Australasian Journal of Philosophy
Vol. 50, No. 1; May, 1972

DAVID LEWIS

UTILITARIANISM AND TRUTHFULNESS[1]

A demon has seized two highly rational act-utilitarians—call them 'You' and 'I'—and put them in separate rooms. In each room there are two buttons, a red one and a green one. The demon has arranged that by both pushing our red buttons or by both pushing our green buttons we bring about the Good; but by pushing one red button and one green button (or by pushing both buttons or neither button in one of the rooms) we bring about the Bad. The demon has made sure that we both know all the facts I have listed so far, that we both know that we both know them, and so on.

You manage to send me a message, and the message is 'I pushed red'. But, strange to say, that does not help. For I reason as follows. 'You are a highly rational utilitarian. You act in whatever way you think will have the best consequences, with no regard to any other consideration. This goes for sending messages: you send whatever message you think will have the best consequences, caring not at all about truthfulness for its own sake. So I have not the slightest reason to believe your message unless I have reason to believe that you think that truthfulness will have the best consequences. In this case, you must know that truthfulness has the best consequences only if I have some reason to believe you and to act accordingly. If not, there is nothing to choose between the expected consequences of truth and untruth, so you have no reason whatever to choose truth rather than untruth. I have not the slightest reason to believe you unless I have reason to believe that you think that I have reason to believe you. But I know that you—knowledgeable and rational creature that you are—will not think that I have reason to believe you unless I really do have. Do I? *I cannot show that I have reason to believe you without first assuming what is to be shown: that I have reason to believe you.* So I cannot, without committing the fallacy of *petitio principii*, show that I have reason to believe you. Therefore I do not. Your message gives me not the slightest reason to believe that you pushed red, and not the slightest reason to push red myself.' Arguing thus, I push at random. By chance I push green.

Such is the disutility of utilitarianism, according to D. H. Hodgson.[2]

We might better say: such is the disutility of *expecting* utilitarianism, and it is not sufficiently compensated by the efforts to maximize utility that fulfil the

[1] This research was supported by a fellowship from the American Council of Learned Societies.

[2] *Consequences of Utilitarianism* (Oxford University Press: Oxford, 1967), pp. 38-46.

expectation. Hodgson says that knowledgeable and rational act-utilitarians would have no reason to expect one another to be truthful, not even when the combination of truthfulness with expectation of truthfulness would have good consequences; so they would forfeit the benefits of communication. Similarly they would forfeit the benefits of promising; for an example of this, just change the message in my example to 'I will push red'. More generally, it seems that Hodgson's utilitarians would forfeit the benefits of all the conventions whereby we coordinate our actions to serve our common interests. The conventions of truthfulness and of promise-keeping are but two of these.

But to talk myself into ignoring your message 'I pushed red' is absurd. My example has no special features; it is just a simple and stark instance of the general situation Hodgson says would prevail among knowledgeable and rational act-utilitarians. I conclude that Hodgson is wrong in general. Where, then, is the flaw in my Hodgsonian argument that I ought to ignore your message? Every step up to the italicised one seems true, and every step beyond that seems false.

I think the argument went wrong when I tacitly assumed that I could not have reason to believe you unless I could show, using nothing but the facts set forth in the first paragraph—our situation, our utilitarianism and rationality, our knowledge of these, our knowledge of one another's knowledge of these, and so on—that I did have reason to believe you. But why must my premises be limited to these? I should not use any premise that is inconsistent with the facts of the first paragraph; but there is nothing wrong with using a premise that is independent of these facts, if such a premise is available.

The premise that you will be truthful (whenever it is best to instill in me true beliefs about matters you have knowledge of, as in this case) is just such a premise. It *is* available to me. At least, common sense suggests that it would be; and our only reason to suppose that it would not is the Hodgsonian argument we are now disputing. It is independent of the facts listed in the first paragraph. On the one hand, it is *consistent* with our rationality and utilitarianism, our knowledge thereof, and so on. For if you are truthful (except when it is best that I should have false beliefs), and if I expect you to be, and if you expect me to expect you to be, and so on, then you will have a good utilitarian reason to be truthful. You will be truthful without compromising your utilitarianism and without adding to your utilitarianism an independent maxim of truthfulness. On the other hand, it is not *implied* by our rationality and utilitarianism, our knowledge thereof, and so on. For if you are systematically untruthful (expect when it is best that I should have false beliefs), and if I expect you to be, and if you expect me to expect you to be, and so on, then you will have a good utilitarian reason to be untruthful. I am speaking, of course, of truthfulness and untruthfulness *in English*; I should mention that systematic untruthfulness in English is the same thing as systematic truthfulness in a different language *anti-English*, exactly like English in syntax but exactly opposite in truth conditions.

Therefore I should have decided that I did have reason to believe your message and to push red myself. This reason is admittedly not premised

merely on our situation, our rationality and utilitarianism, our knowledge of these, and so on. But it is premised on further knowledge that I do in fact possess, and that is perfectly consistent with these facts.

Received May 1971

Princeton University