Would rational act-utilitarians keep their agreements? Suppose two highly rational act-utilitarians agree to meet for a walk in the park. The best outcome they can achieve is to meet as agreed, and the next best is for both to stay home and read. Because each would find it distressing to come and not find the other, the worst outcome they can achieve is for one to come to the park and the other to stay at home. Suppose everything above is common knowledge between them, in the sense that each knows it, each knows the other knows it, each knows the other knows that he knows it, and so on ad infinitum. Under these circumstances, either person would keep the agreement if he knew the other would, for he could then conclude that his coming as agreed would yield the best outcome, that they meet, whereas his staying at home would yield the worst outcome, that one comes and the other stays home. On the other hand, either would break the agreement if he knew that the other would break it, for in that case he could reason that if he came as agreed, he would achieve the worst outcome, that one comes and the other stays at home, whereas if he stayed home, he would achieve the intermediate outcome that both stay home. All this will be common knowledge between the two. Does any of it, though, give either of them grounds for keeping the agreement?

According to D. H. Hodgson (1967, pp. 38–50), neither would have any more reason for keeping the agreement than he would have for coming to the park without an agreement — than he would have if meeting at the park had only been mentioned as a possibility. Precisely because both are highly rational, knowledgeable act-utilitarians, he argues, agreements between them are pointless: agreeing to an arrangement will have no effect on what they actually do. Even if the consequences of meeting would be splendid, they will have no way of bringing these consequences about.

The lesson of this case applies to whole societies. According to Hodgson, if by common knowledge a society consisted of highly rational act-utilitarians, that society would fall apart, because its members would be unable to
coordinate their actions by agreement. Such act-utilitarians would produce few good consequences indeed.

David Lewis (1969, Chs. I–III; 1972) argues that nothing so surprising is the case. He concludes, in effect, that if the society of the two who agreed to meet for a walk in the park had a history of agreement-keeping, their agreement to meet would give them each good reason to come to the meeting-place as agreed. I shall discuss the arguments both of Hodgson and of Lewis in the next section. It is Lewis's argument that provides the starting point for the work in this paper.

The question I want to ask is whether, in the circumstances Hodgson considers, the inability of act-utilitarians to make agreements binding would cause any loss of good results. Suppose a society consists of highly rational act-utilitarians, and this fact is common knowledge among them. Call a society of this kind openly act-utilitarian. Now imagine that members of the society could somehow bring it about — perhaps by magic — that any agreement they made would be kept. There are many agreements they would then avoid. They would make no agreement they could improve: they would reject an agreement if the consequences of making it binding would not be as good as the consequences of making some alternative agreement binding. Call the agreements they might indeed make in these circumstances optimal. An optimal agreement, then, is an agreement act-utilitarians might make if they could somehow make binding any agreement they chose. This term defined, my question now concerns an openly act-utilitarian society of a less astonishing kind, where there is no magical way of making agreements binding. Are there, I shall ask, optimal agreements that, if made in such an openly act-utilitarian society, would be broken no matter how extensive the history of agreement-keeping in that society might be?

The answer may be crucial to the appeal of act-utilitarianism. Act-utilitarianism presumably appeals to those who think that morality is grounded in the production of good consequences. Now suppose that an openly act-utilitarian society would attain less good than would that same society if it had a morality that made optimal agreements binding independently of consequences. That may give someone who thinks morality a matter of doing good a reason for preferring a morality which makes optimal agreements binding. This line of attack on act-utilitarianism fails if it is indeed the case that, in an openly act-utilitarian society, any optimal agreement can be made as good as binding by a suitable history of agreement-keeping.

In this paper, the answers I give to my question are preliminary. I shall state conditions for a practice to be, in a sense, self-perpetuating in an openly