Steven T. Kuhn
Department of Philosophy
Georgetown University
Washington, D.C. 20057

Agreement-Keeping and Indirect Moral Theory[1]

On a social contract view, all moral action is agreement-keeping and agreement-keeping is a means of securing mutual advantage. The second half of this claim is plausible even if one denies the first: even if one believes that morality is not always (or even generally) mutually beneficial in the sense social contract theorists have elucidated, it still seems reasonable that the portion dealing with agreement-keeping should be. Mutual benefit, one would think, is the *raison d'etre* of agreements. This paper discusses two problems with the thesis that the behavior morality dictates for parties to an agreement is to their mutual advantage. The first of these problems seems not to have been previously identified. The second has been, but proposed solutions fall short. The ethics of agreement-keeping, I suggest, is best accommodated by what has been called *indirect* moral theory: to determine what we should *do*, we should ask what we should *teach*, to ourselves and others. The behavior that moral theory prescribes to contractors is *not* mutually advantageous, but the prescription itself may be.

I. Morality, advantage and fidelity

The classic account of how morality may serve to benefit its practitioners is David Gauthier's "Morality and Advantage"[2] (henceforth *M&A*). *M&A* maintains that the prisoner's dilemma game illustrates how it is possible for a system of principles to be "advantageous to

everyone if everyone acts on it" and yet to require "that some persons perform disadvantageous acts." The prisoner's dilemma game can be summarized by the payoff matrix shown below.

		2	
		a	v
1	a	(3,3)	(0,5)
	v	(5,0)	(1,1)

Two players, Ms. One and Mr. Two, are each faced with a one-time choice of performing actions a or v. Their choices are independent: neither can know the other's choice before making his own. In Gauthier's version, action a is adhering to an arms control treaty and action v is secretly violating the treaty. Each square in the matrix contains a *payoff pair* representing the value to One and Two, respectively, of their jointly performing the actions corresponding to that pair. For example, if One chooses to adhere to the treaty while Two violates it then the resulting payoff pair is (0,5), indicating that the state resulting from the joint actions has zero units of value for One and five for Two.

The game illustrated by the above matrix has two important characteristics:

1. Both One and Two benefit more by mutual adherence than by mutual violation. (Mutual adherence gives each three units, mutual violation gives each one unit.)

2. No matter what Two does, One loses by choosing adherence over violation. (If Two adheres, One loses two units by adhering; if Two violates, One loses one unit by choosing

adherence.) Similarly, no matter what One does, Two loses by adhering rather than violating.

The second characteristic implies that mutual violation is the only *nash equilibrium*--i.e., the only outcome such that neither player would benefit by choosing another action, given the action of the other player. Thus players whom Gauthier labels *prudent*, i.e., players whose action choices are based solely on personal advantage, will violate. But according to the first characteristic, then, the state that will be reached by prudent players is *pareto suboptimal*. Both players would be better off if they both adhered. This explains the sense in which each person can benefit by everybody's following a principle that requires some (and in this case all) to perform "disadvantageous" acts. Gauthier calls a person who is willing to fulfill agreements, the common fulfillment of which would be to his or her advantage a *prudent but trustworthy* person. The example shows that a group of prudent but trustworthy persons can obtain advantages for themselves that a group of purely prudent persons can not. The prudent but trustworthy persons who play prisoner's dilemma with one another other each get three units; the purely prudent persons get only one.

Morality, according to *M&A*, is partly--but not wholly--a system of rules of the seemingly paradoxical character that the prisoner's dilemma game clarifies. One respect[3] in which morality goes beyond what is required by any such system, *M&A* notes, is its apparent insistence on *trustworthiness*--agreements must be kept even when mutual adherence is not mutually advantageous in the prisoner's dilemma sense. The term "trustworthy" as a label for the characteristic of keeping agreements through thick and thin has some drawbacks. It suggests that the characteristic is marked by something deserved from others rather than by something one has or does oneself. And it suggests something always positive: it does not seem possible to be too

trustworthy (though others can certainly be too trusting of us). I do not wish to prejudge the issue of whether one is obligated to keep every agreement, so I will use the more neutral terms "faithful" and "fidelity". A person is *faithful* (has *fidelity*) to the extent that he or she keeps agreements[4]. A person is *strongly faithful* (has *strong fidelity)* if he or she keeps agreements even when they are not, in an appropriate sense, mutually advantageous. One question raised by Gauthier's paper, then, is whether morality requires strong fidelity. Put another way, the question is this: when fidelity conflicts with advantage does morality side with the former or the latter?

II. The problem of impure dilemmas

A game in which conditions 1 and 2 above hold is a prisoner's dilemma. In any agreement with the structure of a prisoner's dilemma, both parties benefit more by mutual adherence than by mutual violation. One might suppose that under these conditions mutual adherence is pareto optimal--i.e., that there is no outcome that both parties prefer to mutual adherence. In fact, for the game matrix shown above, that supposition is correct. If one allows *mixed* strategies, however, whereby each player can adhere with probability p and violate with probability 1-p, then there are prisoner's dilemmas in which mutual adherence is not pareto optimal. Suppose, for example, that the payoff for violating when the other party adheres is seven units rather than six, so that the matrix looks like this.

		2	
		a	v
1	a	(3,3)	(0,7)
	v	(7,0)	(1,1)

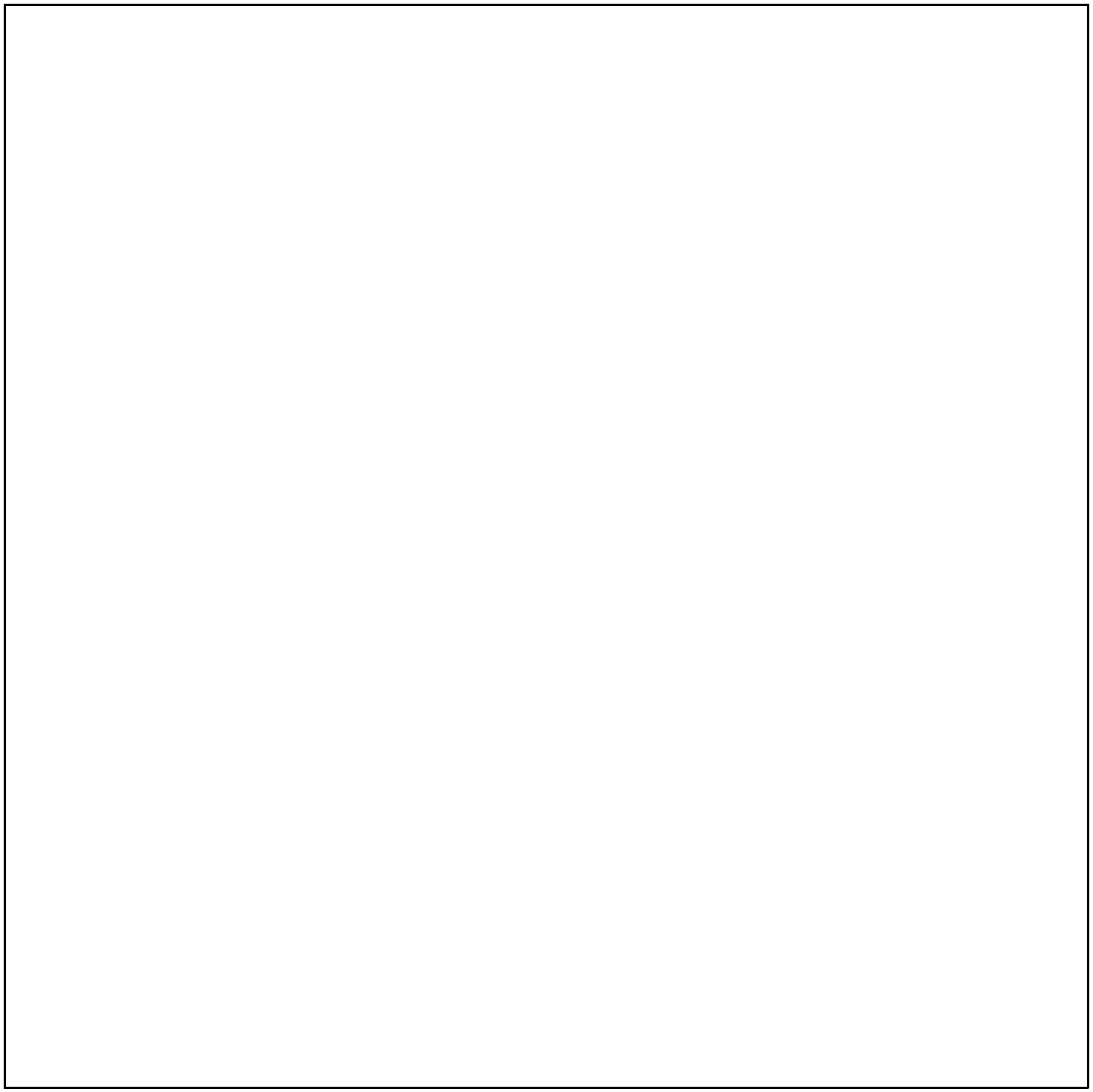
If the two players each adheres with probability one, they will each receive three units. But suppose that each player independently chooses to adhere with probability 5/6 In that case, there is a 25 in 36 chance that they will achieve mutual adherence, a one in 36 chance that they will achieve mutual violation a ten in 36 chance that One will adhere while Two violates and a ten in 36 chance that Two will adhere while One violates. So the *expected* payoffs for both One and Two in this case is (25/36 x 3)+(1/36 x 1)+(10/36 x 7)+(10/36 x 0) which works out to slightly more than four units apiece. So, while it seems irrational to choose a strategy that assures mutual violation when both parties prefer mutual adherence, it seems equally irrational to choose a strategy that assures mutual adherence when there is an outcome that gives both parties an even greater expected payoff. Over the long haul, if many games of this sort of structure were played, each of the players would do better by violating with probability one sixth than by adhering with probability one. And this claim does not have anything to do with recompense, retribution, establishing reputations, or other considerations that enter into discussion of repeated prisoner dilemmas[5]. It holds even if the games were played on each occasion with new opponents who had no access to the results of previous games.

Call a prisoner's dilemma for which certain cooperation by both parities is a pareto-

optimal outcome a *pure* prisoner's dilemma. My colleague Serge Moresi and I have shown[6] that a prisoner's dilemma is pure if and only if it meets the condition:

$$P) \quad \frac{T_2-R_2}{R_1-S_1} < \frac{R_2-S_2}{T_1-R_1}$$

where, for i=1,2, P_i and R_i are the "punishment" and "reward" payoffs that i gets for mutual adherence and mutual violation and T_i and S_i are the "temptation" and "sucker" payoffs that i gets for violating when the other adheres and adhering when the other violates. Applying condition ***P*** to the game described in the first payoff matrix above, for example, yields 2/3<3/2 so that game is pure. Applying it to the game of the second matrix yields 4/3<3/4 so that game is *impure*. The content of condition ***P*** is revealed in the figure below.



One's utility is represented on the horizontal axis and Two's utility on the vertical axis. The four large dots represent the outcomes that are feasible when mixed strategies are not allowed. If the

point (R_1,R_2) lay outside the triangle formed by the other three points (say at the x), the game is pure. In this case, if mixing is allowed, the feasible outcomes are all the points on or within the convex quadrilateral bounded by the four points. Since (R_1,R_2) would be the northeast corner of this region, it would represent a pareto optimal outcome. Here, however, (R_1,R_2) lies strictly inside the triangle formed by the other three points. If mixing is allowed now, the feasible points occupy a more complicated region, bounded on the northeast by a curve of three distinct parts. (R_1,R_2) lies in the interior of this region, so that there are feasible outcomes northeast of it, i.e., feasible outcomes both parties prefer to mutual adherence. Keep in mind that the outcomes in this region represent *independent* mixed strategies. If the players could correlate their moves, so that they could achieve (T_1,S_2) with probability p and (S_1,T_2) with probability 1-p, they could do even better--they could reach the outer dotted line. But the point here is that, even without the possibility of correlation, certain mutual adherence is not a pareto optimal outcome.

The existence of impure prisoner's dilemmas raises a problem for the thesis that morally correct behavior for parties to an agreement coincides with the mutually advantageous behavior. Consider an impure dilemma like the one shown in the second payoff matrix above. Purely prudent contractors will achieve only one unit apiece. Completely faithful contractors will achieve three units. Contractors who are only 83% faithful, on the other hand, have an expectation of over four units. Yet nothing in moral experience seems to ask for randomized behavior. Moral rules require actions of a particular kind, not particular probabilities of such actions. The thought of someone conscientiously consulting a table of random digits to determine whether he adheres to his agreement or violates it does not warm us with Hume's "pleasing sentiment of approbation."

Possible reactions to the problem of impure dilemmas can be divided into two categories: those that champion advantage and those that champion fidelity. A response in the former category might maintain that the recent critics of saintliness, or the recent defenders of virtue ethics over rule-based ethics, have recognized intuitively a principle for which game theory provides justification: that we should not value constant-and-inevitable adherence to rules as highly as almost constant-and-inevitable adherence. But the alleged allies here are unlikely to join the fight. Critics of saintliness and advocates of virtue theory typically see themselves as defending common sense and, when it comes to agreement-keeping, advocating advantage would be radically revisionary. We often *do* value strong fidelity in ourselves and others. It is conceivable that this is because we don't recognize it for what it is, i.e., we don't realize that we would often all be better off with probable adherence than with certain adherence. Once the philosopher or game theorist demonstrates this, the argument might go, we would cease to value detrimentally strict fidelity. A response of this kind is possible, but not probable.

A response in the latter category concedes a gap between moral behavior and advantage. But it might still preserve the idea that agreement-keeping norms are means of securing mutual advantage: our *doing* what morality requires sometimes results in a pareto suboptimal state; our *trying* to do what morality requires does not. We return to this alternative at the paper's end.

III. The problem of new information

Although *M&A* does not recognize the existence of impure dilemmas, it does recognize that there may be times when morality seems to call for strong fidelity. The phenomenon that worries Gauthier is that conditions can change between the time an agreement is made and the

time it is to be kept. Suppose One and Two make an arms control agreement that appears to have the payoff structure in the first matrix above. One then develops a defensive weapons system that would enable it to overpower Two if both should violate. Since one party prefers mutual violation to mutual adherence, the payoff structure is now no longer that of a prisoner's dilemma. Adherence is no longer advantageous in the appropriate sense but, as *M&A* notes, it is still morally required.

The problem is somewhat more general than the example suggests. External conditions need not change, as long as the contractors' information about the conditions does. One and Two might have entered into an agreement thinking the payoffs had a prisoner's dilemma structure only to find out later that it did not. Furthermore the original game need not have been a prisoner's dilemma. Suppose, for example, that One and Two are faced with a choice of whether or not to keep their agreement to go camping next weekend. The situation is depicted in the following matrix.

		2	
		c	n
1	c	(3,3)	(0,0)
	n	(0,0)	(0,0)

A joint camping trip is more valuable to each, than is a solitary camping trip or no camping at all[7]. If One and Two are prudent they each have good reasons (and they collectively have a good

reason) to keep their agreement. But some time before the weekend, One realizes she had a slightly better opportunity (say, a camping trip with someone whose company she slightly prefers to Two's). Now the payoff matrix looks as follows.

		2	
		c	n
1	c	(3,3)	(0,0)
	n	(4,0)	(4,0)

If the players are merely prudent there is no good reason for One (or for One and Two, collectively) to keep the agreement. But surely it is still wrong for One to break the agreement. It appears again that morality requires fidelity even though joint fidelity is not mutually advantageous.

On the other hand, if One's better opportunity were enormously better, we may well feel differently. If One finds out she has won a million dollar lottery prize that must be claimed on the day of the trip we would probably conclude that she is not morally required to keep her agreement.[8] There is a question of *demarcation* here. One would like to know more precisely the conditions under which fidelity is morally required.

The "better opportunity" examples arise when utilities of the contractors increase after the agreement is made. There are parallel examples involving utility decreases. If a small blister

causes me to value the camping trip slightly less than staying home, I should probably still keep my word. If an illness makes the trip painful and dangerous I may surely break it. Again there is a question of demarcation.

A. Expected Values

One solution to problem of changing information (and one answer to the demarcation question) is intimated in *M&A*.

> "It is likely that there are advantages available to trustworthy men that are not available to merely prudent but trustworthy men. For there may be situations in which men can make agreements which each expects to be advantageous to him, provided he can count on the others' adhering to it whether or not their expectation of advantage is realized. But each can count on this only if all have the capacity to adhere to commitments regardless of whether the commitment actually proves advantageous."

The examples requiring explanation are games with two stages--an agreement-making stage with one set of payoffs and an agreement-keeping stage with a new set of payoffs. But the contractors at the initial stage should have considered the possibility that the payoffs might change. Consider the camping trip again. Let us suppose that the probability that the better opportunity would arise for One was 25%. Let us also suppose that there was an equal and independent likelihood that a similar opportunity would arise for Two. Now One and Two have three choices: go camping, don't go camping, and go camping unless the better opportunity arises. The second choice never produces a positive payoff, so both parties may safely ignore it. The matrix now looks like this.

		2	
		f	p
1	f	(3,3)	(2.25,3.25)
	p	(3.25,2.25)	(2.6875,2.6875)

Here f is the "faithful" alternative of going camping no matter what and p is the "prudent" alternative of camping unless the better opportunity arises. If One alone is faithful, there is a 25% chance Two's opportunity will arise and One will camp alone and a 75% chance Two's opportunity will not arise and One will go camping with Two. So the payoff to One is (.75 x 3) + (.25 x 0), or 2.25. For Two, on the other hand there is a 25% chance of an opportunity worth 4 units and a 75% chance of the joint camping trip. If Two alone is prudent, their situations are reversed. The situation in which both are prudent requires a slightly more extended analysis. There is a 1 in 16 chance that both will get better opportunities, a 3 in 16 chance that only One will get a better opportunity, a 3 in 16 chance that only Two will, and a 9 in 16 chance that neither will. So the expected value of this situation to One is (1/16 x 4)+(3/16 x 4)+(3/16 x 0)+(9/16 x 3), or 2.6875. Since the matrix is symmetric, its value to Two is the same.

Notice that the new version of the camping example, unlike the original, does have a prisoner's dilemma structure. It is "advantageous" to everyone if everyone is faithful in the sense that mutual fidelity provides a higher *expected* value for both players than mutual prudence does. In a particular case one player or the other might actually benefit from mutual prudence. But over the long haul, if many games with this sort of structure were played, each of the players would do

better if they were all faithful. (Again, this has nothing to do with retribution or other considerations that enter into discussions repeated games, where players in each round have access to the results of previous rounds.)

The expected value solution to the problem of new information is to note that fidelity really does contribute to *expected* advantage, provided the contractors have accurately assessed the probabilities of new information and made an agreement in which the *expected* payoffs have a (pure) prisoner's dilemma structure. In such cases, we should keep our agreements regardless of how the payoff structure looks at fulfillment time. This suggests an answer to the demarcation question as well: an agreement must be kept if and only if there was an expectation of mutual advantage at the time it was made.

This answer fits nicely with the intuition that agreements may be broken if payoff changes are very large, but not if they are trivial. In certain other respects, however, it is counter-intuitive. Suppose that the improbable slightly-better opportunity arose for *both* One and Two? Must they still keep their agreement? The normal answer, I think, is that agreements may be broken if both parties give their consent. A reasonable explanation for this common idea is that mutual consent is usually a convenient, reliable and public way of establishing mutual advantage. Human nature being what it is, we cannot trust One to make an accurate assessment of the value to Two of a broken agreement. When we would benefit from a broken agreement ourselves we may be excessively disposed to believe that other parties would also benefit. Consequently, before we break the agreement we are required to obtain permission from the other party. The idea that mutual advantage, rather than mutual consent, is what really sanctions agreement-breaking gets some support from examples where the two diverge. If the jilted Two consents to One's breaking

the camping agreement to pursue something marginally more valuable, we are still likely to condemn One for breaking her agreement. Two doesn't really want the agreement broken. His consent doesn't reflect his true preferences and One is a cad to take his consent as license to welch. At any rate, I think it is reasonable to suppose that agreements can be broken when it is suitably established that both parties benefit thereby.[9] Some people (perhaps especially those who subscribe to deontological views of ethics) might have different intuitions when the agreement is especially solemn. Is it morally permissible for a couple to break marriage vows when both would slightly gain thereby (supposing, of course that no children or other third parties are involved and that the couple's action will not encourage divorce among couples for whom it would not be mutually advantageous)? Opinion may be divided here. Nevertheless, the general idea that agreements may be broken when both sides prefer seems reasonable. Yet the expected value demarcation requires that we stick to those agreements whose *expected* payoffs were advantageous even if their actual payoffs are mutually disadvantageous.

The defender of the expected value answer can point out that in these cases the original agreements themselves were not optimally advantageous. In addition to the options of "unconditional camping" and "camping unless my special opportunity arises," One and Two should have considered the option of "camping unless both our special opportunities arise". The third option would be never be less advantageous, and sometimes more advantageous, than the unconditional option. There are really two demarcation questions: what agreements should be made? and what agreements should be kept? It may be that these questions have similar answers, each having to do with mutual advantage. The questions themselves, however, seem to belong to different realms--one prudential the other moral. Although we may censure those who, by

deception, are able to enter agreements that benefit themselves at others' expense, those who enter disadvantageous agreements because of faulty estimates of probabilities or values or poor understanding of the agreements possible consequences do not generally merit moral condemnation. Those who do so out of a desire to help others may merit moral praise.

Let us call an agreement *clearly prudent* when the expected payoffs have a prisoner's dilemma structure and the "adherence" option is optimal in the sense that no unconsidered option would benefit both parties at least as much in any eventuality and more in some eventualities. If the demarcation question is restricted to clearly prudent agreements, the expected value answer provides a simple and plausible condition for keeping agreements--"always". If One's lottery prize were large enough so that the expected payoff of breaking her agreement was positive, then the agreement was not mutually advantageous in the first place. The conditions under which the expected value answer permits agreements to be broken are conditions under which they should not have been made.

To restrict our attention in this way to clearly prudent agreements is to avoid some serious questions. A moral theory of real value must tell us what to do about agreements that are not clearly prudent, even those that are obviously imprudent. It must answer the full demarcation question. The expected value answer seems to fall short here. One shortcoming, as we have seen, is that it requires couples who have imprudently agreed unconditionally to go camping to honor their agreement even when both prefer not to. Common sense suggests that they be permitted to break it. Another shortcoming is that it treats probabilities and values in a parallel way, while our moral assessments seem to treat them differently. Take the camping example again. The expected value answer reasonably suggests that a very valuable special opportunity

makes infidelity permissible whereas a trivial opportunity does not. But it also suggests that a highly probable special opportunity makes infidelity permissible whereas a highly unlikely one does not. That seems wrong. Suppose One had a very high likelihood that a better camping opportunity would arise. Then of course she should not have made the agreement with Two. But given that she did make it, shouldn't she be required to honor it? ("Surely she should have been aware that this might happen.") Or suppose One has to collect her lottery prize on the day of the trip. Our willingness to excuse her breaking the agreement may increase with the size of the jackpot to be collected, but does it really increase with odds of winning? The expected value answer does not distinguish between agreements that should not have been made and agreements that may be broken. When a special opportunity is highly probable we seem more likely to complain that the agreement should not have been made (but must still be honored). When it is highly valuable we seem more likely to say that the agreement may be broken (and perhaps, if its odds were low, to be understanding towards the contractors for making their imprudent agreement).

A defender of the expected value answer might choose to supplement it with a "recision" argument for disadvantageous or suboptimal agreements. Suppose One makes an agreement with Two that is not to Two's expected advantage (i.e., suppose Two's expectation is greater if neither keeps the agreement than if both do). The reason Two has made the agreement must be that he wrongly estimates certain probabilities or values. But there is probably also some likelihood that One would have made similar mistakes and that the bet would really have been disadvantageous to *her*. Perhaps both sides benefit by allowing some sort of right of "recision" when mistakes of this kind are made. Something similar occurs when children allow each other to "take back"

careless moves in a board game. And something like this is behind mandated "cooling off" periods during which important contracts can be canceled. For this move to be convincing, however, we would have to show that conditions under which it is advantageous to rescind the imprudent agreements are the same as the conditions under which it is morally permissible to violate them. This seems unlikely. For when the advantaged party is absolutely certain of the relevant probabilities and values, a recision clause cannot be mutually advantageous. But violation in such cases still seems permissible. It is permissible for you to violate our camping agreement in order to collect your lottery prize even when you knew exactly the odds of winning and exactly what the prize would mean to you.

B. Reputation and translucence

A second response to the problem of new information might be termed the "reputation solution". If One does not keep her agreements in the face of slightly better opportunities her reputation will suffer. People who learn how lightly she takes agreements will be less likely to make agreements with her. Reputations are far more valuable than camping trips. The matrix illustrating the camping story does not properly reflect that fact. A more realistic matrix might have (-1,0) and (-1,-5) on its bottom row, rather than (4,0) and (4,0). Contrary to initial appearances, fidelity in the face of small utility changes is mutually advantageous. More generally, the reputation solution to the problem of new information is to note that to break agreements when payoffs change in your favor is, contrary to appearances, not usually advantageous. By implication, this also answers the demarcation question. An agreement may be broken if and only if it has ceased to be advantageous even when damage to reputation is

properly considered.

The reputation solution is plausible only if breaking the agreement really does damage One's reputation. In many circumstances that would appear not be so. Suppose for example that One's word is already dirt or that, by artful excuses, she can keep a good reputation unsullied. These circumstances would make infidelity advantageous, but they would surely not make it right.

Gauthier's recent discussions of "translucence"[10] (and Robert Frank's discussions of similar notions[11]) can provide a more sophisticated variation of the reputation solution. We might think that in talking about One's "reputation" we could be talking about commonly held beliefs that bear little relation to One herself. As a matter of fact, however, people to some extent wear their souls on their sleeves. Cues from behavior and physiological responses can enable us to "perceive" a person's character traits. The image may be a little cloudy but it does provide real information. One important trait that we expose in ourselves and perceive in others is what might be called *responsiveness*[12]: the tendency to be faithful in dealing with other responsive people and not so in dealing with unresponsive people. The property of having one's responsiveness perceivable is what Gauthier labels translucence.[13]

For the same reasons that a good reputation is valuable to anybody, a responsive disposition is valuable to a translucent person who interacts with responsive people. If One is translucent her responsiveness will be detected and other responsive people will make and keep mutually advantageous agreements with her. If she is not responsive, they will not. So as long as One is translucent and her compatriots are responsive she has good reason to make herself responsive. The link between fidelity and a responsive disposition is not quite the same as the

link between fidelity and a good reputation. An act of infidelity damages one's reputation under many circumstances (although, as was pointed out, these need not coincide with the circumstances in which it is wrong). A *tendency* to infidelity under certain conditions *defines* non-responsiveness. To conclude that infidelity in the face of small temptations is disadvantageous we need the additional assumption that acts of fidelity and infidelity help to produce in us the corresponding tendencies[14]. This assumption, or something like it, is a renowned component of Aristotle's ethics, and it is certainly not implausible. With it we can make the sort of argument about translucency that was made about reputation. In calculating the payoffs for One's infidelity we must weigh the damage to her responsive disposition. Since she is translucent, such damage will be perceived by others and she will find herself deprived of future opportunities to cooperate.[15] When damage to disposition is included we will again find that examples that appear to show that fidelity in the face of small temptations is not mutually advantageous are misleading.

The translucency solution is immune to many of the examples that trouble the crude reputation solution. While artful people can manipulate their reputations, translucent people cannot hide their characters. Analogous objections, however, are not difficult to find. It is unreasonable to think that all of us have the same degree of translucency. On the translucency solution one would expect that more opaque people should be permitted a more relaxed standard of fidelity. Yet we would scarcely permit somebody to welch on an agreement because he was particularly inscrutable. Similarly, we expect variation in *sensitivity*, the degree to which a person's character is affected by acts of agreement-keeping or agreement-breaking, but we find no corresponding variation in culpability. Again, we expect variation in people's ability to

perceive character traits. Great advantage might be gained by having (and showing) the traits of fidelity-to-the-perceptive and infidelity-to-the-blind. But we find no corresponding moral intuition.

IV. Indirect moral theory

Two difficulties for the thesis that the behavior morality requires of agreement-makers is just behavior that is mutually advantageous have now been identified. If the agreement has the structure of an impure dilemma then the rule for producing the most mutually advantageous behavior would require randomly breaking agreements. Morality does not. If the agreement is one in which payoffs may change, advantage would require it to be broken when one party has higher expected payoff (including damage to reputation and observable disposition) from mutual violation than from mutual adherence. Again, morality does not.

The factors that affect whether fidelity is morally required in real-life dilemmas are varied and complicated. One promises to marry Two, but then discovers he is an alcoholic. Is she obliged to keep the agreement? They both discover he has diabetes. Does the answer change? One (a professional salesman of sports memorabilia) offers to sell Two a baseball card for $12.00 and then finds out that it is worth $1200 on the open market. Is she morally required to go through with the deal?[16] The answers seem to depend on the solemnity of the agreement, the amounts and kinds of information shared or not shared by the parties, the magnitude and direction of utility changes between time of agreement and time of fulfillment, the nature of the activities involved, and probably many other factors. None of our proposed answers to the demarcation question draws the line in exactly the right place.

The most promising approach to these difficulties, I believe, is to adopt a theory of the kind that Richard Brandt has aptly labeled "indirect"[17]. Brandt calls the strain he favors an indirect *optimific* theory. He traces versions of it to Berkeley[18] and Mill[19] and finds R.M. Hare's views[20] closely allied. It "*roughly* holds that any *other-person-involving* act is morally permissible if it would be *best* for the moral motivations of (roughly) all agents to *permit* acts of that type in its circumstances, and that an other-person-involving act is an agent's moral duty if it would be best for the moral motivations of (roughly) all agents to *require* acts of that type in those circumstances"[21] or that "what we ought to do is follow the requirements of an *optimal moral code*--a benefit-maximizing set of *moral motivations*, or conscience."[22] Thus Brandt's indirect theory is, as he now calls it, "conscience utilitarianism". But other indirect theories are possible. In particular I want to examine here an indirect *contractarian* theory to the effect that what we ought to do is to follow rules that rational people would agree to teach to themselves and others, or perhaps to follow the conscience we would have if we were given the moral education that it would be rational for us to jointly provide. I will try to show below that indirect social contract theories prescribe different behavior than standard ones, just as indirect utility theories prescribe different behavior than standard ones.

Indirectness exists at two levels. It may be that to produce one kind of behavior we have to inculcate a conscience or moral motivation that calls for a second kind of behavior. That is level one. But it may also be that to inculcate the proper motivation we have to encourage or promote a third kind of behavior. That is level two. Consider, for example, the question of what we should do when we find money in circumstances that make it difficult to determine the owner. Suppose it turns out that we would all benefit most if people turned in amounts of twenty

dollars or more and kept smaller amounts. To get this behavior, however, it might be that we would have to feel a little guilty in keeping amounts of ten dollars. And to cause people to feel guilt at this point, we might have to teach each other that as little as one dollar should be turned in. At the first level, an indirect theory requires that we follow the dictates of an ideal conscience. At the second level, it requires that we follow the dictates of an ideal moral education. Brandt's formulation above may suggest that he is concerned with level one. But he states that when choosing an *optimal* moral code he intends us to consider the costs of teaching it, which suggests that his real concern is level two. He is apparently not asking us to follow that *internal* code or conscience which, however acquired, would yield optimal behavior, but rather to follow the code that it would be optimal to teach, taking into account both direct and indirect effects of the teaching. Whatever Brandt's intention, my own concern will be with level two indirectness.

The greatest difficulty for indirect theories, I believe, is to spell out what one should do when the teachings and motivations in one's society diverge from the ideal that the theory calls for. But even without a faith that this difficulty can be met, one can accept a *part* of the indirect theory: that *if* certain teachings and motivations of one's society do approach the ideal of that theory, then one should try to follow them. I want to suggest that this is the case for our ethics of agreement-keeping.

Another difficulty for *contractarian* indirect theories, is to specify the principles of rationality that would allow us to identify the appropriate rules and teachings. If it turns out that the only education it is rational for us to agree to provide is one whose provision produces more total utility than any other, then the indirect contractarian and the indirect optimific theories

would be the same. But even if one believes otherwise (as I do), one can still accept a part of the optimific theory. If one moral curriculum is pareto superior to another in the sense that teaching the first would leave none worse off and some better off, then it would not be rational to choose the second over the first. In what follows I will try to rely only on this weak principle of rationality.

A great virtue of an indirect contractarian theory, I believe, is that it clarifies and explains the relation between moral theory and our pre-theoretical beliefs and attitudes. The standard accounts of this relation are unsatisfactory. On one view some or all of our pre-theoretic beliefs and attitudes reflect "intuitions" apprehended by a special faculty, which moral theory must try to systematize. On another view beliefs and attitudes about particular acts and objects and more general theoretical views are supposed to cause each other to oscillate with diminishing amplitudes until we achieve a reflective equilibrium in which particular beliefs and attitudes and general theory are harmonized. The first view owes us an account of how we distinguish intuitions from prejudices and an explanation of the nature and operation of the special faculty. The second owes us answers to questions like: why should the magnitudes of belief change diminish at every stage? why must they reach equilibrium? why must such an equilibrium be unique? why should the equilibrium correctly determine what is right and wrong, good and bad, virtuous and vicious? An indirect contractarian view explains both why our pre-theoretical beliefs and attitudes ordinarily carry weight and why they should sometimes be discounted. Over the years we have gotten pretty good at figuring out how to inculcate ideas and attitudes that will further our joint and individual aims. Some of these, like those involving agreement-keeping, are quite subtle and hard to summarize. Not all of our moral beliefs and attitudes were

formed by a rational education, however. If any is known not to be, the mere fact that we happen to have it should not affect our choice of moral theory.

This idea can be made clearer by considering a short fable. There is a species of individuals that are basically rational in the sense that they are usually capable of finding and doing what will best achieve their ends (within the limits of their powers of discrimination and calculation). Their characteristics are to some extent "hard-wired" and to some extent "programmable." For example, each has a strong natural tendency to promote his own survival and well-being, a weaker natural tendency to promote the survival and well-being of other members of his species (especially close relatives), and even, perhaps, some natural tendency to make and keep mutually and optimally advantageous agreements with others. There are good evolutionary explanations of why they have the natural tendencies that they do. But by some sort of "education" or "conditioning" the members of this species are able to strengthen or diminish these natural tendencies in each other and to create other tendencies, some of them highly context-specific. For example, they are able to enhance the tendency to keep agreements in some conditions and reduce it in others. Programming requires a certain consistency. If individuals are subjected to conflicting "educations" their dispositions will not be altered at all. Best results are achieved if everyone with whom an individual interacts is trying to inculcate the same dispositions. Different tendencies are secured by different mechanisms. The tendency to eat sufficient food to maintain health is promoted by a particular kind of discomfort that is felt when more food is needed and a particular kind of satisfaction felt when it is obtained. The tendency to maintain healthy body temperature is promoted by different sorts of discomforts and

satisfactions. Discomforts and satisfactions associated with programmed dispositions feel somewhat different than those associated with hard-wired ones. The creatures have come to call the former "moral" sentiments. They are aware that jointly they can influence which circumstances will and will not trigger moral sentiments among their fellows. Indeed they realize that they can affect the strength of the sentiments that are will be evoked by circumstances of a particular kind. They campaign for various "curricula" of moral education that would alter the conditions that trigger the sentiments and thereby change their fellows' dispositions. These curricula take the form of pronouncements that certain kinds of action are "right" or "wrong" or "supererogatory" or "very wrong" and certain characteristics are "good" or "bad" or "morally excellent" or "morally terrible."

I want to argue that the beliefs and attitudes we should take seriously as moral data are those that reflect the curriculum that a species like this would adopt. It seems plausible, moreover, that our beliefs and attitudes about agreement-keeping fit this category. One might think that this view will lead to the same results as more standard contractarian views. Given its strong concern for its own survival and well-being each individual would like to campaign for the curriculum that benefits itself the most. Since it knows that the others know that their interests don't coincide with its own, each realizes that this curriculum has no chance of being adopted. If agreement is possible at all, it must be agreement on a mutually advantageous curriculum. If the creatures are rational the agreement must at least be pareto optimal. But there is no reason to expect an ideal curriculum to be a curriculum that calls for ideal behavior. In fact, there are several reasons to expect the contrary.

One important reason is that ideal behavior is often achieved already by the hard-wired

dispositions. Moral education has costs. There no point in bearing these costs to inculcate a disposition already universal. J. Urmson, in a famous essay on Mill, argues that Mill could not be what we might now call a simple hedonistic act utilitarian: His tone is indignantly dismissive:

> "... on this view a man who, *ceteris paribus*, chooses the inferior of two musical comedies for an evening's entertainment, has done a moral wrong, and this is preposterous. If this were in fact the view of Mill, he would indeed be fit for little more than the halting eristic of philosophical infants."

Urmson is right in the sense that there is no need for a curriculum that censures unfortunate choices of musical comedies. We already have the necessary disposition. The act utilitarian, on the other hand is also right in a sense: one really shouldn't go to the inferior of two musical comedies (though it would not be evil to do so). But if his theory is supposed to be about *morality*, then Urmson has found a defect. An unconvincing utilitarian reply to this sort of objection has been to argue that we are confusing the "praiseworthiness" or "blameworthiness" of actions and dispositions with their moral worth.[23] A more plausible response has been to admit that they are *not* talking about moral worth in the ordinary sense at all.[24]

A second reason that an ideal curriculum may not call for ideal practice is that an individual's capacity to be programmed is limited. Suppose that, among the creatures in our fable, absolute, unconditional fidelity were optimally advantageous. But suppose also that no amount of moral education would cause very many of these individuals to remain faithful in the face of temptations as great as T. A curriculum that required fidelity in the face of T-sized temptations would cause a great deal of guilt and discomfort to no good end.[25] An ideal curriculum would be one that called for less than ideal behavior. Our weakness as moral pupils necessitates a lenient

curriculum.[26]

Our weakness as pupils also necessitates a stricter curriculum. Suppose the creatures of the fable had a tendency towards backsliding. To make them keep agreements in the face of temptations up to size T we may need to endorse a curriculum that calls for fidelity in the face of temptations up to some larger size T'. To get the ideal effect, we might have to ask for something more than ideal.

I don't propose to try here to work out a detailed account of how considerations like these explain all of our complex and subtle intuitions about the morality of agreement-keeping and agreement-breaking. But I think that they are suggestive and they help us to understand the cases in which morality seems to call for strong fidelity.

Consider first the case of impure dilemmas. Suppose that the creatures in our fable knew that they would play each other in various impure prisoner's dilemma games. What strategies might their ideal curriculum endorse? One option would call for them to estimate the mixed strategy that would result in optimal advantage and try to follow it. This option has drawbacks. The creatures of the fable find it difficult to calculate the correct strategy for a given imperfect game and, without access to randomizing devices, they find it difficult to follow a particular mixed strategy. They find it even more difficult to cause appropriate negative feelings to be triggered by, say, adhering with probability 98% instead of the required 96%. Indeed it is impossible for a creature to know whether an act of adherence (by itself or another creature) is an instance of 98% adherence or 96% adherence. Furthermore, as we argued in the case of agreement-keeping, a certain weakness in the creatures' educational potential can cause an ideal curriculum (supposing it *could* be followed) to have less than ideal results. A more workable

option would be to teach that choosing violation is always wrong, but that the *degree* to which it is wrong (and the strength of the psychological penalties and rewards that should be inculcated) depends on the payoffs. Knowing that the odds of violation decrease as the psychological penalties increase, the creatures could arrange things so that the ideal probability of violation is attained.

Now consider the case of new information. Why should morality distinguish so sharply between high-value temptations and high-probability temptations? Because it would be extremely difficult for potential contractors to imagine and accurately evaluate all the remote possibilities of huge temptations that could arise. We don't want this difficulty to prevent their making mutually beneficial agreements. The solution is to arrange morality so that all agreements become tacitly conditional. Contractors must consider all of the reasonably probable outcomes and remain faithful if any of them should arise. If a remote-but-huge temptation arises, however, violation is permitted. We would all be better off if we made and kept all and only the agreements called for in the expected-value solution. This might require making agreements that are explicitly conditional: "We will go camping unless you win the lottery or I win the lottery, or you get pneumonia or I get pneumonia, or ...". A curriculum that calls for this kind of ideal, impractical behavior, however, is counterproductive.

Why does the translucency solution fail? Because moral education is not needed to make me keep agreements whose violation would hurt me. I am naturally disposed to do so. Small wonder we don't acknowledge a looser standard of fidelity for inscrutable individuals, insensitive individuals and those who deal exclusively with the imperceptive. These are exactly the people

whose natural tendency toward infidelity would be strongest. The difficulty will be to induce them to keep ideal agreements, not to induce them to break inferior ones. Conversely, although we might well benefit if the transparent, the sensitive and the partners-of-the-perceptive maintain slightly higher standards of fidelity, there is no need to teach them to do so.

The loose talk of curriculum and practice should not be construed to mean that I am advocating some form of hypocrisy, whereby an individual is supposed to teach one thing to others while he tries to do something else himself. One wants a curriculum that will produce optimally advantageous behavior. If I am a backslider of the sort described above, for example, then I should endorse the strict standard for myself as well as others just as, if I am habitually five minutes late, I should set my watch five minutes early. Sidgwick is one philosopher who recognized that what it would be best for us to teach may diverge from what it would best for us to do.[27] But morality for Sidgwick had more to do with the latter than the former: his utilitarianism was direct rather than indirect. Consequently Sidgwick was forced to admit that persons with "exceptional qualities of intellect, temperament or character" could permissibly break the rules that others must live by, so long as they did so secretly.[28] Such departure from accepted principles of universality and publicity[29] make his theory less plausible.

There is admittedly something odd about indirect theories of morality. The desired results are obtained if what they prescribe is internalized, taught, endorsed, accepted, believed, acted-on, and so on, but (in some cases at least) not if what they prescribe is actually done. If moral philosophy is a practical subject, then surely its primary concern should be with the former kind of thing (i.e., with that which brings about the desired states). As "lovers of wisdom," however, we may want to examine the latter (that which actually comprises the desired states). The

information provided by such purely theoretical investigations would help in the design and evaluation of practical curricula. But there is also a danger. We must be careful that the beliefs inculcated during a proper moral education are not subverted by the knowledge of what this education is designed to bring about. If I become keenly aware of the fact that my watch is five minutes fast, setting it ahead may cease to have the desired effect. And it is important to keep in mind that our pre-theoretical beliefs and attitudes about what is required or permitted *by morality* can (if appropriately formed) be taken as evidence for the practical subject, but not for the theoretical one.

1.I have benefitted from the comments of Luc Bovens, Richard Brandt, Alisa Carse, David Copp, Wayne Davis, David DeGrazia, Ted Lockhardt, Serge Moresi, and Henry Richardson.

2. *Philosophical Review* volume 76 number 4 (October 1967), pp. 460-75.

3. As *M&A* notes, a second respect in which morality goes beyond these requirements is fairness. In Gauthier's more recent writings (*Morals by Agreement*, Oxford: Clarendon Press, 1986, and *Moral Dealing: Contract, Ethics and Reason*, Ithaca: Cornell University Press, 1990) fairness gets more attention than fidelity. The emphasis here will be on fidelity.

4.Perhaps people are *trustworthy* if they keep exactly the agreements that should be kept, or in Aristotelian terms, if they achieve the mean in fidelity. In replacing the notion of trustworthiness by that of fidelity I do not mean to be *narrowing* the area of concern. I take it that fidelity, like trustworthiness, applies to implicit agreements as well as explicit ones. This encompasses a broad and central area of morality on any view. On a social contract view it encompasses all of morality.

5.See, for example, R. Axelrod, *The Evolution of Cooperation*, New York: Basic Books, 1988.

6.Kuhn and Moresi, “Pure and Utilitarian Prisoner's Dilemmas,” *Georgetown University Working Papers*, 1994, Department of Economics, Georgetown University, Washington, D.C. A condensed version of that paper, without the proofs, is forthcoming in *Economics and Philosophy* under the same title. Some of the discussion in those papers and this one is anticipated in J Howard Sobel, "Constrained Maximization", *Canadian Journal of Philosophy* Volume 21, Number 1 (March 1991), pp 25-32. Sobel notes that a necessary condition for what is here called purity is that $R_i \geq \frac{1}{4}(S_i+T_i+R_i+P_i)$, for i=1 or 2 and that $R_i \geq \frac{1}{2}(S_i+T_i)$, for i=1 or 2, is necessary for there to be no mixed *correlated* strategy that dominates mutual cooperation. (The last concept is the one that most interests him.) It is shown in Kuhn and Moresi, *op cit*, that condition ***P*** is necessary and sufficient both for purity and for the non-dominance of mutual cooperation by mixed joint strategies.

7.I am assuming that neither party is hurt by the mere fact of being stiffed by the other. Perhaps the pride in demonstrating greater fidelity just compensates the indignity of being wronged and the disappointment of unfulfilled expectations.

8.That is, unless we are willing to consider the possibility that One's has a moral duty to keep the agreement that is *outweighed* in this case by non-moral considerations. Not many have been willing to adopt this line (but, see Robert Louden, “Can We Be Too Moral?” *Ethics* vol 98 number 2 (January 1988), pp. 361-378, for a critical discussion of several who have been.)

9. It seems reasonable, at least, if benefit is tied to *preference*. David DeGrazia has forcefully reminded me that under more "objective" construals of benefit breaking an agreement for mutual benefit can be regarded as excessively paternalistic and insufficiently sensitive to the other party's autonomy. It is also worth noting that the process of seeking and granting consent serves to inform the parties that the agreement is being broken, knowledge which may itself be of benefit. I want to

maintain that considerations of advantage may explain when consent is required to break an agreement in a particular case, and that in general consent alone is neither necessary nor sufficient for breaking agreements.

10. Chapter VI of *Morals by Agreement, op cit.*

11. "If Homo Economicus could choose his own utility function, would he want one with a conscience?" *American Economic Review* volume 77 number 4 (September 1987), pp 593-604, and *Passions within Reason*, New York: W.W. Norton, 1988.

12.Gauthier uses the term *cooperativeness*.

13.Gauthier does not offer translucence as a solution to the demarcation problem, but as a way of demonstrating to Hobbes' fool that it is rational for an individual to be moral (or at least that it is rational for him to try to attain a responsive disposition). Unlike some of Gauthier's critics--for example, John Harsanyi, in his review in *Economics and Philosophy* volume 3 number 2 (October 1987) pp 339-373--I find the assumption that what goes on inside us is partially perceivable by others quite a reasonable one to make. Excellent poker players are rare human beings. And Gauthier and Frank give persuasive arguments that translucence has survival value. It seems unlikely that Gauthier's responsiveness is the only or even the main trait whose visibility is important. In the present context similar conclusions could perhaps be derived from the assumption that we can estimate a person's tendency to be unfaithful in the face of temptations of particular sizes. (See note 15.) I find the notion that translucence is what makes morality rational less plausible. To the extent that people are transparent, morality is really not needed at all. Self-interest will dispose us to be responsive. Morality is most needed among more opaque people whom self-interest disposes to be unresponsive.

14. There is a prior assumption that it is *possible* for someone with a certain responsive or unresponsive disposition to choose whether to be faithful. An entirely responsive disposition would presumably require certain fidelity. For such a person it is pointless to talk about what is morally required. We are assuming that it is possible for people to act in ways that are unexpected given their character traits and that it is possible for them to change their character traits. The argument also requires that behavior be predictable from character traits, however, so these possibilities must be improbable.

15.In Gauthier's version of the argument this is because responsive people, by definition, will not cooperate with her. A stronger argument can be obtained by observing that people are often *in competition* to be partners in agreements. Two may not particularly care who is camping partner is, in which case he is likely to make the agreement with the most faithful partner he can find. This version of the argument would seem to make fidelity even more valuable, and therefore lead to a higher moral standard.

16. Some lawyers apparently think not. According to the *Washington Post* (April 4, 1991, page A4), when a salesman in Addison, Illinois sold 13 year old Bryan Wrzesinski a 1968 Nolan Ryan rookie card for $12.00, the owner of the store sued Wrzesinski for $1188.00.

17.Richard Brandt, "Fairness to indirect optimific theories in ethics" *Ethics* volume 98 number 2 (January 1988), pp 341-60. Reprinted in Brandt (ed), *Morality, Utilitarianism and Rights*, New York: Cambridge University Press 1992.

18.George Berkeley, *Passive Obedience* (1712), reprinted in M.W. Calkins (ed.), *Berkeley*, New York: Charles Scribner, 1929, esp. pp 432-440.

19.John Stuart Mill, *Utilitarianism*, chapter 5.

20.R.M. Hare, *Moral Reasoning*, Oxford: Clarendon Press, 1981, esp. pp 35-52.

21.Brandt, op cit. p 342

22.ibid p.347

23.See Sidgwick, *Methods of Ethics*, Book IV, Chapter III, §2. It is unconvincing because going to the wrong musical comedy *isn't* merely not worthy of blame, it is not morally wrong.

24. Consider the following passage from J.J.C. Smart's "Extreme and Restricted Utilitarianism", *Philosophical Quarterly*, vol 6 number 2 (October 1956) pp 344-354:

> "The restricted utilitarian might say that he is talking only of *morality*, not of such things as rules of the road. I am not sure how far this objection, if valid, would affect my argument, but in any case I would reply that as a philosopher I conceive of ethics as the study of how it would be *most rational* to act. If my opponent wishes to restrict the word 'morality' to a narrower use he can have the word. The fundamental question is the question of rationality of action *in general*."

It will be argued below that the questions like the one the utilitarian may be addressing, though of some interest, are less important than the question of what is morally right.

25.The creatures of our fable start with just their hard-wired dispositions. The concern raised here is that these native dispositions might interfere so as to make the results of the proposed curriculum suboptimal. In a more realistic story, the creatures might start with flawed programmed dispositions as well as the native ones. These could similarly interfere with a proposed new curriculum. Perhaps the creatures could then be appropriately reprogrammed or perhaps the new moral education could be provided only for the next generation. But it could also turn out that it was again more advantageous to make the proposed new curriculum more lenient.

26.Even if the capacity to be programmed were there, the costs of the programming *and the costs of being programmed* might still favor the lax curriculum. We might all be better off behaving as Puritans, but if it requires constant contemplation of Duty or chronic guilt, it's not worth it.

27.Sidgwick, *op cit*, Book IV, Chapter 5, §3.

28.One wonders how sincerely this view was held. Sidgwick noted that one consequence of it was that "the opinion that secrecy may render an action right which would not otherwise be so should itself be kept comparatively secret; and similarly it seems expedient that the doctrine that esoteric morality is expedient should itself be kept esoteric." And yet Sidgwick published this conclusion in *Methods of Ethics*, a work which he kept in print over six editions from 1874 through 1901.

29.See, for example, John Rawls, *A Theory of Justice*, Cambridge: Harvard University Press, 1971 pp 132-133.