The Southern Journal of Philosophy (2003) Vol. XLI

Can They Say What They Want? A Transcendental Argument against Utilitarianism

Olaf L. Mueller
Georg-August-University

In the first chapter of *Reason, truth and history*, Hilary Putnam famously argues that the language of an envatted brain must be radically different from our language (Putnam 1981, 1–21). It is less well known that in chapter 6 of the same book, Putnam has a parallel argument with respect to consistently practicing utilitarians: According to Putnam, their language is also radically different from our own (139–41). Now in chapter 1, Putnam's point leads to a direct refutation of a certain extreme skeptical scenario, namely to a refutation of our eternal envatment (7). In chapter 6, by contrast, Putnam's point does not lead to the refutation of utilitarianism. Rather, he uses his point in order to prove that facts and values cannot be disentangled (141). I agree with this metaethical conclusion. But I want to show that the parallel between envatted brains and consistently practising utilitarians goes further than is indicated in *Reason, truth and history.* My aim is to show that complete obedience to utilitarianism is as impossible as our eternal envatment. This paper has four parts: In part 1, we shall circumscribe the sort of utilitarianism that will be the target of my argument. In part 2, we will be concerned with three preliminary thought experiments so as to become familiar with the dialectical techniques that will be needed for refuting utilitarianism. The actual refutation of utilitarianism will be given in part 3. In part 4, we shall compare my anti-utilitarian argument to Putnam's argument against brains in a vat.

Olaf L. Mueller teaches logic and philosophy at Georg-August-University, Goettingen, Germany. He writes on topics in epistemology, metaphysics, ethics, and the philosophy of language.

1. Utilitarianism and the Ideal Level of Ethical Thought

Act utilitarianism tells us that an action is morally right in a given situation only if it produces consequences that are better than (or at least as good as) the consequences of every alternative action open to us in that situation. There is considerable disagreement among act utilitarians as to how the notion of better consequences has to be spelled out. Fortunately, we need not pay attention to these differences because what I shall try to show in this paper will apply to every version of act utilitarianism: If the refutation that I propose works against, say, hedonistic versions of act utilitarianism, then it will work against its more formal counterparts such as preference utilitarianism as well.[1]

A notorious problem for act utilitarians is the following. The moral theory that they want us to follow does not seem to be suitable for guiding our moral behavior in real life situations. This is so because regretfully we humans are limited in two crucial respects: On the one hand, we don't know enough, and on the other we are not good enough for really succeeding in doing what we should do, according to act utilitarianism. We are both epistemically and motivationally restricted. True, these all-too-human limitations also make it difficult to act in accordance with non-utilitarian ethics; moral conduct is not easy anyhow. But in the case of utilitarianism the difficulty seems insuperable. To see why, just imagine yourself facing a moral choice. You would need a lot of factual and counterfactual knowledge if you really wanted to determine which action among those open to you maximized utility.[2] But you are not omniscient; therefore, you have little hope of finding out what your utilitarian obligation consists in.[3]—Worse, even if you had no such epistemic limitations, your situation would be no better. As a utilitarian agent you would have to neglect yourself and those close to you whenever you could produce more utility by way of helping people who need it more. Unfortunately there are almost always numerous anonymous people who need you more than your friends and relatives. A being with superhuman moral powers might well be motivated to behave as altruistically as utilitarianism demands. But because our motivations are more modest than those of such a happy creature, utilitarianism does not seem to be made for us.

Admittedly, all this shows merely that act utilitarianism cannot *directly* guide our moral deliberations in real life situations. Nonetheless, act utilitarianism could be right because it might prove to be an *indirect* guide to our moral obligations.[4]

The most promising way to flesh out this idea is taken by utilitarians who propose to distinguish between two levels of ethical thought.[5] On the *ideal level* we abstract from our human limitations. Ethical norms formulated on the ideal level are not addressed to true human beings but to an ideal agent, that is, to an omniscient being with highly altruistic motivations. On the *everyday life level* of ethical thought, however, we try to formulate ethical norms that are addressed to us, with all our epistemic and motivational limitations. Unlike ideal norms these norms have to be suitable for guiding our behavior in real life situations. But this is not the only property they must have.[6] In addition, norms that are acceptable on the everyday life level of ethical thought must be *justified in light of ideal norms.*

It is obvious why utilitarians feel attracted to such a two-level picture of ethical thought. Confronted with the problem of human limitation, they will claim that utilitarianism, when properly understood, was always meant to be an ideal norm, that is to say, a norm that is addressed to highly idealized agents, who are omniscient and altruistically motivated. Thus understood, utilitarianism need no longer seem implausible in light of both our epistemic and motivational limitations. Nonetheless, utilitarians who wish to appeal to this line of thought still have to work out how exactly norms are to be justified on the everyday life level of ethical thought; this may well turn out to be quite a complicated story. Happily we need not go into the details of this story because I have an argument that, if correct, renders the whole quarrel superfluous.

My thesis is as follows. The ethical theory of an ideal agent cannot possibly be act utilitarianism: The notion of an ideal agent who practically and theoretically subscribes to act utilitarianism is incoherent. That is to say, an ideal agent cannot both behave in accordance with and believe in act utilitarianism.[7] If this is right, then utilitarian philosophers can no longer appeal to an ideal level of ethical thought on which utilitarianism is supposed to be more convincing.

2. Three Preliminary Thought Experiments

Before we can proceed to the proof of my thesis with respect to utilitarianism, I want to demonstrate analogous theses with respect to three "ethical" rules that are simpler than my main target. The structure of my reasoning will always be as follows. I'll ask you to imagine speakers whose moral and verbal behavior is completely in accordance with the ethical rule in question. In each of the three cases, the ethical rule will seem rather strange from our moral point of view; so we'll have to imagine speakers whose behavior differs radically from our own. (For simplicity we shall assume that the speakers are exactly like ourselves in all aspects of their lives that are *not* affected by the ethical rule in question.) The crucial difference between us and the speakers to be imagined will concern the use of language. As we'll see, it is impossible to use language in the way these speakers are supposed to use it. We shall conclude from this that the three ethical rules under discussion are incoherent. In the next part of this paper we shall seek to detect the same sort of incoherence in utilitarianism.

First example. Imagine a speech community whose members subscribe to a moral norm that obliges them to always say the opposite of what they believe. When they believe that p, they assert *not-p*. Is such a speech community possible? Not at all. Its impossibility is of course not due to political, sociological, economical, or psychological reasons; it is due to philosophical reasons. The philosophical reasons I have in mind derive from Quine's celebrated principle of charity: Whenever you want to make sense of a speech community's verbal behavior, you'd better try to maximize agreement between your own beliefs and the assertions that you ascribe to the members of the interpreted community.[8]

Equipped with this principle, we can see what kind of mistake I made when describing the community that I asked you to imagine. My description did not maximize, it *minimized* agreement between us and the members of that community, and it thus depended upon an understanding, or interpretation, of the community's language which cannot be right. A better, more charitable interpretation of this very language is readily available: Simply remove the negation sign (with the widest scope) from every "assertion" that was ascribed to the natives under the original interpretation.[9] (Remember that whatever the speakers were "asserting" according to the original interpretation had the form *not-p*).

Given this new interpretation, the members of the imagined community can no longer be understood as saying the opposite of what they believe. On the contrary, they now follow the Eighth Commandment as much as we do. Conclusion: When properly understood, a community of eternal liars is not a community of eternal liars, or, less paradoxically: There cannot be a community of eternal liars.

Second example. If it is impossible to turn the Eighth Commandment on its head, then the same holds good for the obligation to keep one's promises. Thus it is impossible to imagine a speech community whose members subscribe to a moral norm that obliges them always to do the opposite of what they promise. The community of alleged promise-breakers must be re-interpreted; whenever a member from this community seems to say "I promise to do X", she has to be understood as saying: "I promise *not* to do X."

Striking as the analogy between our reasoning concerning eternal liars and permanent promise-breakers may seem, we should not fail to notice that the principle of charity takes a different form in the two cases. In its original version, the principle aimed at maximizing *agreement*. This was fine within the context of our first example because agreement is an aspect of assertoric language use and because in this example we were indeed dealing with assertoric speech acts. But in the second example, we are dealing with non-assertoric speech acts, that is, with promises, and an adequate interpretation of promises cannot be said to maximize *agreement* between interpreter and interpretee. Still, an interpretation on which the speakers are *always* unfaithful to their promises clearly does not accord with the spirit of Quine's principle of charity. How are we to generalize the principle so as to cover such speech acts?

Speakers who always break their promises have something in common with speakers who always say the opposite of what they believe: They are extremely unreliable in using language, that is, they do not comply at all with the linguistic rules that are conventionally connected with their utterances. But it is certainly not charitable if, without need, someone calls your behavior altogether unreliable. Therefore, the principle of charity should take the following form:

> Any plausible interpretation of utterances of a certain speech act type should care for the speaker's overall compliance with the linguistic rules defining that same speech act type.

There are good arguments for this version of Quine's principle. They rest on an insight that lies behind the truism that *meaning is use*.[10] Admittedly, meaning should not be equated with use; the insight behind that slogan nonetheless remains valid: Meaning and use of linguistic expressions cannot be divorced from one another. If we apply this idea to the linguistic devices that indicate speech act types (like assertion, promise, etc.), it follows that an adequate interpretation of these devices cannot be completely independent of their use. Thus, suppose your interpretation says that a certain grammatical construction indicates assertions; this interpretation is refuted, if it turns out that the speakers *never* comply with the rules governing the exchange of assertions, that is, if it turns out that they do not care for truthfulness at all. Of course, from time to time the speakers may well break these rules. So the principle of charity should allow for exceptions, and it should be amended with a *ceteris paribus* clause:

> Other things being equal, interpretation A is more plausible than interpretation B if, under interpretation A, the speakers can be seen to follow the rules governing their speech acts more reliably than under its alternative B.

Let us see how Quine's principle in this new form works when applied to another case.

Third example. Imagine a community whose members always toss a coin in order to determine whether or not they will do what they have promised. Again, our principle tells us that there cannot be such a community: Whenever its members seem to say "I promise to do X", they should not be interpreted at face value; rather, they should be interpreted as follows: "I shall toss a coin to determine whether or not I'll do X."

And this is not an expression of a promise; nor is it an expression of an assertion, as it were, about the future—it is a speech act of an altogether different type that might be called an *accidental promise*. Notice that this label is a little misleading; we should not think of accidental promises as being a species of promise. On the contrary, an accidental promise is no promise at all.

One might object: Why shouldn't it be a promise? Couldn't an accidental promise to do X be analyzed as a genuine promise to toss a coin to determine whether or not to do X?The answer is to the negative. Genuine promises belong to a practise which is more complex than the accidental practice of the community we are trying to imagine. One important characteristic of that practice is this: Whenever you have promised something, you could just as well have promised the opposite. I want to demonstrate that if this is right, then the objection cannot hold good.

Suppose that a member of the imagined community has given an accidental promise to do X. According to the objection under discussion, her utterance can be analyzed as a genuine promise to toss a coin to determine whether or not to do X. If, however, it were to be a genuine promise, then the speaker should have been able to genuinely promise *not* to toss a coin to determine whether or not to do X. And this she cannot possibly do, because in the charitable understanding of whatever she might say, the negation sign cannot take the position that is needed. For example, if she says:

(*) I promise not to do X,

then she must be interpreted as accidentally promising not to do X, which would be tantamount to the "genuine" promise to toss a coin in order to determine whether or not to do *not*-X. But what we need is the genuine promise *not* to toss a coin in order to determine whether or not to do X.
And if, on the other hand, she says:

(**) I don't promise to do X,

then she must be interpreted as not accidentally promising to do X, which would be tantamount to *not* giving the "genuine" promise to toss a coin in order to determine whether or not to do X. But again, this is not what we need because not giving the promise to do Y is different from giving a promise not to do Y.

Let us conclude from these considerations that, in the third community that I asked you to imagine, it is impossible to give genuine promises. *A fortiori*, then, it is also impossible that its members handle their genuine promises accidentally. To repeat, in that community there are no genuine promises to be handled in this or the other way. Thus the idea of a community subscribing to the moral rule "Always toss on your promises" is incoherent.

3. Utilitarianism Refuted

Our last example was perhaps a bit too playful. Thus, the next example should and will have more practical significance. Let us imagine a speaker who deals with her promises not accidentally but in the utilitarian fashion: Whenever she promises to do X, she does X only if this maximizes (expected) utility.[11] Let us suppose for the sake of argument that she always succeeds in complying with this utilitarian commandment. We can assume her to be an ideal agent, that is, omniscient and altruistically motivated.

Now, quite often the greatest utility cannot be obtained by keeping one's promise. As we all know from textbooks on ethics, there is a systematic range of cases where our speaker must break her promise—if she wants to act in accordance with utilitarianism. And this means that she will have to *systematically* break her promises; she is not complying with the rules that define the speech act type *promise*. She does not even intend to comply with these rules; her intention is directed toward maximizing utility.

How are we to interpret the alleged promises of our utilitarian speaker? I don't think we should interpret them at face value because doing so would turn her into an unreliable language user. But the utilitarian speaker is not unreliable in general. True, she is an unreliable promise-keeper, but she remains a reliable utilitarian nonetheless. This gives us the clue for her adequate interpretation. We must look for some kind of utilitarian speech act, that is, for a speech act type that is governed by utilitarianism because utilitarianism is the norm which our speaker in fact follows when she seems to make a promise.

In our language there is no particular convention for expressing speech acts of this utilitarian type. This lack of expressive force in our language need not surprise us; after all *we* do not practice utilitarianism in everyday life. Thus we have to invent the kind of speech act in question. Let us not be confused by the fact that our linguistic creativity is called for at the present stage of our investigation; we wish to describe linguistic behavior that deviates radically from our own. If there is any truth to the slogan that meaning is use, then it is to be expected that in our language we do not have resources for indicating the strange speech act we are after. The best we can do is to make use of our own linguistic resources for *circumscribing* the speech acts that are determined by utilitarian usages of language.

Suppose our utilitarian speaker says "I promise to do X." How should we translate this so as to ascribe to her a sufficiently high degree of reliability? Our first attempt departs from the observation that the speaker will, reliably, perform X only if performing X maximizes utility. This observation suggests the following interpretation:

(*) I'll perform X only if performing X maximizes utility.

As before, this shouldn't be taken to be a genuine promise; nor is it an assertion about the future. Rather, when interpreted in the manner of (*), the speaker is seen as *emphasizing* her utilitarianism with respect to X. Indeed this interpretation of the speaker's apparent promise has it that the speaker displays a high degree of reliabilty because her conduct *is* emphatically utilitarian.

Nevertheless the proposal is not convincing. It does not highlight the speaker's reliability *with respect* to her utterance because she will in any case reliably perform X only if doing so maximizes utility, whether or not she has uttered the words "I promise to do X" beforehand. Let us, therefore, see whether we can find a rule that reliably governs the speaker's use of the words "I promise to do X." We shouldn't be misled by the fact that *we*, when we utter those words, already have an intention to perform X; in our case this very intention is intrinsically connected with the utterance of those words. Not so in the case of our utilitarian. To her, performing X is one possible item on her utilitarian agenda, while uttering the words "I promise to do X" is another possible item on that agenda; the two items are kept separate in the speaker's deliberations because each stands in need of its own utilitarian justification.[12] We are concerned here with interpreting the first item only. And the rule our speaker is following with respect to the utterance in question is this:

> It is correct to say "I promise to do X" only if saying so maximizes utility.

It is not easy to see what kind of speech act is defined by this rule. Of course, it is not a genuine promise; nor is it an assertion. Still it seems to be a speech act—if it is right to view speech acts as possible ways of realizing Austin's famous phrase *How To Do Things With Words*.

We do have speech acts in our language that are somewhat similar to what our utilitarian is doing with her words. For instance, we have certain phrases that we use for consoling the bereaved. Consolation is a speech act that aims at improving the mood of the listener; by contrast, our utilitarian performs speech acts that aim at improving the situation of *everybody*. And unlike consolation, the utilitarian speech act is not restricted to the occasion of funerals; it could be performed on any occasion, at any time (so long as it maximizes utility).

It might be interesting to think more about utilitarian speech acts, but we need the remainder of our time for drawing some conclusions. As our utilitarian speaker cannot express genuine promises, it follows that she cannot possibly apply her moral theory to promises she herself has given.[13] The utilitarian rule "Break your promises if doing so maximizes utility" is incoherent.

This will not yet impress the utilitarian much. She'll reply that she has been opposed to the institution of giving promises all along. And indeed, quite often during its long history, utilitarianism has taken a revolutionary line of opposing bad old institutions that have to be overcome in the name of the general good.

Why shouldn't we free ourselves from the institution of promises? Let's revolutionize our language so as to get rid of even the linguistic resources for expressing them! Perhaps we could do that, but it would not be the end of the story. The utilitarian revolution of our language must go much further than that. It must go beyond what we can afford, even beyond what utilitarianism itself can afford.

To prove this I'll try to convince you that utilitarianism is not only incompatible with promises but with assertions as well. If I am right, utilitarian speakers cannot possibly make assertions! As we shall see, this puts an end to utilitarianism.

Let me first provide you with some intuitive evidence in favor of my claim. There have been several points on our way where it became evident that assertions and promises can be dealt with similarly. First of all, we appreciated that neither the rules governing assertions nor those governing promises can be inverted: Both the community of eternal liars and that of permanent promise-breakers are philosophical impossibilities. Second, the reason for this was the same in both cases: There is a generalized version of Quine's principle of charity covering not only assertions but also promises. Third, our reasoning against tossing on promise-keeping allows for a parallel with respect to assertions: It is impossible to imagine a community whose members always toss a coin in order to determine whether they assert what they believe or its negation.[14]

Now, we have just seen that consistent utilitarianism leaves no linguistic room for the speech act of promising. Thus, if there is a parallel between promises and assertions, we may expect that utilitarianism leaves no linguistic room for the speech act of asserting either. So much about the intuitive evidence for my claim. Let us proceed to its proof.

As it turns out, we need not do much to prove my claim. Let us suppose that a practising utilitarian says something that sounds like an assertion from our language, for example, "There is lots of mineral water in Karlovy Vary." How is this to be translated? Before we try to get clear about the propositional content of this utterance, we'd better find out what kind of speech act it exemplifies. As we have seen, each speech act type is defined by particular linguistic rules, and we have seen that only if the speaker really follows those particular rules with great reliability, can her language be said to be equipped with the speech act type in question. So let us ask: What are the rules that reliably govern the speaker's usage of the words "There is lots of mineral water in Karlovy Vary"? Because our speaker is a utilitarian, there is but one rule that she is following in all her verbal and non-verbal behavior: utilitarianism. Thus, with respect to the utterance in question we obtain:

> It is correct to say "There is lots of mineral water in Karlovy Vary" only if saying so maximizes utility.

Not surprisingly this rule displays exactly the same pattern as the rule we found in the speaker's behavior with respect to alleged promises:

> It is correct to say "I promise to do X" only if saying so maximizes utility.

So we see that within utilitarian language, the rules governing utterances that sound like assertions are in no way different from those governing utterances that sound like promises. What's more, due to the speaker's utilitarianism, *every* utterance U from her idiolect reliably follows a rule of the very same form:

> It is correct to utter U only if doing so maximizes utility.

But if it is true that speech act types are defined by the rules governing them, then we can conclude that there exists but one uniform speech act type in the language of our utilitarian speaker. It is a kind of speech act that we cannot express but only circumscribe in our language. As we have seen, the speech act in question aims at the people's happiness—on the one hand resembling the speech act of consolation while on the other differing from it by being more general.

For our purposes we don't need to know much about this bizarre kind of speech act; for us it suffices to see that the speech acts performed by practicing utilitarians cannot be understood as assertions. And indeed it seems clear that they cannot be so understood: If they resemble consolations then they are far away from what *we* do with words when we express our beliefs, describe something, state an opinion etc.

But are the utilitarian's speech acts really so different? There are at least five objections which might be raised against my claim.

First objection. In everyday life the utilitarian will follow the Eighth Commandment as reliably as we do. She won't calculate the consequences of every possible utterance because this would take too much time and money, that is, happiness. Rather, she'll stick to the Eighth Commandment as a rule of thumb. Her usage of the utterance "There is lots of mineral water in Karlovy Vary" will be similar to ours and, thus, will qualify as an assertion proper.—Reply. On the level of everyday life ethical thought this is perhaps right. But we are concerned with refuting utilitarianism on the *ideal* level of ethical thought. An ideal agent is omniscient and does not need any rules of thumb. If she acts in accordance with utilitarianism, then her usage of utterances such as "There is lots of mineral water in Karlovy Vary" is likely to deviate radically from our usage of such utterances.

Second objection. Even on the ideal level of ethical thought the difference in usage is not as grave as needed for my argument. We shouldn't situate the ideal utilitarian agent, whom we want to interpret, within a speech community whose members are all ideal utilitarian agents, too. If we did so, the thought experiment of ideal agency would lose its point: We don't want to know what an ideal agent would do in an ideal world (where everyone is omniscient and altruistically motivated); we are interested in moral questions from our world, whose inhabitants are far from perfect. But if, so the objection proceeds, the ideal utilitarian, whose language we are interpreting, stands alone in a speech community whose members are not ideal, then she will have to speak this very community's language. Otherwise she could not possibly interact with its members and, thus, could not maximize utility among them.

Reply. It is true that we ought to imagine the ideal agent within an unideal speech community;[15] it is not true, however, that the mere interaction with members of a certain speech community makes you speak their language. Whether or not you really speak a certain language does not depend on how effectively you are able to reach your goals *via* exchanging sound waves with speakers of that language. You could be quite successful in verbally manipulating those speakers without following the linguistic rules that they follow. But if you do not follow their rules, then you are not speaking their language.

Third objection. But the ideal agent, who is utilitarian, would have to follow the linguistic rules of her speech community; otherwise she would be excluded from that community and couldn't maximize its utility any longer.—Reply. This proves at best that our ideal agent must speak and behave *as if* she were complying with those rules. But to follow a rule is not the same as *apparently* following a rule. To see this, notice how much the ideal agent's intentions with respect to "assertions" differ from those of the members of that speech community. The former are always directed at maximizing utility, while the latter, normally and as a rule, aim at truth.

Fourth objection. Granted, the ideal agent must aim at maximizing utility—if she wants to be utilitarian. But why does this preclude her from simultaneously aiming at the truth when she utters something that sounds like an assertion?—Reply. Sometimes (perhaps even rather often) maximizing utility might demand from her that she utter something sounding like an assertion and which is true in the language of the speech community concerned. In this case one might say that she aims at truth *for the sake of* maximizing happiness. But there will be numerous factual and counterfactual cases where truth and general happiness fall apart. It is these cases that call for an interpretation more charitable than the standard one.

Fifth objection. Couldn't it be that utilitarianism *implies* truthfulness, so that, contrary to the preceding reply, general happiness and truth *cannot* fall apart?—Reply. If we could count on some sort of preestablished harmony between the goal of truth and that of general happiness, then my argument against utilitarianism wouldn't work. But as far as I can see, there is not the slightest reason to believe in the preestablished harmony that my opponent is invoking. Anyway, many utilitarians have assumed that they were fighting against strict obedience to rules such as the Eighth Commandment. It would come as an unwelcome surprise to them, should it turn out that their own moral theory implied, rather than weakened, commandments from the Bible.

In sum, it seems reasonable to think that an ideal agent, who behaves in accordance with utilitarianism, cannot possibly express assertions. What is more, she cannot possibly make any of the speech acts *we* are familiar with.[16] As we have seen, the only kind of speech act open to her is something that is slightly similar to, but more general than, consoling the bereaved.

If it is true that there is but one, unheard-of speech act type that the ideal utilitarian agent can perform, then we may start wondering whether she is in command of a language at all; her verbal behavior seems to be an activity of an altogether new sort.

Considerations like that may already cast considerable doubt on the plausibility of utilitarianism. But they depend too much on debatable intuitions about the nature of language. Let us try to proceed on dialectic ground that is firmer.

In addition to what I have shown so far, I need one more premise to *refute* utilitarianism. The additional premise is this: Anyone fully subscribing to utilitarianism must not only *act* in a certain way (i.e., act so as to maximize utility), she must also *entertain certain beliefs*. This seems plausible. If you want to qualify for being a utilitarian, then it is certainly not enough that you in fact succeed in maximizing utility; this might be a matter of good luck—or even bad luck, in case it runs counter to your intentions. For being utilitarian you need to have specifically utilitarian intentions: you must do what you do *because* you believe that your very action will maximize utility.[17]

At least, an *ideal* agent has to entertain beliefs of this kind—if she is to qualify as a utilitarian. Although our discussion takes place on the ideal level of ethical thought, the additional premise can even be defended on the everyday life level of ethical thought. To defend it on this level, it suffices to name just one belief that every utilitarian must hold. Here is one such belief: "Actions have consequences." How could you be utilitarian without believing in this truism? I conclude that my additional premise is plausible on both levels of ethical thought.

So let us use it for completing our argument against utilitarianism. On the one hand, we have shown that an ideal agent who is utilitarian cannot possibly state assertions; that is to say, she cannot possibly express her beliefs. On the other hand, we have seen that every utilitarian must hold certain beliefs, for example, the belief that actions have consequences. We can already sense the tension between these two points; I want to convince you that they are incoherent.

This should be an easy task. We all believe in some version of Wittgenstein's celebrated private language argument. Although I must admit that it is difficult to formulate it precisely, I think we can rely on its conclusion nonetheless. Applied to our problem, Wittgenstein's conclusion tells us that someone who *cannot* publicly express her beliefs cannot entertain them privately either. This is precisely the situation of the utilitarian we have been interpreting: Whatever she might say, she cannot possibly state her beliefs because her language is not equipped with the speech act of assertion. However the details of Wittgenstein's argument have to be combined, it seems clear to me that it would be a magical mystery if our speaker could secretly entertain beliefs that are forever excluded from public access.[18] (Even an omniscient radical interpreter could not guess at the contents of those alleged beliefs!)

If this is right, utilitarianism must break down. As we have seen at the opening of our discussion, utilitarianism cannot be practiced in everyday life.[19] But now we know that things look even worse for utilitarianism on the ideal level of ethical thought. On the ideal level, utilitarianism is incoherent: If an ideal agent really always acts in accordance with utilitarianism, even while she speaks, then she cannot entertain the beliefs necessary for being motivated by that very same utilitarianism. In short, if an ideal agent practically subscribes to utilitarianism, then she cannot subscribe to it theoretically as well. The practical side and the theoretical side of utilitarianism do not agree.

4. Brains in a Vat and Utilitarianism

As is well known, Putnam is not fond of utilitarianism; I take it that his rejection of utilitarianism is grounded in firm moral intuitions against the structure and against implications of utilitarian ethics. He does not need an argument against utilitarianism that comes from theoretical philosophy. I must confess that my own moral intuitions tend strongly toward utilitarianism. In my experience, utilitarians often find themselves in a reflective equilibrium against which one cannot argue from within moral philosophy. That's why philosophers with moral intuitions similar to mine might be more surprised about my argument than Putnam.

But even though my argument is not related to Putnam's moral philosophy, it has a parallel in his theoretical philosophy. As it happens, my argument makes use of dialectic techniques that were first introduced by Putnam in his famous argument against the brain-in-a-vat scenario. Let me conclude our discussion with a comparison of the two arguments.

Both arguments are designed to refute certain extreme scenarios invented by philosophers: Putnam's argument refutes an extreme, skeptical scenario (viz., eternal envatment), while my argument refutes an extreme moral scenario (viz., complete obedience to utilitarianism).

The crucial step in both arguments lies in re-interpreting languages: neither the language of an envatted brain nor that of an ideal utilitarian agent can be taken at face value. In fact, neither of the two arguments depends on a positive claim about the interpretation of the speaker concerned: for Putnam's argument it suffices to deny that an envatted brain can refer to brains;[20] for my argument it suffices to deny that an ideal utilitarian agent can make assertions. In short, both arguments deny the speaker's linguistic ability to come to terms with her extreme situation.

Putnam's argument needs a particular premise from the philosophy of language; so does mine. But whereas my argument invokes a generalized version of Quine's principle of charity, Putnam's argument appeals to the causal theory of reference (or anyway, to the denial of magical theories of reference[21]). This apparent difference doesn't threaten the parallel between the two arguments. If you ascribe magical referential powers to a speaker, then this may be a nice compliment; it is nonetheless not in accordance with Quine's principle of charity. Remember that the magical interpretation of an envatted brain makes most of its beliefs false and, thus, minimizes agreement. (That's why Davidson's version of the argument against envatted brains works without appealing to causal theories of reference; rather it invokes the principle of charity in the very same way as does my argument against utilitarianism.[22])

Both arguments, Putnam's and mine, yield a conclusion that is not analytic but can be known by *a priori* reasoning. How is this possible? The trick is, in both cases, performed by some sort of transcendental technique. Both arguments depend on the insight that the conditions for the possibility of describing one's own situation are not fulfilled, neither in the case of an envatted brain nor in the case of an ideal utilitarian agent.[23]

Having mentioned philosophical conceptions such as the *synthetic a priori* and transcendental reasoning, I can proceed to the final point of comparison between the two arguments. It is not easy to swallow the result that we *cannot* be brains in a vat; Putnam's argument is surrounded by the air of philosophical dubiousness. Mine is too, isn't it?[24]

Notes

[1] Notice, however, that in what follows I shall not attempt to refute rule utilitarianism.

[2] I shall use this term as a dummy for covering whatever it is that utilitarianism wants us to maximize. The term is neutral with respect to the competing versions of act utilitarianism mentioned in the preceding paragraph. Furthermore, for brevity, I shall often speak of "utilitarianism" instead of using the more exact label "*ideal act* utilitarianism."

[3] Admittedly, omniscience is merely a sufficient and not a necessary condition for success in determining which action maximizes utility. (You need not be omniscient for example with respect to the past in order to find out what your utilitarian obligation consists in.) But although less than omniscience is needed, it seems clear that a successfully practicing utilitarian must know much more than human beings can be expected to know. For the sake of brevity, I shall not repeat this clarification in the main text.

[4] One might wonder whether an ethical theory needs to have a guiding function at all, be it direct or indirect. Couldn't utilitarianism simply tell us which property in fact distinguishes right from wrong, without any indication as to how we human beings should ever be able to do what is right and avoid what is wrong? I don't think this is a good idea. If an ethical theory had no guiding function whatsoever, then it would be difficult to claim that it really *is* an ethical theory, that is, a theory about the right and the wrong, rather than a theory about, say, the *ight* and the *wong*. Thus, imagine a tribe of utilitarians who call an action "ight" whenever they believe that it maximizes utility. If the natives are not inclined or motivated to perform actions which they call "ight", how, then, should an omniscient radical interpreter find out that "ight" means "morally right"? Isn't it more plausible to translate the natives' word "ight" by the phrase "maximizes utility"?

[5] This line of thought has been made prominent by Richard Hare (1981, 25ff). I shall concentrate on Dieter Birnbacher's more recent version of the same strategy, cf. Birnbacher 1988, 16–23.

[6] There are many norms suitable for guiding our behavior in real life situations. Here is one such norm: "Do always what you like."

[7] Thus, my argument does not apply to those versions of act utilitarianism that do not aim at practical significance but content themselves with a theoretical claim only. According to these—externalist—versions of utilitarianism, you can describe a certain action as morally right (because you think it maximizes utility) without being motivated to perform that action when it is open to you. (See, for example, Brink 1989). I am indebted to Tatjana Tarkian and Jay Wallace for directing me to this externalist point of view. (For lack of space I cannot give substantial reasons for why I find such views implausible. But compare footnote 4.)

[8] Cf. Quine 1960, 59; compare also Davidson 1984, 196.

[9] A similar interpretation has been considered by Lewis 1986, 340–42.

[10] The truism goes back to Wittgenstein 1984, §43.

[11] The probabilities necessary for calculating expected utilities are to be understood as being conditional on the fact that the others expect her to do X.

[12] If there is a connection between the speaker's words "I promise to do X" and her performing X, then it cannot be an intrinsic connection, that is, a connection mediated by a rule. It can only be an extrinsic connection, based on contingent facts. For instance, uttering "I promise to do X" might cause certain people to expect the speaker to do X, and it might be that it is these expectations that are decisive for X's maximizing utility.

[13] We are assuming a speaker who is utilitarian throughout her life.

[14] The parallel in ethics between asserting and promising goes still further than these three points indicate. For instance, consider the best cases for an application of Kant's categorical imperative. His arguments are much more convincing when directed against lying and breaking one's promises than when directed against, say, suicide.

[15] Thus, unlike Putnam (1981, 139-41), I propose to interpret an isolated utilitarian speaker within a non-utilitarian speech community. I find it difficult to imagine a *stable* language spoken by

a group of utilitarians. (It may well be that their "language" will soon cease to exist). But can I really do without debatable speculations concerning the stability of utilitarian language as a social institution? Happily the answer is to the positive. Remember that my target is *act* utilitarianism. Unlike rule utilitarians, act utilitarians are not committed to the claim that all members of a community could simultaneously act (and speak) in accordance with utilitarianism. Therefore, we can leave it open how the language of a utilitarian speech community might change in time.

[16] This result resembles a similar claim by Hodgson (1967, 50-62). Hodgson's considerations concern promises and assertions of an act utilitarian agent in a non-utilitarian society; his claim is that the act utilitarian must produce suboptimal consequences because she cannot successfully communicate with her peers; this in turn (Hodgson thinks) amounts to a refutation of act utilitarianism. I think this is too hasty; in my view, an ideal act utilitarian agent can always produce noises that maximize utility. The problem is, rather, how those noises have to be interpreted. If it is right that they cannot be understood as assertions or promises, then this by itself does not suffice for refuting act utilitarianism. This is why I have to invoke additional considerations (in the remainder of the present part) that do not have a parallel in Hodgson's book. (For a detailed criticism of Hodgson's arguments see Singer 1972 and Lewis 1986.)

[17] We can leave it open whether you also have to believe in utilitarianism itself; plausible as this might seem on first sight, there are—so-called non-cognitivist—philosophers who deny that ethical systems can possibly be objects of belief, that is, Ayer 1946, ch. 6.

[18] Here is a possible counterexample to my claim that is due to Charles Travis (oral communication): An autistic child cannot express her beliefs even though we might still wish to ascribe certain beliefs to her. (A similar point can be made with respect to higher animals.) If this is so, then the beliefs in question must be basic and cannot have sophisticated structures. They must be more simple than the belief in complicated counterfactuals. But practicing utilitarians must believe in extremely complicated counterfactuals. Thus, the objection forces me to reformulate my argument in more specific terms. It seems clear that this can be done.

[19] By the way, if in the end you do not find the arguments from part 1 against the practicability of utilitarianism convincing, then we don't need the ideal level of ethical thought. In this case, we can repeat my argument on the everyday life level of ethical thought.

[20] This has been shown in Crispin Wright's careful reconstruction of Putnam's argument, see Wright 1994. Wright's version of the argument is still in danger of begging the question against the skeptic (see Mueller 2001a, 302–3). But this danger can be avoided, see Mueller 2001a, 313–15.

[21] Cf. Putnam 1981, 3–5, 16.

[22] Cf. Davidson 1986, 313, 316–19.

[23] As I have shown elsewhere, Putnam's argument cannot convince the skeptic unless it exhibits transcendental features. There are several ways of formulating such transcendental arguments; either you concentrate on the conditions for the possibility of *referring* to elements from the skeptic's scenario (see Mueller 2001a, 307), or alternatively, on the conditions for the possibility of

describing her scenario by means of terms with the appropriate *intension* (see Mueller 2001b, 530–37). By contrast, the transcendental argument from the present paper concentrates on the conditions for the possibility of making *assertions*.

[24] This is a modified version of a paper presented at the *Fifth Karlovy Vary Symposium on Analytic Philosophy (Swimming in XYZ, Supervised by Hilary Putnam),* September 14th-18th, 1998. I should like to thank the symposium's participants for stimulating discussions. (German versions of the paper were presented at Georg-August University Goettingen, Duesseldorf University, Free University Berlin, University of Bonn, and I am grateful to the audiences at those various places for quite a lot of interesting controversy.) Many thanks to Thomas Schmidt, and Kathi Koellermann for various suggestions that helped me to improve the paper. I am grateful to David Hyder for linguistic advice; thanks also to an anonymous referee, who prevented me from a serious exegetical mistake. I am most grateful to Hilary and Ruth Anna Putnam for generous encouragement.

References

Ayer A. J. 1946. *Language, truth and logic.* 2d ed. London: Victor Gollancz.

Birnbacher, Dieter. 1988. *Verantwortung fuer zukuenftige Generationen.* Stuttgart: Reclam.

Brink, David. 1989. *Moral realism and the foundations of ethics.* Cambridge: Cambridge University Press.

Davidson, Donald. 1984. On the very idea of a conceptual scheme. In Davidson, *Inquiries into truth and interpretation.* Oxford: Clarendon Press, 183–98.

———. 1986. A coherence theory of truth and knowledge. In *Truth and interpretation: Perspectives on the philosophy of Donald Davidson,* ed. Ernest Lepore. Oxford: Blackwell, 307–19.

Hare, Richard. 1981. *Moral thinking: Its levels, method and point.* Oxford: Clarendon.

Hodgson, D. H. 1967. *Consequences of utilitarianism. A study in normative ethics and legal theory.* Oxford: Clarendon.

Lewis, David. 1986. Utilitarianism and truthfulness. In *Philosophical Papers: Volume II,* ed. Lewis. Oxford: Oxford University Press, 340–42.

Mueller, Olaf. 2001a. Does Putnam's argument beg the question against the skeptic? Bad news for radical skepticism. In *Erkenntnis* 54: 299–320.

Mueller, Olaf. 2001b. Der antiskeptische Boden unter dem Gehirn im Tank: Eine transzendentale Fingeruebung mit Intensionen. In *Zeitschrift fuer philosophische Forschung* 55: 516–39.

Putnam, Hilary. 1981. *Reason, truth and history.* Cambridge: Cambridge University Press.

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

Singer, Peter. 1972. Is act-utilitarianism self-defeating? In *The Philosophical Review* 81: 94–104.

Wittgenstein, Ludwig. 1984. *Philosophische Untersuchungen.* In Wittgenstein, *Werkausgabe Band 1.* Frankfurt: Suhrkamp, 225–618.

Wright, Crispin. 1994. On Putnam's proof that we are not brains in a vat. In *Reading Putnam*, ed. Peter Clark and Bob Hale. Cambridge, Mass.: Blackwell, 216–41.