, from this procedure of understanding the speaker as she would understand herself, we end up with an interpretation according to which the most collectively entrenched consistent set of commitments associated with her words will turn out to be true. It follows that Charity, the requirement that we assign semantic values to the interpretee's words that maximize the truth of her commitments, follows directly from the fact that we should try to understand the interpretee as she would understand herself.[51] Charity is thus not simply a matter of being generous to the interpretee; it follows from constraints internal to the interpretational project. It is something that makes interpretation what it is, not merely something that makes interpretations possible.

VI. CONCLUSION

Charity, so understood, represents a type of minimal individualism that is closely tied to self-knowledge and first-person authority. We can be mistaken in the use of our words, but these mistakes must be explained in terms of factors that we would endorse (i.e., other commitments that we take to be more central).[52] Purely physicalistic or social accounts of meaning would not, for instance, satisfy this constraint.[53] The speaker's self-interpretation can be treated as wrong if we can appeal to facts about conflicts among her commitments that she is unaware of, but barring such facts, it should be accepted. Consequently, the speaker's idealized self-interpretation is constitutive of what she means. To the extent that an interpretation violates the Principle of Charity, it will fail to capture the interpretee's point of view and fail to provide the type of understanding that interpretation seeks.

Interpretational accounts of meaning are thus not essentially third-person in the way they are frequently criticized for being.[54] Charity is a constraint that captures the first-person perspective and is essential to the interpretive process. Charity is unavoidable in the first person case, and endorsing it in our interpretations of others amounts to a commitment to capturing, from the third-person starting point, their first-personal point of view.

ENDNOTES

I'd like to thank Robert Brandom, Joe Camp, Johnathan Cohen, Robert Epperson, Joshua Finkler, Mark McCullaugh, John McDowell, Richard Manning, Ram Neta, Arthur Ripstein, Roy Sorensen, Gerald Vision, and members of the audience at York University, the University of Guelph, the 1998 APA Central division meeting, the 22nd annual Mid-South philosophy conference, and the 2001 meeting of the Canadian Philosophical Association for comments on earlier versions of this material.

[1] Wilson (1959, 532).

[2] Ibid., 528.

[3] Ibid., 531.

[4] This is another sense in which "commitment" may differ from "belief." For a discussion of this, see Brandom (1994).
[5] This is stressed by Wilson himself (1959), 535.
[6] This notion of "entrenchment" is discussed in greater detail in Gärdenfors and Makinson (1988), and Rott (2000). The degree to which a given commitment is entrenched will typically reflect the strength of the evidence in favor of that commitment, but other less evidential factors can help determine the entrenchment of our commitments. Such other factors are discussed in greater detail in Jackman (1996; 1998a).
[7] In Jackman (1999a), I argue that what we intuitively take ourselves to mean varies in just this way. The fact that the judgments that determine the comparative entrenchment of our commitments are not well ordered may suggest that the notion of maximizing the truth of our weighted set of commitments will be an unattainable goal. However, while such lack of ordering may prevent us from "optimizing" the truth of the set (i.e., finding an interpretation that produces more truth than *any* other), it does not prevent maximization, which only requires that we not pick an interpretation that has an alternative that produces more truth than it. For a discussion of the difference between maximization and optimization, see Sen (2000). This choice of terminology is not mandatory, however, and Davidson argues that the fact that there are infinitely many commitments involved favors the use of "optimize" over "maximize" (Davidson 1975, 169). Davidson's preference here may result from his conception of the constitutive role of rationality in interpretation leading him to assume that our commitments will be well ordered.
[8] For a discussion of this, see Davidson (1973, 136).
[9] This use of "semantic value" draws on Dummett (1975) and Evans (1982). However, since I tend to understand the semantic value of some of our words in terms of what they refer to, I will occasionally speak in terms of referents rather than semantic values.
[10] Quine (1960, 59). See Evnine (1991, 104–105) for endorsement of this understanding of Charity.
[11] Quine (1970, 82).
[12] Quine (1970, 82). Quine then adds that "every logical truth is obvious, actually or potentially." This emphasis of Quine's on the logical truths is stressed in Davidson (1984, xvii, 136).
[13] For a discussion of these strands in Davidson's conception of Charity, see Goldman (1986), Devitt and Sterelny (1987), and Evnine (1991, 110–111). For further discussion of Davidson on this issue, see Jackman, (1996b), forthcoming.
[14] Davidson (1974c, 237).
[15] Of course, maximizing the truth of a set of commitments will automatically bring some increase to the rational coherence of the commitments, since a set of massively inconsistent beliefs could not contain very many truths. Nevertheless, the two projects are different, and the interpretation that maximizes the truth of the agent's beliefs will not necessarily be one that maximizes their rational coherence.
[16] See, for instance, Davidson (1973, 137; 1975, 159; and 1994, 232).
[17] It is defended extensively, for instance, by Davidson himself, and to a limited extent in Cherniak (1986), Dennett (1978; 1987), Lewis (1974), Loar (1981), and McDowell (1986).
[18] See, for instance, Davidson (1975, 159–160; 1985, 138). "Simulation" and "Theory" theories of interpretation endorse claims of a similar form.

19. See, for instance, Goldman (1989). Goldman himself endorses a version of the simulation theory, which he equates with an endorsement of Grandy's principle of Humanity. Quine seems to withdraw from a truth-driven notion of Charity towards something more like the "theory" and simulation theories not only in his "save the obvious" maxim, but also in his (1970b, 16–19; 1990a, 46; and 1990b, 158). Davidson (1984, xvii) may be seen as taking steps in this direction.

20. See, for instance, the discussion of the "Paradox-of-the-preface" in Goldman (1989, 12), and for an extensive criticism of the claim that any *robust* rationality constraints guide our interpretive practices based on the limits of our computational capacity, see Cherniak (1986).

21. For various forms of proposal (1), see Grandy (1973) and Lukes (1982). The Principle of Humanity suggests that we interpret others in the manner described by the simulation theory. Davidson seems to move in this direction in Davidson (1974b, 196), and see Evnine (1991, 103, 109) for a reading of Davidson where Charity is much more like Grandy's Principle of Humanity and (like Quine) requires no more than attributing to the interpretee a shared sense of what is obvious. For a clear example of (2), see, for instance, Cherniak (1986). Loar (1981) seems to assume that the more minimal conception is what Davidson actually endorses. See McDowell (1986) for a criticism of this assumption. Quine's "Save the obvious" maxim might also be interpreted as such a minimal version of Charity about truth. An instance of (3) can be seen in Hacking's downgrading of Charity and Rationality constraints to mere "commonsense rules of thumb that might, like all common sense, sometimes offer bad advice" (1975, 149–150). Grandy, by contrast, may in fact be moving to a proposal of type (4) when he characterizes Humanity not as a constitutive norm, but as a "pragmatic constraint" (1973, 443), which, if we were omniscient about physical and design facts, we wouldn't need. For a useful attempt to resist these four trends, see Ramberg (1989, chap. 6).

22. As a result, the Principle will not suggest that, say, an atheist should interpret all other speakers as atheists as well. There is nothing charitable in interpreting people who hold true sentences like "God exists" as atheists.

23. Much of this section will read like an attempt to "internalize" familiar forms of semantic externalism. This should not be surprising, since externalistic considerations have been what have lead many to question the plausibility of the Principle of Charity. The project of bringing semantic externalism within the methodologically individualistic restrictions of the Principle of Charity is also of independent interest.

24. See McGinn (1977) for an early and influential presentation of this view that semantic externalism presents problems for the Principle of Charity. The points about proper names should be familiar from Kripke (1972), and similar arguments relating to proper names and natural kinds can be found in Donnellan (1972) and Putnam (1975). The evidence in favor of "social" externalism (Burge, 1979) is considerably less conclusive, and some are willing to endorse Charity's purported recommendations here, and simply reject social externalism because of this sort of case. See, for instance, Bilgrami (1992).

25. Consequently we need not, *pace* Evnine (1991, 108), appeal to public meanings to account for Charity's compatibility with such cases of individual error. Indeed, given Davidson's occasional hostility to such social norms, such a suggestion seems highly problematic as an exposition of Davidson's views. For a discussion of Davidson's ambivalence on this issue, see Jackman (1998). It should thus also be clear that Charity is not something that should, *pace* Grandy (1973, 445), bear heavily on the

interpretation the speaker's reference of phrases like "The man drinking the martini," though it may be relevant for the semantic referents of the terms involved.

26. Of course, the Radical Interpreter cannot simply *assume* that speakers have such a commitment. However, if evidence for such a commitment is absent, one would also have evidence that the sorts of ascriptions Kripke calls our attention to would be inappropriate for such speakers.

27. This makes more palatable what Gareth Evans refers to as "Russell's principle," namely, the assumption that "a subject cannot make a judgment about something unless he knows which object his judgment is about" (Evans 1982, 89). For instance, Evans asks us to imagine a subject who briefly sees one ball rotating by itself on one day, and another on a later day, but "retains no memory of the first episode, because of a localized amnesia." If the subject were to reminisce about that shiny ball he once saw, he would be unable (if "asked which ball he is thinking about") to "produce any facts which would discriminate between the two." Evans draws from this the highly counterintuitive conclusion that such a subject could not have thoughts about the ball that he remembers (ibid., 90). However, Evans admits that the subject could specify the ball in question if he *reflected* on the role of memory in the informational system, but he claims that, "thinkers . . . will not in general resort to this [causal] way of identifying the object of their thought" (ibid., 117). Nevertheless, while the typical subject will not *appeal* to the object's role in the informational system, there is still a sense in which he implicitly "knows" that the object must play such a role, and can thus be said to know which object he is thinking of. When such implicit knowledge is brought into play, Russell's principle is much less restrictive. We can allow that *merely* being causally connected to an object isn't enough to have thoughts about it (and thus reject the "photograph model" that Evans opposes to Russell's principle), while still insisting that our implicit understanding of the informational system allows us to think of those items from which our memories derive.

28. If socially sensitive content ascriptions are justified by the presence of such an implicit picture of language, and these implicit commitments are manifested in such deference behavior, then it is of considerable importance that we do defer in such a fashion. Explanations of deference solely in terms of, say, a pragmatic desire to communicate effectively would, if true, suggest that there are no such implicit commitments, in which case non-individualistic content attributions would be much less plausible. For a discussion of this, and other aspects of methodological individualism, see Jackman (1996; 1998).

29. Searle (1991, 237). See also, Searle (1983), and Dreyfus (1991).

30. Searle (1983, 207). In his Twin-Earth case, Putnam (1975) argues that our counterparts on a planet whose water was phenomenologically just like ours but had a different chemical structure would mean something different by "water" than we do.

31. In spite of his own warnings against doing so. (Searle 1983, 19; 1991 230, 232–233.)

32. One should also note that the only way that Searle can account for the fact that I and my Twin-Earth counterpart have thoughts with different contents is by appealing to the fact that we are different people and thus that our thoughts make reference to different thought/experience tokenings. Such an account is, however, unable to explain twin-cases in which we think of the *same* experience as arising in different counterfactual situations. Just as Searle claims that I would have thoughts with the very same contents that I now have even if I were a brain in a vat (Searle 1983, 154),

he must also say that I would have thoughts with the very same contents if the "water" on this planet had actually been XYZ rather than H_2O.

[33] Searle (1983, 243–244, 250).

[34] For a much more exhaustive list of purported problems instability brings to holistic theories (including its apparently disastrous consequences for the intelligibility of disagreement, change of mind, psychological and scientific explanation), see Fodor and LePore (1992). This "instability thesis," and its relation to holism, is discussed in greater detail in Jackman (1999). The defense of holism sketched below is presented in considerably more detail in that paper.

[35] Consider, for instance, the claim that one's final letter grade in a class is a function of one's grades on one's exams and homeworks. The truth of this claim certainly doesn't entail that no two people could have the same final grade unless they had precisely the same score on all of their homeworks and exams. Nor does it entail that any change to one of one's homework grades will produce a change in one's final grade. The function from contributing grades to final grades is many-to-one, and thus allows a good deal of stability in the output in spite of the possibility of tremendous variation in the input.

[36] Some holistic theories of meaning, such as those that identify a thought's content with its inferential role (Block, 1986; Field, 1977; Harman, 1973; and Sellars, 1974) may, of course, be committed to semantic instability.

[37] Wilson (1959, 531–532).

[38] Davidson (1973, 137).

[39] Davidson (1974b, 197). See also (1974a, 153; 1975, 168). He makes a similar claim about the necessity of finding a large degree of rationality in (1975, 159; and 1994, 232). Ramberg (1989, 70, 77) presents this sort of defense of Charity as well.

[40] Failing to choose the non-maximal interpretation may, of course, violate the general principles of Davidsonian interpretavism, but the fact remains that doing so is perfectly compatible with Davidson's purported *justification* of those principles.

[41] See the formulation in Kripke (1972, 64–65, 71). However, it should be noted that the most prominent cluster accounts do not, strictly speaking, require that the referent satisfy a majority of the descriptions associated with a name (Searle 1958, 94; Strawson 1959, 191). Davidson himself, however, occasionally seems to suppose that if we mean anything at all by a term, that most of the beliefs we associate with it must be true (e.g.,1975, 168). Wilson by contrast never suggests that most the beliefs associated with a name must be true, though he occasionally claims that *all* couldn't be false (1959, 527, 533, 535).

[42] The same point holds for appeals to evolutionary theory to show that most of our beliefs must be true. For a discussion of such views, see Dennett (1987). Such arguments, even if they were sound, would not be enough to establish the Principle of Charity.

[43] Lewis (1974, 112). Evnine (1991, 112) claims that Davidson himself endorses this approach.

[44] Lewis (1974, 112).

[45] By "self-interpretation" I mean here our understanding of our *current* mental states. Our understanding of our *past* selves is in many respects more like our understanding of another person. Furthermore, while one can treat even one's current beliefs as less "transparent" to the world, but this amounts to taking a third-person perspective on one's current mental states. For a discussion of this, see, Moran (1988, 1994).

[46] Under non-ideal conditions, of course, we may fail to come up with the most charitable interpretation of ourselves (by failing to grasp our commitments globally, to work

out their full consequences, etc.). However, even in these cases we are *trying* to interpret ourselves charitably; it is only our lack of information that keeps us from doing so successfully. Charity can still be understood as guiding our interpretation, even if we lack the means to be successfully guided.

47. One could argue that a general endorsement of Charity would follow from adopting a simulation account of interpretation. Charity is unavoidable in first-person case, so if we treat as correct the commitments of the interpretee that we would endorse were we in *her* place, we should treat her most heavily entrenched commitments as true. However, simulation is meant as a *description* of our interpretive practices, it is not meant to be a constitutive account of meaning. For a discussion of this, see Stich (1981; 1984). Consequently, it could only show that we *do* interpret others in accordance with Charity, not that we *are correct* in doing so.

48. I will, for simplicity's sake, be assuming that the interpreter's conception of the world is an accurate one, and so will be treating preserving agreement and truth as if they were the same. Otherwise the topic could be discussed in terms of mapping their world onto our own and maximizing agreement.

49. Some might argue that determining what someone's commitments would be if she had information that she currently lacks is hardly a way of seeing the world through her eyes. One should not, for instance, interpret me as believing that something is a salamander when I've said "that it is a lizard," simply because one knows that it *is* a salamander, and that I would change my commitment upon learning this. However, selves are temporally extended, and one's "self-interpretation" can refer to either one's interpretation of one's current mental states, or one's interpretation of one's past mental states. When I claim that interpretation requires "trying to make our interpretation match what her self interpretation would be if she were aware of all that we were," the type of idealization involved creates a context that shares properties of both cases. If I say, "that is a lizard," but would learn in the idealized context that "that" is a salamander, my interpretation of my *own* original utterance will involve the false claim that a particular salamander is a lizard. Third-person interpretation should try to capture this, not what the idealized agent at would take to be their *current* beliefs. One should also remember that Charity is about the reference of our terms, not what "thin" commitments we happen to have. The reference of "that" and "lizard" will be the same at both contexts, it is only one's attitude towards the claim that that *is* a lizard that will change.

50. For ease of exposition, these conflicts (and their resolution) are described as taking place at a local level. A more accurate, but more cumbersome account would reflect the fact that the relevant decisions would ideally be made globally.

51. This does not commit one to any foundational or reductive project where third-person charity can be *derived* from something completely independent of it. Claiming that self-interpretation grounds third-person Charity doesn't require that there could be self-interpretation without third-person interpretation (or commitments without self-interpretation). Self-interpretation could have this role even if no one who was not the interpreter of another could be a self-interpreter.

52. This minimal, or "methodological" individualism is discussed further in Jackman (1996, 1998). The type of semantic self-knowledge involved comes from the partially constitutive role the interpretee's point of view has for what she means. On such an account, first-person authority is explained in terms of constraints on interpretation, and self-knowledge is understood in terms of first-person authority. For a related discussion, see Bilgrami (1999). This contrasts with an alternate mode of explanation in

which first person authority is explained in terms of self-knowledge and self-knowledge is explained in terms of privileged access. It is this latter order of explanation that can make the relationship between externalism and self-knowledge seem problematic, since how we could have privileged access to these external factors seems mysterious. No such problems exist with the former order of explanation, since we are still "authoritative" about whether or not such external factors are relevant. For instance, if we are not committed to social usage being relevant to what we mean by a particular word, then such usage is not relevant.

53. See Devitt (1996) and Devitt and Sterelny (1987) for the former, and Kripke (1982) for the latter.

54. See, for instance, Taylor (1980) and Searle (1987). For a discussion of this, see McDowell (1994).

BIBLIOGRAPHY

Bilgrami, A. 1992. *Belief and Meaning*. Cambridge, Mass.: Blackwell.

————. 1999. "Why Is Self-Knowledge Different from Other Kinds of Knowledge?" In *The Philosophy of Donald Davidson*. Edited by L. E. Hahn. Chicago: Open Court, 1999.

Block, N. 1986. "Advertisement for a Semantics for Psychology." In *Midwest Studies in Philosophy X: Studies in the Philosophy of Mind*. Edited by French, Uehling and Wettstein. Minneapolis, University of Minnesota Press.

Brandom, R. 1994. *Making it Explicit*. Cambridge, Mass.: Harvard University Press.

Burge, T. 1979. "Individualism and the Mental." In *Midwest Studies in Philosophy IV: Studies in Metaphysics*. Edited by French, Uehling, and Wettstein. Minneapolis, University of Minnesota Press.

Cherniak, C. 1986. *Minimal Rationality*. Cambridge, Mass.: MIT Press.

Davidson, D. 1973. "Radical Interpretation." In Davidson, 1984.

————. 1974a. "Belief and the Basis of Meaning." In Davidson, 1984.

————. 1974b. "On the Very Idea of a Conceptual Scheme." In Davidson, 1984.

————. 1974c. "Psychology as Philosophy." In Davidson, *Essays on Actions and Events*. Oxford: Oxford University Press, 1980.

————. 1975. "Thought and Talk." In Davidson, 1984.

————. 1984. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

————. 1985. "Deception and Division." In *Actions and Events*. Edited by E. Lepore and B. McLaughlin. Oxford: Blackwell, 1985.

————. 1994. "Donald Davidson." In *A Companion to the Philosophy of Mind*. Edited by S. Guttenplan. Cambridge: Blackwell, 1994.

Dennett, D. 1978. *Brainstorms*. Cambridge, Mass.: MIT Press.

————. 1987. *The Intentional Stance*. Cambridge, Mass.: MIT Press.

Devitt, M. 1996. *Coming to Our Senses*. New York: Cambridge University Press.

———— and K. Sterelny, eds. 1987. *Language and Reality*. Oxford: Blackwell.

Donnellan, K. 1972. "Proper Names and Identifying Descriptions." In *Semantics of Natural Languages*. Edited by Davidson, D. and G. Harman. Boston: D. Reidel, 1972.

Dreyfus, H. 1991. *Being-in-the-World.* Cambridge, Mass.: MIT Press.

Dummett, M. 1975. “Frege’s Distinction between Sense and Reference.” In Dummett, *Truth and Other Enigmas*. London: Duckworth, 1978, 116–44.

Evans, G. 1982. *The Varieties of Reference*. New York: Oxford University Press.

Evnine, Simon. 1991. *Donald Davidson*. Stanford, Calif.: Stanford University Press.

Field, H. 1977. “Logic, Meaning and Conceptual Role.” *The Journal of Philosophy* 74.

Fodor, J. and E. LePore, eds. 1992. *Holism, a Shoppers Guide*. Cambridge, Mass.: Blackwell.

Gärdenfors an Makinson. 1988. “Revision of Knowledge Systems using Epistemic Entrenchment.” In *TARK ’88: Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*. Edited by Mosche Vardi. Los Altos, Calif.: Kaufmann, 1988, 83-95.

Goldman, A. 1986. *Epistemology and Cognition*. Cambridge, Mass.: Harvard University Press.

————. 1989. “Interpretation Psychologized.” In Goldman, *Liaisons: Philosophy Meets the Cognitive and Social Sciences*. Cambridge, Mass.: MIT Press, 1992.

Grandy, Richard. 1973. “Reference Meaning and Belief.” *The Journal of Philosophy* 71: 439–452.

Hacking, I. 1975. *Why Does Language Matter to Philosophy?* New York: Cambridge University Press.

Harman, G. 1973. *Thought*. Princeton, N.J.: Princeton University Press.

Jackman, H. 1996. *Semantic Norms and Temporal Externalism*. Ph.D. thesis, University of Pittsburgh.

————. 1996b. “Radical Interpretation and the Permutation Principle.” *Erkenntnis* 44: 317–326.

————. 1998. “Individualism and Interpretation.” *Southwest Philosophy Review* 14.

————. 1998a. “James’ Pragmatic Account of Intentionality and Truth.” *Transactions of the C. S. Peirce Society* 34.

————. 1999. “Moderate Holism and the Instability Thesis.” *American Philosophical Quarterly* 36.

————. 1999a. “Holism, Meaning, and Context.” *Proceedings of the Ohio Philosophical Association*, 1999.

————. 2000. “Belief, Rationality and Psychophysical Laws.” *Proceedings of the Twentieth World Congress of Philosophy,* volume 9, *Philosophy of Mind*.

Kripke, S. 1972, 1980. *Naming and Necessity*. Cambridge, Mass.: Harvard University Press.

————. 1982. *Wittgenstein on Rules and Private Language*. Cambridge, Mass.: Harvard University Press.

Lewis, D. 1974. “Radical Interpretation.” In Lewis, *Philosophical Papers,* volume 1. New York: Oxford University Press, 1983.

Loar, B. 1981. *Mind and Meaning.* Cambridge, Mass.: Cambridge University Press.

Lukes, S. 1982. “Relativism in Its Place.” In *Rationality and Relativism.* Edited by Hollis and Lukes. Cambridge, Mass.: MIT Press, 1982

McDowell, J. 1986. “Singular Thought and the Extent of Inner Space.” In *Subject, Thought, and Context*. Edited by Pettit and McDowell. New York: Oxford University Press, 1986.

————. 1994. *Mind and World*. Cambridge: Harvard University Press.

McGinn, C. 1977. "Charity, Interpretation, and Belief." *The Journal of Philosophy* 74.

Moran, R. 1988. "Making Up Your Mind: Self-Interpretation and Self-Constitution." *Ratio* 1: 135–151.

————. 1994. "Interpretation Theory and the Fist Person." *The Philosophical Quarterly* 44: 154–173.

Putnam, H. 1975. "The Meaning of 'Meaning.'" In Putnam, H, *Mind Language and Reality*. New York: Cambridge University Press, 1975.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, Mass.: MIT Press.

————. 1970, 1986. *Philosophy of Logic*. Cambridge, Mass.: Harvard University Press.

————. 1970b. "Philosophical Progress in Language Theory." *Metaphilosophy* 1 (1970): 2–19.

————. 1990a. *The Pursuit of Truth*. Cambridge, Mass.: Harvard University Press.

————. 1990b. "Comment on Harman." In *Perspectives on Quine*. Edited by Barrett and Gibson. Cambridge: Blackwell.

Ramberg, B. 1989. *Donald Davidson's Philosophy of Language*. New York: Blackwell.

Rorty, R. 1979. *Philosophy and the Mirror of Nature*. Princeton, N.J.: Princeton University Press.

————. 1991. *Objectivity, Relativism and Truth*. New York: Cambridge University Press.

————. 1998. *Truth and Progress*. New York: Cambridge University Press.

Rott, H. 2000. "Two Dogmas of Belief Revision." *The Journal of Philosophy* 97.

Sawyer, S. 1998. "Privileged Access to the World." *Australasian Journal of Philosophy* 76.

Searle, J. 1958. "Proper Names." *Mind* 67: 166–173.

————. 1983. *Intentionality*. New York: Cambridge University Press.

————. 1987. "Indeterminacy, Empiricism, and the First Person." *The Journal of Philosophy* 84: 123–146.

————. 1991. "Response: Reference and Intentionality." In *John Searle and His Critics*. Edited by E. LePore and R. Van Gulick. Oxford: Blackwell.

Sellars, W. 1974. "Meaning as Functional Classification." *Synthese* 27.

Sen, A. 2000. "Consequential Evaluation and Practical Reason." *The Journal of Philosophy* 97.

Stich, S. 1981. "Dennett on Intentional Systems." In *Mind and Cognition*. Edited by W. Lycan. Cambridge, Mass.: Blackwell, 1990.

————. 1984. "Relativism, Rationality and the Limits of Intentional Description." *Pacific Philosophical Quarterly* 65: 211–235.

————. 1990. *The Fragmentation of Reason*. Cambridge, Mass.: MIT Press.

Strawson, P. F. 1959. *Individuals*. London: Methuen.

Taylor, C. 1980. "Theories of Meaning." In Taylor, *Human Agency and Language*. New York: Cambridge University Press, 1985.

Wilson, N. L. 1959. "Substance without Substrata." *Review of Metaphysics* 12: 521–539.