[A shortened version of this paper was published in *The Southern Journal of Philosophy*, 41 No. 2 (June 2003), pp. 241-259].

A Transcendental Argument against Utilitarianism

Olaf L. Müller

(Www.GehrinImTank.de)

CONTENT.

ABSTRACT. I want to explore a new way of refuting act-utilitarianism. My claim is that nobody maximizing utility can possibly be said to believe in act-utilitarianism. In *section I*, I shall circumscribe the sort of utilitarianism with which we'll be concerned: Act-utilitarianism on the ideal level of ethical thought. *Section II* is devoted to an earlier attempt of refuting act-utilitarianism, which resembles the argument from this paper. I shall try to show that the proposed refutation (due to Hodgson) is not convincing because it leaves out half of the story. In that section, I shall also motivate the theoretical (Quinean) background of my own argument. The main hero of *section III* will be Quine's principle of charity; we'll be concerned with three preliminary thought experiments so as to become familiar with the dialectical techniques that will be needed for refuting act-utilitarianism. The core of my paper can be found in *section IV*. I shall demonstrate that act-utilitarian agents can express neither promises nor assertions. After having defended the latter claim (as regards assertions) against six small objections (*section V*) and one big objection (*section VI*), I shall complete my argument in *section VII.*

NOTE. This is a modified version of a paper originally presented on September 16th at the *Fifth Karlovy Vary Symposion on Analytic Philosophy (Swimming in XYZ, Supervised by Hilary Putnam,* September 14th-18th, 1998).

A Transcendental Argument against Utilitarianism

In this paper, I want to explore a new way of refuting act-utilitarianism. I shall not try to mobilize moral intuitions against certain implications of act-utilitarianism because this strategy is not likely to impress hardcore utilitarians; they keep themselves in reflective equilibrium by replacing recalcitrant intuitions by intuitions that agree better with their utilitarian theory. To impress these act-utilitarians, I shall try to argue *from within* against utilitarianism. If I am right, it can be shown that there is an incoherence in act-utilitarianism, which, as far as I know, up to now has escaped the notice of both act-utilitarians and their critics. The incoherence in question can be brought out by way of invoking some premisses from the philosophy of language, such as Quine's principle of charity. Utilitarian philosophers could of course attack these theoretical assumptions if they wanted to challenge my argument. But even for them the argument might be instructive, as they could learn from it which philosophy of language they had better avoid. (And I shall try to indicate why avoiding it won't be easy for act-utilitarians). Let me conclude these introductory remarks with the following metaphilosophical *caveat*. I do not expect that with this paper I'll convince act-utilitarians that they are wrong; conclusive proofs are extremely rare in philosophy. Still, in describing what is going on in this paper, I shall use phrases such as "refutation", "incoherence", etc. They should be taken *cum grano salis*. My aim is not to deliver a knock-down argument but to pose a serious problem for utilitarians—so that they must say *something* in order to meet the challenge from my argument. Thus I shall not consider it a failure of the paper, should it succeed in eliciting substantial objections.

I. 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 which 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 which 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 act-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 act-utilitarianism demands. But because our motivations are more modest than those of such a happy creature, act-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 act-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 the light of ideal norms*.

It is obvious why act-utilitarians feel attracted to such a two-level picture of ethical thought. Confronted with the problem of human limitation, they will claim that act-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, act-utilitarianism need no longer seem implausible in the light of both our epistemic and motivational limitations. Nonetheless, act-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; and 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 which, if correct, renders the whole quarrel superfluous.

My claim 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 act-utilitarian philosophers can no longer appeal to an ideal level of ethical thought on which act-utilitarianism is supposed to be more convincing.

Before proceeding to the proof of my claim, I want to look at a similar proof, which tries to manage without taking into account the *theoretical* aspects of act-utilitarianism. That proof will be failure; I shall argue that it is not enough merely to seek an incoherence within act-utilitarian *practice*.

II. Hodgson's Attempt and an Answer from Quine

A good starting point for our anti-utilitarian endeavour can be found in the book *Consequences of Utilitarianism* by D. H. Hodgson. Hodgson does not say explicitly that his considerations take place on what I have called the ideal level of ethical thought. But this is implicit in his reasoning throughout the whole book.[8] He invites us to imagine what a sophisticated act-utilitarian might consider when asking herself (in a given situation) whether or not she should tell the truth or keep a promise. If both the agent and her adressee are highly rational, then (Hodgson claims) they are doomed to end up in a vicious circle—to the effect that neither assertion nor promise deserves trust. Here is what Hodgson says with regard to the case of keeping promises:

> [...] a promised act could have greater (comparative) utility (than it would have had if it had not been promised) only if the promisee has a greater expectation that it would be done (than he would have had if it had not been promised); but there would be a good reason for such greater expectation only if (in the promisor's belief) the act would have such greater utility. Being highly rational, the promisor would know that the greater expectation was a condition precedent for the greater utility; and so would not believe that the act would have greater utility unless he believed that the promisee had greater expectation. Also being highly rational, the promisee would know this, and so would not have greater expectation unless he believed that the promisor believed that he had greater expectation. And this, of course, the promisor would know.[9]

And so the circle goes on. Trying to escape from it amounts to "bootstrap-tugging", says Hodgson.[10] He offers similar considerations of truthfulness, which drive us to the conclusion that we have no rational reason whatever to believe in the truth or truthfulness of an assertion made by someone whom we know to obey to act-utilitarianism.[11]

So the act-utilitarian agent faces an unhappy choice. Either she must risk sacrificing the social benefits which she wished to produce by way of making an assertion that is likely to be believed (i.e., the benefits from successful exchanges of information, helping to coordinate social behavior). Or she must systematically deceive her environment with respect to her moral orientation, thus excluding herself from forming deep human relationships, so that as a result, again she would have to produce consequences worse than those open to non-utilitarians.[12]

Hodgson concludes that act-utilitarianism is "irrational" and even "self-defeating".[13] And that seems right—if indeed it were true that (in some situation) no option open to the act-utilitarian agent is apt to produce best consequences.[14]

But can this be true? It sounds like a paradox because it contradicts a truism from mathematics: From any finite, non-empty set of objects that have been measured on one and the same scale (such as weight, temperature, or cardinal utility), you can always choose a maximal object (i.e., an object such that in the set concerned there are no objects further up on the scale in question). Let us apply this to our case: When an act-utilitarian agent faces several options open to her in a given situation, then *of course* there is at least one option producing best consequences. How could it be otherwise? Still, Hodgson claims to have shown that the act-utilitarian cannot produce best consequences in the situation we have been considering. He seems to sense the paradoxical air surrounding his claim;[15] whence he explains:

> Even though an act-utilitarian might consistently be successful in choosing the act which would have best consequences, this only means that the consequences are the best possible *in the circumstances of his being an act-utilitarian*.[16]

Schematically, Hodgson's picture looks as follows: As long as your moral orientation remains undetermined, you may face a choice between options A, B, ... Z, and there will be at least one option with best consequences, say, Z. However, if you are act-utilitarian, then Z may not be open to you because it can happen that Z consists in (i) not hiding your moral orientation and (ii) making assertions which are likely to be believed. Non-utilitarian agents may be able to combine both features (i) and (ii) from option Z, whereas act-utilitarians cannot possibly do this (Hodgson argues), and thus, have only suboptimal options to choose from, say, A, B, ... M.

Now in a trivial sense Hodgson is right in claiming that obedience to act-utilitarianism limits the number of options. None of the following options is open to an act-utilitarian:

(a) Performing a certain action without having any moral orientation.

(b) Performing a certain action plus entertaining an anti-utilitarian orientation.

(c) Performing a certain action plus truthfully saying "I do not believe in act-utilitarianism".

(d) Performing a certain action plus hiding your true moral orientation while saying "I am an act-utilitarian".

The point here is not that according to her moral orientation an act-utilitarian *should not*, but that according to logic, she *cannot* choose items from the list.[17] And the point is trivial because it can be repeated with regard to *any* moral orientation, say, a Kantian one. So in a certain (trivial) sense it is true that entertaining some moral orientation or other limits your choices. It is true, for example, when the choices are described in terms of those very moral orientations.

But there are several reasons why we had better not describe our options in such terms. First, we do not choose between competing moral orientations in the same way we choose between different courses of action. To confuse the two kinds of choice is a category mistake.[18]

Furthermore, if you describe options (partly) in terms of moral orientations, you may, on the one hand, neglect differences which should matter or, on the other hand, pay attention to differences which should not matter—in the context of utilitarian decision making.

We shall come back to the last point at the end of this section. For the moment, I want to have a closer look at the other point (the danger of neglecting important differences). Let us consider again option Z from our earlier discussion:

Option Z: (i) not hiding your moral orientation, plus
(ii) making assertions which are likely to be believed.

We want to compare a Kantian and a utilitarian, who are alike in all respects apart from moral orientation. Let us suppose each of them has been asked, "Are you obedient to act-utilitarianism?" and let us see what our description of Z tells them to do. The Kantian (choosing Z) would have to say "No!", whereas the utilitarian (*also* choosing Z) would have to say "Yes!"—thus performing a course of action quite opposite to the Kantian one, a course of action moreover, which can lead into very different consequences than those produced by the Kantian. So by itself, our "description" of option Z does not say much about the future conduct of someone opting for Z. Among other things, the description annihilates the crucial difference between "Yes" and "No". And the reason for this is that option Z was described in terms of moral orientation (compare clause (i)).

The lesson from our example can be generalized. In the context of utilitarian decision making we are mainly interested in the causal consequences of different options. In that context, therefore, an option should be described in terms which exhibit its causal impact on the future course of events. Now, an agent's moral orientation (taken in itself) need not be causally efficacious at all—no more than any other item from the agent's belief system. The agent's inward mental activity alone does not affect other people; they are affected only if it is accompanied or followed by overt behavior (verbal or non-verbal). So for utilitarians it is crucial not what an agent thinks or believes but what she does and says. In the context of utilitarian decision making, therefore, options open to an agent should by differentiated, and identified, in terms of overt behavior.[19]

Let me highlight our result from a different perspective. The phrase "overt behavior" is well known from Quine's philosophy of language, more specifically, from his thought experiment of radical translation.[20] This in mind, we can formulate our result as follows. When we think of an act-utilitarian agent and wish to describe the set of options open to her, then we should put ourselves into the position of Quine's radical translator: In principle, if not in practice, we should be able to describe the agent's possible behavior exclusively in terms available to the famous field linguist who knows everything about the agent's overt behavior and nothing about her thoughts, beliefs, orientations, or meanings.

Of course, I cannot prove that this is "the" right terminology for describing options open to act-utilitarian agents. So far, I have only tried to indicate why Quine's naturalism (in the philosophy of language) is a natural choice for act-utilitarian philosophers: The naturalistic regimentation of language à la Quine leaves us with no more and no less than is needed for act-utilitarian calculations of causal consequences. According to naturalism, the world should be described in terms of the causal sciences; the common-place utilitarian conception of actions and their causal consequences fits very well into this naturalistic conception of the world. Act-utilitarianism and Quine's naturalism seem to be made for one another. The deeper reason for this is that both spring from one and the same philosophical worldview: from enthusiastic respect for the achievements of modern science. (Call this worldview "scientism" if you want to show it in a less positive light).

Here is a more concrete reason why act-utilitarian philosophers should welcome Quinean regimentation of language. It offers a convincing diagnosis of what went wrong in Hodgson's paradoxical attack on utilitarianism. Let us remember: Hodgson claimed that in certain situations an act-utilitarian agent cannot choose the best option Z because this is only open to non-utilitarian agents. Now suppose that option Z, as performed by some non-utilitarian agent, indeed maximizes utility: on one hand, the agent utters a true sentence for the sake of coordinating future behavior; on the other hand, she makes clear that she is not a follower of act-utilitarianism. How does she accomplish all this? In Quinean terms, the answer runs as follows. The agent maximizes utility by uttering certain words, accompanied with body language that looks trustworthy.

If this is the proper description of her behavior, however, then it becomes obvious that our act-utilitarian agent can behave exactly the same way: She can utter the very same words and also combine them with body language that looks trustworthy. True, in her case this should not be described as a choice in favour of option Z. But who cares? Her overt behavior cannot be distinguished from that of her non-utilitarian counterpart. Now overt behavior is what matters causally.[21] So from a causal point of view, both the act-utilitarian agent and her non-utilitarian counterpart can perform the very same course of action. If (as Hodgson rightly assumes) the latter can choose an option maximizing utility then the former can do so, too. Hodgson failed to see this because (appealing to moral orientations) he differentiated options which are causally equivalent, and thus, should not be differentiated (in the context of utilitarian decision making).

Our initial intuition regarding Hodgson's argument has proved to be reliable: There is a paradox in his result. For *conceptual reasons*, it is not to be expected that in some unhappy situation there will be no alternative producing best consequences. On the contrary, it is always possible to maximize utility—even for agents obedient to act-utilitarianism. The practical side of act-utilitarianism does not comprise a contradiction.

Is that good news for act-utilitarians? Not exactly. Hodgson's considerations provide only half of what is needed for refuting act-utilitarianism. Instead of trying to prove that act-utilitarian behavior is (sometimes) impossible, I shall try to prove that act-utilitarian behavior cannot be combined with act-utilitarian conviction. Instead of revealing a contradiction within act-utilitarian practice, I want to reveal a contradiction between act-utilitarian practice and act-utilitarian theory. No one, I claim, can at the same time believe in and behave in accordance with act-utilitarianism.

To prove my claim, I shall appeal to the very same doctrines from the philosophy of language which have helped us to defend act-utilitarianism against Hodgson. Roughly, the idea behind my refutation is this. When we describe options open to an act-utilitarian agent in terms of overt behavior (that is, from the standpoint of radical translation), then this amounts to starting without any assumption concerning the agent's beliefs, opinions, orientations—and meanings. According to Quine, the proper understanding of the agent is a function of her overt behavior.[22] If the agent's verbal and non-verbal behavior maximizes utility, then (I shall claim) her utterances cannot be interpreted so as to ascribe to her those beliefs that constitute utilitarian theory. In short: Nobody maximizing utility can possibly be said to believe in act-utilitarianism; act-utilitarianism is indeed incoherent—but in a more subtle way than Hodgson tried to prove.

III. Three Preliminary Thought Experiments

Before we can proceed to the proof of my claim with respect to act-utilitarianism, I want to demonstrate analogous claims 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 later sections of this paper we shall seek to detect the same sort of incoherence in act-utilitarianism.

First example. Imagine a speech community whose members subscribe to a moral norm which 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. More concrete, it is due to 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 which you ascribe to the members of the interpreted community.[23]

Equipped with this principle, we can see what kind of mistake I made when describing the community which 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" which was ascribed to the natives under the original interpretation.[24] (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 which 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 speakers' 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 which lies behind the truism that meaning is use.[25] 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 which 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; then this interpretation is refuted, should it turn 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. (For simplicity, let us assume that in the community concerned there is but one linguistic form for expressing promises: "I promise to do X"). 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 which might be called an *aleatoric promise*. Notice that this label is a little misleading; we should not think of aleatoric promises as being a species of promise. On the contrary, an aleatoric promise is no promise at all.

One might object: Why shouldn't it be a promise? Couldn't an aleatoric 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 practice which is more complex than the aleatoric 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.

So suppose that a member of the imagined community has given an aleatoric 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 aleatorically 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 aleatorically 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 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 aleatorically. 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.

IV. Can an Act-Utilitarian Make Promises or Assertions?

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 aleatorically but utilitaristically: Whenever she promises to do X, she does X only if this maximizes (expected) utility.[26] 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, i.e., 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 act-utilitarianism. And this means that she will have to *systematically* break her promises; she is not complying with the rules which define the speech act type *promise*. She does not even intend to comply with these rules; her intention is directed towards maximizing utility.

How are we to interpret the alleged promises of our act-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 act-utilitarian speaker is not unreliable in general. True, she is an unreliable promise-keeper, but she remains a reliable act-utilitarian none the less. 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 which is governed by act-utilitarianism, because act-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 practise act-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.

So suppose our act-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 act-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 act-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 which 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 act-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.[27] 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 which are somewhat similar to what our act-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 act-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 act-utilitarian speaker cannot express genuine promises, it follows that she cannot possibly apply her moral theory to promises she herself has given.[28] The utilitarian rule "Break your promises if doing so maximizes utility" is incoherent.

This will not yet impress the act-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, act-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 act-utilitarianism itself can afford.

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

Let me first provide you with some intuitive evidence in favour 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.[29]

Now, we 've just seen that consistent act-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 act-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 act-utilitarian says something which sounds like an assertion from our language, e.g., "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 which reliably govern the speaker's usage of the words "There is lots of mineral water in Karlovy Vary"? Because our speaker is an act-utilitarian, there is but one rule which she is following, in all her verbal and non-verbal behavior: act-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 which sound like assertions are in no way different from those governing utterances which sound like promises. What's more, due to the speaker's act-utilitarianism, *every* utterance X from her idiolect reliably follows a rule of the very same form:

> It is correct to utter X 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 act-utilitarian speaker. It is a kind of speech act which we cannot express but only circumscribe in our language. As we have seen, the speech act in question aims at the people's happiness, thus on the one hand resembling the speech act of consolation while on the other differing from it by being more general.[30]

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 practising act-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 act-utilitarian's speech acts really so different? There are at least seven objections which might be raised against my claim.

V. Objections

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, i.e., 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 act-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 act-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 act-utilitarian agent, whom we want to interpret, within a speech community whose members are all ideal act-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.[31] But if, so the objection proceeds, the ideal act-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;[32] 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 an act-utilitarian, would have to follow the linguistic rules of the speech community in question—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 act-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. And it is these cases which call for an interpretation more charitable than the standard one.

Fifth objection. Couldn't it be that act-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 act-utilitarianism wouldn't work. But as far as I can see, there is not the slightest reason to believe in the preestablished harmony which my opponent is invoking.[33] And anyway, many act-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.

Sixth objection. Of course there is no preestablished harmony between the goal of utility and that of truth; aiming at utility does not always and not necessarily lead to correct assertions only. But it could still be the case that an ideal act-utilitarian agent might follow both the *strategic* aim of maximizing utility and the *tactical* aim of making truthful assertions (as one way among others of working towards that strategic aim). To speak truthfully could be a *subsidiary* disposition that one has to develop in order to maximize utility.—Reply. The objection should not be understood as recommending the disposition to truthfulness merely as rule of thumb, helpful only to limited beings like us. Remember that our discussion takes place on the ideal level of ethical thought. So the tactical aim of speaking truthfully is supposed to amount to more than a rule of thumb. It must play a stronger role in the deliberations and motivations of our ideal agent, and yet at the same time, it must be weak enough to allow being silenced whenever the strategic aim of maximizing utility conflicts with it. But if, in the case of conflict, such a tactical aim is always silent, and if in the case of non-conflict, the tactical aim anyway calls for the same conduct as the strategic aim—then we can as well assume that the tactical aim plays no genuine role at all in the deliberations of the agent concerned. (How could Quine's field linguist tell that the agent really entertains the supposed tactical aim?)

Seventh objection. The reply to the preceding objection is not sufficiently sensitive to the phenomenon of *exceptions*. If someone does not *always* pursue a certain goal but sometimes acts against it, then this does not show that he does not have that goal. So despite the exceptions (that are demanded by the strategic aim of maximizing utility), our act-utilitarian agent could still have the tactical aim of truthfulness. After all, exceptions prevail everywhere—and yet there exist aims, rules, practices, institutions, and so forth. For example, our own (non-utilitarian) institution of assertoric language use does not demand that we *always* speak the truth and nothing but the truth; our assertoric practice allows for exceptions such as making a joke, being sarcastic, or telling a lie so as to not hurt someone's feelings. If the institution of assertoric language use can survive a certain range of exceptions then (the objection concludes) this institution may also be compatible with the exceptions to truthfulness demanded by act-utilitarianism.[34]

Reply. The objection is right in complaining that so far my reasoning depended on an unrealistic picture as to the prevalence of truthfulness in our own linguistic practice. In order to defend my point I have to do more than revealing essential differences between an act-utilitarian speaker and speakers who are *always* perfectly truthful; I have to reveal essential differences between an act-utilitarian speaker and *us*, who are far from perfection with regard to truthfulness. So I have to make plausible the claim that our own, actual dispositions concerning the utterance of assertions are radically different from those of an act-utilitarian speaker. This is the burden of the next section, where we will be forced to speculate about a new version of Quine's indeterminacy of translation.[35] The basis for these speculations will be more decidedly Quinean than the considerations from the other parts of this article. Up to now, we have not been invoking much more than Quine's principle of charity. But the present objection is so strong that to meet it we need to become more specific in our assumptions from the philosophy of language. Let me emphasize that Quine's philosophy provides only one of the suitable frameworks for the my argument. (For reasons of limited space we shall not explore alternative frameworks).

VI. Indeterminacy of Translation: A New Case?

According to the final objection from the preceding section, an act-utilitarian's speech dispositions can be interpreted assertorically. Even if this were so, it could still be true that her speech dispositions can *also* be interpreted in terms of our new (utilitarian) speech act type; both interpretations could be right. In this case we would have detected a fresh and radical example for Quine's indeterminacy of translation (not so much an example for the indeterminacy of translating *propositional content* but rather an example concerning *speech act types*). Let us for a moment postpone the question whether the present case really allows for such radical indeterminacy. (I shall deny this in the end). Let us better ask: Is it at least *possible* to interpret the act-utilitarian non-assertorically, namely in the utilitarian way we have indicated?

I think we can easily see that such an interpretation is feasible. For one of the criteria for the feasibility of interpretations demands that they should help in predicting the interpretee's verbal reactions. And the utilitarian interpretation is perfect for predicting future utterances of our speaker. This is so because the whole interpretation comprises just one condition for correct utterances from the speaker's language:

(U) It is correct to utter X only if doing so maximizes utility.

And indeed you do not need more information about the speaker (you do not need a better interpretation, that is) if you want to predict her verbal reactions. The "only" additional bit of information you need is perfect causal knowledge about the rest of the world.[36]

So our (utilitarian) interpretation, as given by rule (U), is feasible. But even if this is right, I have to admit that more needs to be shown for completing my argument against act-utilitarianism. What I must show is that we cannot find an assertoric interpretation for the act-utilitarian speaker. That's a bigger claim than the mere possibility of interpreting her otherwise; *several* interpretations might be feasible. (Remember Quine's indeterminacy of translation!)

Such indeterminacy is precisely what the final objection from the preceding section implies. If the objection is right then the act-utilitarian's speech dispositions are so similar to our own speech dispositions that they allow for an assertoric interpretation. (According to the objection, one aspect of the similarity concerns exceptions: Neither the act-utilitarian nor our own practice of making assertions should be seen as total submission to truthfulness). Yet I think the objection tries to make the similarity between us and the act-utilitarian bigger than it is, which has implausible consequences. For similarity is a *symmetric* relation. If an act-utilitarian's speech dispositions are sufficiently similar to our own speech dispositions, so that she can be (like us) interpreted as making assertions, then, symmetrically, our dispositions should be sufficiently similar to hers, so that *we* can be (like her) interpreted as making utterances in accordance with rule (U). (And again, we should be prepared to grant some exceptions).

In short, if the objection were right then our new indeterminacy of translating speech act types would not only affect act-utilitarian speakers; it would extend to our own language—it would "begin at home".[37] But this is not plausible. Even if we indulge ourselves in some exceptions, it is obvious that we cannot be understood as following the one rule that constitutes act-utilitarian language use:

(U) It is correct to utter X only if doing so maximizes utility.

Our own (assertoric) language use is more complex and more liberal than is mirrored by that rule. To see this, suppose someone were told that you are to be interpreted along the lines of rule (U); give him in addition all causal knowledge about (the rest of) the world that he wishes—would he be able to predict even a decent portion of your actual verbal reactions? I submit we must deny this for two reasons.

First of all, in the vast majority of cases your utterances will not produce best consequences. Notice that normally there will be just *one* utterance producing *best* consequences,[38] while you tend to choose rather carelessly between different verbalizations of quite different thoughts. It would be tantamount to winning the first price in the lottery if you succeeded regularly in producing best consequences while you speak. So the exceptions to (U) will be the rule, and most of the predictions guided by (U) must fail. Secondly, rule (U) is by far too crude; it cannot do justice to all those fine patterns, little nuances, and surprising idiosyncrasies which make up your individual web of belief. There is no hope that the expression of your beliefs can be subsumed under any short rule—let alone under rule (U). (If, by contrast, someone understands your language in the standard way and knows what you believe, then he will be rather successful in predicting your verbal reactions. True, he need not be able to predict all your jokes, sarcasms, and lies. But in many situations he will be able to tell in advance how you will react on questions).

I conclude that our own speech dispositions cannot be feasibly interpreted along the lines of rule (U). Our dispositions are not similar enough to those of an act-utilitarian speaker (whose verbal behavior is well reflected by rule (U)). Our way of using language leaves no room for an indeterminacy of translating speech act types (or anyway not for the kind of indeterminacy that we have been trying to imagine). But, for reasons of symmetry, there would have to be such an indeterminacy—if the act-utilitarian speaker could be feasibly interpreted as making assertions. Therefore and contrary to the final objection from section V, she cannot be so interpreted. Ideal agents who are obedient to act-utilitarianism cannot make assertions.

VII. Act-utilitarianism Refuted

The lesson from the preceding sections is this: It seems reasonable to think that an ideal agent, who behaves in accordance with act-utilitarianism, cannot possibly express assertions. What is more, she cannot possibly make any of the speech acts *we* are familiar with. As we have seen, the only kind of speech act open to her is something which 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 act-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 act-utilitarianism. But to my taste they depend too much on debatable intuitions about the nature of language. Let us try to proceed on dialectic ground which is firmer than that.

In addition to what I have shown so far, I need one more premise to *refute* act-utilitarianism. The additional premise is this: Anyone fully subscribing to act-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 an act-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 an act-utilitarian you need to have specifically utilitarian intentions: you must do what you do *because* you believe that your very action will maximize utility.[39]

At least, an *ideal* agent has to entertain beliefs of this kind—if she is to qualify as an act-utilitarian. Although our discussion anyway 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 which 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 act-utilitarianism. On the one hand we have shown that an ideal agent who is an act-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, e.g., 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 none the less. 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 act-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 which are forever excluded from public access.[40] (Even an omniscient radical interpreter could not guess at the contents of those alleged beliefs!)

If this is right, act-utilitarianism must break down. As we have seen at the opening of our discussion, act-utilitarianism cannot be practised in everyday life.[41] But now we know that things look even worse for act-utilitarianism on the ideal level of ethical thought. On the ideal level, act-utilitarianism is incoherent: If an ideal agent really always acts in accordance with act-utilitarianism, even while she speaks, then she cannot entertain the beliefs necessary for being motivated by that very same act-utilitarianism. In short, if an ideal agent practically subscribes to act-utilitarianism, then she cannot subscribe to it theoretically as well. The practical side and the theoretical side of act-utilitarianism do not agree.[42]

Notes

[1] My main target are utilitarian theories (in the narrow sense) which measure consequences either in terms of positive mental states (such as feelings of happiness) or in terms of fulfilled preferences of human beings. (Theories combining these two approaches will be under discussion, too). Although it might be an interesting task, I shall not explore the possibility of extending my argument to consequentialist theories (that is, to utilitarian theories in the broad sense). So we shall not be concerned with theories which include, say, knowledge, or truthfulness, or friendship into the list of things that are good in itself.—Notice, in addition, 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 preceeding footnote.—Furthermore, for brevity, I shall sometimes 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 e.g. with respect to the past in order to find out what your act-utilitarian obligation consists in). But although less than omniscience is needed, it seems clear that a successfully practising act-utilitarian must know much more than what 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 act-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, i.e., a theory about the right and the wrong, rather than a theory about, say, the *ight* and the *wong*. Thus, imagine a tribe of act-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 Hare (1981), pp. 25ff. I shall concentrate on Dieter Birnbacher's more recent version of the same strategy, cf. Birnbacher (1988), pp. 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 which do not aim at practical significance but content themselves with a theoretical claim only. According to these—externalist—versions of act-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] Compare, e.g., Hodgson (1967), p. 3 and p. 51. He makes clear from the outset that he is not interested in cases of "misapplication" of utilitarianism (p. 3); this already indicates that he wants to abstract from both our motivational and epistemic limitations. Later he says: "We assume that he [i.e., the act-utilitarian agent] attempts always to act in accordance with the act-utilitarian principle, and that he is highly rational" (p. 51).

[9] Cf. Hodgson (1967), p. 41. This is the core of Hodgson's argument; it concerns two act-utilitarians in interaction. Compare p. 51-53 for the parallel argument concerning an act-utilitarian who interacts with non-utilitarians.

[10] Cf. Hodgson (1967), p. 42.

[11] Cf. Hodgson (1967), pp. 42-44, 53.

[12] Cf. Hodgson (1967), pp. 51-53.

[13] Cf. Hodgson (1967), p. 60 and p. 3.

[14] As Singer rightly observes, there is no inconsistency in the two statements: (i) "An act is right if and only if it would have best consequences" and (ii) "If *people* accepted (i) it would not have best consequences, even if *they* applied it correctly" (cf. Singer (1972), p. 94; my emphasis). An inconsistency only arises when we change (ii) into (ii*) "If *I* accepted (i) it would not have best consequences, even if *I* applied it correctly". (If act-utilitarianism demands you to produce best

consequences, yet at the same time leads you to do worse, then—*because* of act-utilitarianism—you should not be act-utilitarian indeed. In short: If act-utilitarism were right, then it would be wrong; so it cannot be right). Therefore, Hodgson's case is stronger when applied to individuals instead of societies. Consequently, I shall mainly discuss passages from Hodgson's book where he concentrates not on an act-utilitarian society but on an act-utilitarian individual. (For a critique of Hodgson's considerations concerning the interaction of several act-utilitarians, compare Singer (1972) and Lewis (1986)).

[15] Cf. Hodgson (1967), p. 3.

[16] Cf. Hodgson (1967), p. 58, Hodgson's italics. Compare also p. 45 for the parallel point with respect to an act-utilitarian society.

[17] Hodgson, too, appeals to logic in the present context, cf. Hodgson (1967), p. 58.

[18] As already quoted, Hodgson speaks of the "*circumstances of [...] being an act-utilitarian*", cf. Hodgson (1967), p. 58. This is another category mistake.

[19] For simplicity, I shall neglect the following complication in the main text: Even without making a detour through overt behavior, sometimes an agent's mental activity can very well influence general happiness—it does so in case it produces happiness *in the agent herself*. We can neglect this complication because most often the agent's happiness is of minor importance when the goal is *general* happiness. (However, the complication is crucial in case of ethical egoism; so my criticism of Hodgson's anti-utilitarian argument cannot be straightaway extended to Hodgson's anti-egoistic argument, cf. Hodgson (1967), pp. 60-62).

[20] Cf. Quine (1960), chapter 2; compare also Lewis (1974).

[21] Is this hardcore behaviorism? Fortunately not. My point does not depend on the dubious claim that there is no inward mental activity but only stimulus and response. What I need is less controversial: If an agent's mental activity (no matter how it has to be analyzed) is supposed to have causal effects on other people, than there must be elements in her overt behavior that produce the very same effects. (We cut off, as it were, the earlier phases of the causal chain in question). In sum, focussing on overt behavior gives us all we need for making an act-utilitarian decision.

[22] If you prefer to put the same point less Quinean you might say that meaning is a function of use. Compare footnote 25.

[23] Cf. Quine (1960), p. 59; compare also Davidson (1984), p. 196.

[24] A similar interpretation has been considered by Lewis (1986), p. 342.

[25] The truism goes back to Wittgenstein (1984), § 43.

[26] 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.

[27] 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, i.e., 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 which are decisive for X's maximizing utility.

28 We are assuming a speaker who is utilitarian throughout her life.

29 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.

30 As far as I know, Hilary Putnam is the only philosopher who has claimed that the language of an act-utilitarian speech community cannot be taken at face value but needs to be reinterpreted, cf. Putnam (1981), pp. 139-41. Putnam makes his point not with regard to those linguistic devices that indicate speech act types. Rather, he considers words for so-called thick ethical concepts such as "honest" (*op. cit*., p. 140), because in his dialectical context he wishes to show that facts and values cannot be disentangled (*op. cit*, p. 141). I agree with this meta-ethical conclusion. But in the present paper, my ambition is directed toward first-order normative ethics (to the refutation of utilitarianism, that is). Despite of this divergence, I think that my goal agrees nicely with the spirit of Putnam's book. Putnam himself hints at a parallel between the case of consistently practising act-utilitarians and his famous case of the brains in the vat (*op. cit*., p. 139). According to Putnam, both the language of envatted brains and that of act-utilitarians must be radically different from our language. In the case of the envatted brains, Putnam derives from this the impossibility of our eternal envatment (*op. cit*., pp. 1-21). But he does not arrive at the parallel negative result with regard to act-utilitarianism. Still, the impossibility of act-utilitarianism fits well into Putnam's doctrines from *Reason, Truth and History.*

31 A similar point is made by Singer (1972), p. 97.

32 Thus, unlike Putnam (1981), pp. 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 act-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 act-utilitarian language as a social institution? Happily the answer is to the positive. Remember that my target is *act-utilita*rianism. 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 act-utilitarianism. (Compare footnote 14). Therefore, we can leave it open how the language of an act-utilitarian speech community might change in time.

33 A harmony between the goal of truth and that of maximizing *utility* (in the broad sense) might occur in the case of (consequentialist) theories which say for example that knowing the truth is itself a component of utility. Such theories are beyond the scope of this paper, see footnote 1.—According to (certain versions of) preference utilitarianism, knowing the truth might be a contingent component of utility, if, for example, the agents' preferences are in favour of knowledge. So

why shouldn't preference utilitarians appeal to such preferences in order to escape from my argument? If they did then their version of utilitarianism committed them to certain *empirical* claims. I take it that a moral theory should hold good in each possible world. (That's why we often invent strange counterfactual stories in order to test moral theories). But this might be controversial. The issue belongs to metaethics and goes beyond the scope of this paper. (I am indebted to Timothy Chappell for pointing at these complications).

[34] This objection and the preceeding one are credited to Timothy Chappell.

[35] The *locus classicus* is Quine (1960), chapter 2. Compare also Quine (1990).

[36] Don't worry about this idealization; it is no less realistic than the whole thought experiment of radical translation. Quine and his followers operate with an interpreter who knows every fact about the world apart from "facts" about the speaker's thoughts, beliefs and meanings, compare e.g. Lewis (1974), p. 331. And the goal of interpretation is accomplished when it systematizes the facts about the speaker in such a way that someone, who knows nothing else about the speaker and who knows everything about the rest of the world, can predict what the speaker will reply to any arbitrary question.—Notice that this Quinean idealization agrees well with the sort of utilitarian idealization which we have discussed at the outset. Both utilitarianism and Quine's philosophy of language share a certain style of thinking.

[37] This in an echo of Quine's famous claim that "radical translation begins at home", cf. Quine (1969), p. 46. After having revealed an indeterminacy in translating "gavagai" ("rabbit" versus "undetached rabbit part" cf. *op. cit.* p. 29-34), he makes clear that the problem carries over to our own word "rabbit" (*op. cit.* p. 46/7). It is beyond the limits of this paper to rehearse the arguments that can be advanced for and against extending the indeterminacy thesis to our home language. Suffice it to say that many philosophers believe that the thesis can be so expanded and that if this is right the same holds good for its counterpart from my text.

[38] This is most obvious in case of utilitarian theories that assign *cardinal* numbers to the value of consequences.

[39] We can leave it open whether you also have to believe in act-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, e.g., Ayer (1946), chapter VI.

[40] Here is a possible counterexample to my claim which 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 practising act-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.

41 By the way: If, in the end, you do not find the arguments from section I against the practicability of act-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.

42 This is a modified version of a paper presented on September 16th at the *Fifth Karlovy Vary Symposion on Analytic Philosophy (Swimming in XYZ, Supervised by Hilary Putnam,* September 14th-18th, 1998). German versions of the paper were presented at Georg-August University Goettingen, Duesseldorf University, and Free University Berlin. I should like to thank the symposion's participants as well as Thomas Schmidt, Kathi Koellermann, Tatjana Tarkian, Joerg Schroth, David Hyder, and an anonymous referee for various suggestions that helped me to improve the paper. I am most grateful to Hilary and Ruth Anna Putnam for their generous encouragement.

References

Ayer A. J. (1946), *Language, truth and logic.* London: Victor Gollancz, second edition.

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.

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 (1974), 'Radical interpretation', in *Synthese* 27 (1974).

Lewis, David (1986), 'Utilitarianism and truthfulness', in Lewis, *Philosophical Papers: Volume II.* Oxford: Oxford University Press.

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

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

Quine, W. V. (1969), 'Ontological relativity', in Quine, *Ontological relativity and other essays.* New York: Columbia UP.

Quine, W. V. (1990), 'Three indeterminacies', in Barrett, Robert / Gibson, Roger (eds.) *Perspectives on Quine.* Cambridge / Mass.: Blackwell.

Singer, Peter (1972), 'Is act-utilitarianism self-defeating?' In *The Philosophical Review* LXXXI.

Wittgenstein, Ludwig (1984), *Philosophische Untersuchungen*, in Wittgenstein, *Werkausgabe Band 1.* Frankfurt: Suhrkamp.

Wright, Crispin (1994), 'On Putnam's proof that we are not brains in a vat', in Peter Clark and Bob Hale (eds.), *Reading Putnam.* Cambridge / Mass.: Blackwell.