MORALITY AND RATIONALITY

by

KATHARINE BROWNE

B.A. (Hons.), University of Toronto, 2003

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

(Philosophy)

THE UNIVERSITY OF BRITISH COLUMBIA

April 2005

ABSTRACT

Hobbes's Foole and Hume's knave raise the fundamental question of why one should be moral. They ask why rational persons should adhere to the rules of morality (or justice) when doing so is not to their immediate benefit. If rationality is understood as the pursuit of one's interest, and morality requires that one constrain that pursuit, then why should rational persons be moral? This essay looks at the attempts of Hobbes, Hume and Gauthier to solve the problem raised by the Foole and the knave.

In Chapter 1, I look at Hobbes's answer to the Foole's objection. Hobbes argues that the constraints characteristic of morality are to our advantage, and that one who violates the rules of morality is liable to forgo the benefits of social cooperation. I argue that Hobbes fails to show that this risk outweighs the benefits of selective violations, and thus that he has no answer to the Foole.

In Chapter 2, I examine Hume's reply to the sensible knave. Hume appeals both to prudence, as Hobbes does, and to moral sentiments (i.e., feelings of guilt generated by unjust acts) in his reply. I argue that neither of these appeals is able to support strict and inflexible adherence to justice, and that he is therefore unable to effectively reply to the knave's objection.

Chapter 3 looks at Gauthier's solution. Gauthier argues that morality and rationality can be reconciled by appealing to dispositions rather than directly to actions. He seeks to show that it is rational, because it is advantageous, to cultivate a disposition of "constrained maximization" (i.e., of keeping one's agreements even when violating them would be in

one's interest), and that non-maximizing actions which flow from that disposition are rational. I argue that Gauthier fails to show that constrained maximization is more utility maximizing than the actions recommended by the Foole and the knave and, thus, that he fails to show that it is rational to be a constrained maximizer.

I conclude that Hobbes, Hume and Gauthier are all unable to reconcile morality and rationality, where rationality is understood as a pursuit of one's self-interest.

TABLE OF CONTENTS

ACKNOWLEDGEMENT

For helpful comments and criticisms, encouragement and advice,

I am most grateful to Dr. Paul Russell and Dr. Andrew Irvine.

INTRODUCTION

David Gauthier has claimed that "the reconciliation of morality and rationality is the central problem of modern moral philosophy."[1] Critical to the project of reconciling morality and rationality is making clear what is meant by "rationality." There are two primary senses in which "rationality" can be conceived. The first is the "self-centred" sense, where one's reasons for action depend on one's desires, preferences, interests, and so on. The second sense is the (to use Gauthier's term) "universalistic" sense. On this view, rationality requires that other people's interests be taken into account and, hence, that one can have reasons for action which do not depend on one's preferences or desires. This essay will focus on the possible reconciliation between morality and the "self-centred" sense of rationality.

It is obvious that "rationality," when understood in the self-centred sense, can provide one with the requisite motivation for action. If an action suitably serves one's interest, preference or desire, then one will be motivated to perform that action. The problem here is to show how this sense of rationality is consistent with morality. If rationality is understood as the pursuit of one's self-interest, and morality (or justice) dictates that we curb self-interested pursuit, then it is hard to see how rationality can be reconciled with morality. As Philippa Foot says, "justice seems rather to benefit others, and to work to the disadvantage of the just man himself."[2] This is an especially acute problem if justice arises through a convention that is adopted to serve our interests, as is held by both

Hobbes and Hume. If justice is founded on interest, then what is the rationale behind sometimes acting contrary to our interests in order to adhere to justice?

The standard solution to the problem of why we, as rational beings, should be just suggests that this incompatibility between justice and rationality is only apparent. Human beings are not entirely self-sufficient and rely for their well-being on social interaction. Peaceful social interaction is established through justice. Thus, justice serves one's interest by enabling one to partake in social interaction. The rules of justice do require one to curb self-interested pursuit but, as the argument goes, this restraint is necessary to further one's overall interest, so long as others are willing to do the same. As Kurt Baier puts it, "moralities are systems of principles whose acceptance by everyone as overruling the dictates of self-interest is in the interest of everyone alike, though following the rules of morality is not of course identical with following self-interest."[3]

It is not clear, however, that adherence to the rules of justice in all cases will be to one's advantage. Hobbes's Foole and Hume's knave ask why one cannot violate a rule of justice in order to maximize one's own utility when one can do so with impunity. This essay will examine the attempts of Hobbes, Hume and Gauthier to answer the objections raised by the Foole and the knave. I will argue that all attempts to reconcile morality and rationality when understood as the pursuit of one's self-interest fail and, thus, that the Foole and the knave cannot be defeated.

Notes

[1] David Gauthier, "Justice and Natural Endowment: Toward a Critique of Rawls's Ideological Framework," *Moral Dealings: Contract, Ethics, and Reason,* ed. David Gauthier, (Ithaca: Cornell University Press, 1990), p. 150.

[2] Philippa Foot, "Moral Beliefs," *Virtues and Vices,* ed. Philippa Foot, (Berkeley and Los Angeles: University of California Press, 1978), p. 125.

[3] Kurt Baier, *The Moral Point of View: A Rational Basis of Ethics,* Abridged edition, (New York: Random House, Inc., 1965), p. 154.

CHAPTER 1

HOBBES AND THE FOOLE

Hobbes tries to reconcile morality and rationality by arguing that, first appearances to the contrary, it is always to our advantage to be just. This view is challenged by Hobbes's analogue of Hume's knave, the Foole. I will argue that the Foole wins the argument.

1. Origin of Morality and Justice

According to Hobbes, morality can be derived from the assumptions that human beings are (1) motivated by self interest, and (2) rational. A sufficiently rational and self-interested individual, on Hobbes's view, will recognize that the benefits of moral behaviour far exceed any benefits of immoral behaviour, and so we can (to use Gauthier's phase) "bargain our way into morality."[1] The argument goes like this. Left on their own, individuals will pursue their own interests. This will bring them into competition with others. Person A will want to take what person B possesses, and person B will want to take what person A has. Recognizing the possibility of attack from B, A will make a pre-emptive strike against B, and B will reason similarly and behave analogously towards A. Individuals will thus be in a state of war, which in a state of nature (i.e., a state in which there is no overarching civil authority) is the natural condition of mankind.[2]

A state of war, however, is to no one's advantage and everyone can do better if they make agreements of non-aggression. But an agreement between A and B would be meaningless unless A and B each had some reason to think that the other would comply, and

as long as A and B are in a state of nature there will be no such reason. Thus it will be in the interest of individuals to erect a Sovereign—a civil authority with sufficient power to enforce agreements. With such an authority, it becomes rational to make and comply with agreements and, with that, justice and injustice first become possible.[3] The Sovereign thus enables human beings to escape from their natural condition of war, in which life is "solitary, poor, nasty, brutish, and short."[4] Thus the very same thing that drives human beings to war—rational self-interest—points the way to peace, the key to which is keeping one's agreements.[5]

2. The Foole's Objection

At this point the Foole enters. Hobbes introduces him in the following words:

> The foole hath said in his heart there is no such thing as Justice; and sometimes also with his tongue; seriously alleging, that every man's conservation, and contentment, being committed to his own care, there could be no reason, why every man might not do what he thought conduced thereunto: and therefore also to make, or not make; keep, or not keep covenants, was not against reason, when it conduced to one's benefit. He does not therein deny, that there be covenants; and that they are sometimes broken, sometimes kept; and that such breach of them may be called Injustice, and the observance of them Justice: but he questioneth, whether Injustice ... may not sometimes stand with that Reason, which dictateth to every man his own good ...[6]

Hobbes's Foole asks why, if we enter into covenants to satisfy our own interests, can we not break covenants if it is our interest to do so. If reason dictates that one pursue one's own interests, and if breaking a covenant is in one's interest, then it is not incompatible with reason to break a covenant. As the Foole says, "... [breaking covenants] can never be against reason, seeing all the voluntary actions of men tend to the benefit of themselves; and those actions are most reasonable, that conduce most to their ends."[7] Justice should be compatible with reason and, thus, if breaking a covenant is not against reason, performing such an act is likewise not against justice; otherwise "justice is not to be approved for good."[8]

It is important to note that the Foole is contending that both in the state of nature and in civil society, it is sometimes in one's interest (and is therefore reasonable) to break covenants.[9] Covenants in the state of nature are those made in the absence of a commonwealth capable of enforcing the covenants that men make. Covenants in civil society, on the other hand, are those made where there exists such a commonwealth. It is further useful to distinguish here, as Gregory Kavka does, between three different cases of covenant violation. These are (1) first party violations in the state of nature; (2) second party violations in the state of nature, where the first party has already adhered to his part; and (3) violations in civil society.[10] Hobbes's ultimate solution to the problem of non-compliance is a political one: to have a Sovereign with sufficient power of surveillance and authority to punish, so as to make non-compliance unattractive. This provides Hobbes with an account of how to handle violations in civil society; but in his reply to the Foole, Hobbes

speaks to the rationality of compliance in a state of nature. This is puzzling, given what Hobbes has previously argued.

Indeed, the appearance of the Foole is altogether puzzling. Hobbes argues that in a state of nature, there is no justice or injustice, and so no reason to keep covenants.[11] He also argues that in a state of civil society, fear of the Sovereign is sufficient to secure obedience. There thus does not appear to be room for the Foole's objection. There is only room for the Foole's objection if the Sovereign is unable to secure obedience or there is some obligation to keep covenants in a state of nature. It is not clear how this can be reconciled, but let us set this seeming inconsistency aside, and take the Foole's objection in its own terms. It is unrealistic to think that a Sovereign will be able to enforce all covenants, and so reasonable to ask why we should not violate them when we think we can get away with it. As will appear in what follows, Hobbes's claim that it is rational for second parties to perform their part of the covenant when the first party has done the same[12] is interesting in its own right, and fills a gap that would otherwise be left in Hobbes's account. In everyday life we make agreements with others that are unenforceable in law (e.g., agreements with friends to meet at a restaurant at 8 p.m., etc.). If the state of nature is defined as a state in which there is no overarching civil authority, these are state-of-nature agreements, and the question then arises whether it is rational to make and keep them. In his reply to the Foole, Hobbes argues that it is. If he is successful, then Hobbes makes a big gain, for he can then argue (as Gauthier subsequently does[13]) that it is rational to cultivate a disposition to keep covenants. This will reduce reliance on a sovereign, together with the costs attendant to that. Self-interested and

rational people could bargain their way into the moral stance. So let us turn to Hobbes's reply to the Foole.

3. Answering the Foole

Hobbes gives two very different arguments for the rationality of keeping covenants. The first is that it is absurd to break a covenant: "For as it is there [i.e., in the schools] called an absurdity, to contradict what one maintained in the beginning: so in the world, it is called injustice, and injury, voluntarily to undo that, which from the beginning he had voluntarily done."[14] Hobbes thus argues, in a way reminiscent of Kant, that if one voluntarily makes a covenant, then it is contradictory to voluntarily break that covenant; and since reasonable people cannot perform self-contradictory acts, reasonable people must keep their covenants.

Hobbes's first argument regarding the rationality of keeping covenants can be dismissed with relative ease. There is no contradiction in making and then breaking a covenant. It is intelligible to make a covenant if it is initially in one's interest to do so and, likewise, it is intelligible to break that covenant if it is later in one's interest to do so. Since the account that we give of why we make and break covenants is intelligible, it cannot be a contradiction to do so, and we therefore cannot rule out breaking a covenant on the grounds that it is an absurdity.

Hobbes's second argument is more substantial, and it serves as his official answer to the Foole. It runs as follows:

> He … that breaketh his covenant, and consequently declareth that he think he may with reason do so, cannot be received into any society, that unite themselves for peace and defence, but by the error of them that receive him; nor when he is received, be retained in it, without seeing the danger of their error; which errors a man cannot reasonably reckon upon as the means of his security: and therefore if he be left, or cast out of society, he perisheth; and if he live in society, it is by the errors of other men, which he could not foresee, nor reckon upon; and consequently against the reason of his preservation; and so, as all men that contribute both to his destruction, forbear him only out of ignorance of what is good for themselves.[15]

Hobbes is not contending here that first-party violations in a state of nature are improper. He agrees that violating one's covenants may be in one's self-interest when one is the first party to perform, as one cannot be sure that the other will adhere to his part.[16] Kavka refers to this type of violation as a "defensive violation".[17] If I refrain from performing my part of a covenant that I have with my neighbour requiring me not to steal his possessions so long as he does not steal mine, for example, on the grounds that I do not have reasonable assurance that he is willing to adhere to his part, my violation is a defensive violation. If my neighbour is not willing to comply with our agreement, then I am putting myself at a disadvantage by adhering to the covenant while he violates it. Hobbes's first law of nature states that we should first and foremost seek peace, but he also holds that we ought to defend ourselves by every means that we can.[18] What is implied here is that, if others are not willing to adhere to agreements that ensure peaceful cooperation, it is not contrary to reason

to violate such agreements. Violating the agreement that I have with my neighbour is a means of protecting myself from his potential attempts to gain at my expense.

Thus for Hobbes, defensive violations may be prudent in the state of nature. But Hobbes disagrees with the Foole's claim that it is reasonable to violate covenants both in the state of nature when the other party has performed (i.e., second-party state-of-nature violations) and in civil society.[19] In the latter case, someone who violates a covenant (in civil society) runs the risk of punishment and of being shunned by others. In the former case (of second-party violations in the state of nature) one would not then be admitted to civil society. Violations of this sort, where the other party has already performed, are what Kavka describes as "offensive violations".[20]

Hobbes argues that by not adhering to a second-party covenant in the state of nature, one will no longer be trusted to carry through with agreements.[21] Others will not be willing to enter into agreements with one who is known to violate such agreements. Without the ability to make covenants with others, the untrustworthy person will be excluded from society, and will be left in the state of nature, "where every man is enemy to every man,"[22] and where there can be no security of survival. Hobbes sees this as incompatible with one's self interest and, hence, argues that to preserve our interests, reason dictates that we must keep covenants. Hobbes does nevertheless recognize that there is a possibility that one will not be found out for one's violation of a covenant, and that one may consequently be received into society; but he argues that it is not reasonable for one to act on such an anticipated outcome, given the threat of being found out for such violations and being forced

to live in the state of nature.[23] In what follows, I will argue that the Foole cannot be defeated so easily.

It should first be noted, however, that the line between offensive and defensive violations is not always clear. Hobbes's second law of nature states that we ought to lay down our rights when others are willing to do the same.[24] Thus, only when others are not willing to lay down their rights are covenant violations permissible. Such violations are defensive violations. This raises some questions that Hobbes seems to fail to address. How sure do we have to be that others are willing to lay down their rights? Beyond all doubt? Beyond reasonable doubt? On balance of probability? And how many others must be willing to lay down their rights to justify laying down one's own? In an agreement made between ten people, are all ten people required to be willing to lay down their rights for one to lay down his? What if only seven are willing? If the number is small enough, the line between defensive and offensive violations becomes blurred, for if some are not willing to give up their rights, then violating one's agreement can be viewed as a defensive violation. On the other hand, if others are willing to lay down their rights, then one's same violation can be viewed as an offensive violation. Hobbes claims that offensive violations are not permissible while defensive violations are. If the distinction between these two sorts of violation is not always clear, then neither will what is permissible and not permissible always be clear.[25] This is a problem for anyone who puts weight on the distinction between offensive and defensive violations, but it is not a problem that need detain us, for I will argue that offensive as well as defensive violations are sometimes rational.

4. Why Hobbes' Answer Fails

It will be useful in analyzing Hobbes's argument for why we should keep our covenants to highlight his key assumptions. His argument runs as follows:

P1: Life in the state of nature is worse than life in civil society.

P2: One who is found out for one's (second-party state-of-nature or civil-society) violation of a covenant will be left in, or will be returned to the state of nature.

P3: It is not reasonable to select a course of action that might leave one in, or return one to, the state of nature.

C: It is not reasonable to violate one's covenant.

In the evaluation that follows, I will challenge each of these three premises and, thus, the conclusion that it is against reason to violate one's covenant.

P1: Life in the State of Nature is Not So Bad

Hobbes places much emphasis on the benefits of civil society, and on the security one achieves there through cooperative pacts one makes with others. Life in civil society, according to him, is far better than the alternative, namely, life in the state of nature. Disqualification from civil society is thus regarded as a significant cost. That life in the state of nature is worse than life in civil society is, however, questionable.

There is evidence that non-human creatures employ certain conditional cooperative strategies in nature, and that these strategies have evolutionary benefits. Some fish, for example, display the characteristics of using a "tit-for-tat" strategy when they survey for predators.[26] The "tit-for-tat" strategy applies when players interact with one another on more than one occasion (as opposed to "one-shot" games, where there is no further interaction). In repeated games, the "tit-for-tat" strategist will cooperate on the first interaction, and mimic the other player's previous move on all subsequent interactions.[27] The tit-for-tat strategist therefore cooperates when the other player cooperates, and defects when the other player defects. In the case of the fish, when two fish survey for predators, one fish will begin to approach the predator. If the other fish follows, the first fish will continue to approach. If the second fish fails to approach, the first fish will then refuse to proceed. The fish appear to cooperate when the other is willing to cooperate, and defect when the other defects. The tit-for-tat strategy has its benefits. In repeated games, where individuals interact randomly with one another on more than one occasion, Robert Axelrod discovered that individuals using the tit-for-tat strategy tend to do, on average, much better than those who always defect.[28] If such cooperative strategies are beneficial and can evolve in non-human creatures, there is no reason to think that they cannot do the same in humans.

Nor is there reason to think that effective defensive pacts in the state of nature cannot be made. Hobbes, in fact, concedes that such pacts are possible when he claims that it is not rational to violate second-party state-of-nature covenants, but seems to think them to be unstable. Recall that he claims that only those covenants made in the state of nature where neither party has performed (i.e., first-party state-of-nature covenants) are invalid. Second-

party state-of-nature covenants, where the other party has performed, on the other hand, are, according to Hobbes, both rational and obligatory.[29] This seems to suggest that agreements are possible in the state of nature. If so, it is not clear that in the state of nature no one will be willing to make agreements with one another and, hence, that there can be no security of survival.

Thus life in the state of nature need not be as bad as Hobbes claims. We also must not ignore the possibility that life in the civil society is not as good as Hobbes claims. Civil society has its defects. The benefits of civil society come from the rule of law, but the rule of law comes at a cost. Laws curtail one's freedom; the enforcement of laws requires surveillance that compromises privacy; there are sure to be unjust laws and unjust enforcement of laws; those in power are liable to be corrupted by power; and so on.[30] If effective defensive pacts can be arrived at in a state of nature, then these problems associated with civil society can be avoided. Hobbes may be right to say that life in the state of nature is worse than life in civil society, but to show this he must show that it is better to accept the above evils of civil society than to try to make do with state of nature defensive pacts. This he does not do.

P2: Detected Violations Will Not Put One in the State of Nature

Even granting that life in the state of nature is worse than life in civil society, it is not clear that all detected violations will leave on in, or return one to, the state of nature. Hobbes, in answering the Foole, appeals heavily to the undesirable consequences that befall one who violates a covenant. In particular, he argues that others will not enter into future agreements

with one who is known to have violated such agreements in the past. It is true that if our aim is to enter into cooperative agreements only with those who can be trusted to carry through with their part of the agreement, then those who cannot be counted on to adhere to their agreements ought therefore to be disqualified from future agreements. It is not true, however, that one who has violated such agreements in the past cannot be trusted to adhere in the future.

Whether one can or cannot be trusted to adhere to future agreements depends on whether one has a disposition to violate. We cannot, however, immediately infer a disposition to violate from all detected violations. If one's past violation is small (e.g., breaks a minor promise), or if one violates for good reason (e.g., to save a life), shows suitable remorse, or has undergone effective behaviour modification (e.g., punishment), it is not clear that one cannot be trusted to carry through with future agreements and thus whether one should be excluded from future agreements with others.

Even if we can infer from one's past violations an untrustworthy disposition, there are other ways that a known violator may still reliably partake in agreements with others. A past violator might, for example, be able to put up a bond of some sort—money, or something else of value—as a means of assurance that he will adhere to his part of the agreement.[31] The untrustworthy person might put forward a substantial sum of money, for example, to make an agreement with his neighbour. In so doing, the neighbour with whom the untrustworthy person makes the agreement can know that either the agreement will be kept or the untrustworthy person will lose the money he put up. If the sum to be lost is

substantial, it is reasonable to assume that the untrustworthy person will keep the agreement he makes.

In the absence of bonds, there still remains hope for the untrustworthy person. Just as one might be willing to gamble by being dishonest if the potential gains from non-compliance are large enough, so others might gamble on making an agreement with an untrustworthy person if they stand to gain substantially in so doing. An untrustworthy person might be so exceptional in other areas that it is worth taking the risk. An employer might, for example, hire a highly skilled candidate knowing that at previous jobs the candidate has not been a model employee. The skills that the candidate can bring to the workplace might be worth taking the risk and hiring the candidate. Just as an employer might take a risk on an untrustworthy candidate if the potential gain is large enough, so one might make a gamble by making an agreement with a known violator if one stands to gain through such an agreement. It is not clear in either case that one would be foolish to do so. Thus it seems that the untrustworthy person remains as someone with whom one might make agreements.

In his discussion of the Foole, David Gauthier comments that "Hobbes needs to say that it is rational to perform one's covenant even when the performance is not directly to one's benefit, provided it is to one's benefit to be disposed to perform. But this he never says. And as long as the Foole is allowed to relate reason directly to benefit in performance, rather than to benefit in the disposition to perform, he can escape refutation."[32] Appeal to dispositions deserves further exploration, which I will turn to in Chapter 3, but if the above

sketch is right, appeal to dispositions will not enable Hobbes to maintain his strict opposition to offensive violations of covenants.

P3: Violating is Not a Bad Gamble

I have argued thus far that detected violations will not necessarily disqualify one from civil society; but even if they did, it is not clear that all violations are a bad gamble. Consider the following kind of case: I have travelled to a foreign country to which I plan never to return. Suppose I make an agreement with a fellow traveller that I have met on the train: I will watch his luggage while he checks into the hostel, so long as he does the same for me. As he checks in, I take off with his luggage, filled with valuable goods, never to be seen again, and never to be found out. Further, one does not have to leave home to make undetected violations. I may sneak into my neighbour's house and take his things or, perhaps even more invisibly, cheat on my income tax or insurance claims. Are these actions irrational?

Hobbes will argue that they are, for if I were to be found out, the benefits reaped through violation would be outbalanced by the consequences of being found out for such a violation. But Hobbes fails to demonstrate that violation of covenants is a bad gamble. There may be cases where it is so unlikely that one will be found out, or the benefits of violation are so great, that the risk is worth taking.

Hobbes argues that it is never the case that the benefits of violating a (second-party state-of-nature or civil-society) covenant are worth the potential costs of so violating. Since Hobbes is concerned with maximizing one's self-interest (or maximizing one's utility), we can put this formally by employing the expected-utility-maximization theory, which tells us

to act so as to ensure the greatest possible utility.[33] To see how this works, let us turn to the following table:

	Violations are detected		Violations are not detected	
Adhere	b1	(p)	c1	(p – 1)
Violate	c2	(p)	b2	(p – 1)

The expected utility of adherence can be calculated with the following formula: b1p + c1(p-1), where b1 represents the benefits of adherence, c1 the costs of adherence, p the probability that violations will be detected, and p – 1 the probability that violations will not be detected. We can likewise calculate the expected utility of violation: c2p + b2(p – 1), where c2 represents the costs of violation, b2 the benefits of violation, and p and (p-1) the respective probabilities of violations being detected and not detected.

For Hobbes to reach the conclusion that it is never rational to violate one's (second-party state-of-nature or civil-society) covenant, the expected utility of non-compliance must be less than the expected utility of adherence. It should be apparent, then, that the values we assign to the relevant benefits, costs and probabilities is crucial in determining the overall expected utility of each act. It should also be clear that there is a great deal of arbitrariness in how we assign these values. If we are dealing with fixed—or very nearly fixed—costs and benefits, and very small probabilities, the costs of non-compliance must be substantially greater than the benefits of non-compliance if Hobbes's view that non-compliance is always a bad gamble is to stand; but it is not obvious that this is a reasonable assumption to make.

To make this clear, let us turn to Pascal's famous wager, which can be represented in a structurally similar way (with a different payoff scheme) to the above gamble. Pascal's wager weighs the benefits of believing in God against the costs of not believing. We can represent the alternatives in Pascal's wager as follows:

	God Exists	God Does Not Exist
Believe	+ infinite (go to heaven)	- 10
Don't Believe	- infinite (go to hell or forgo heaven)	0

Here it is assumed that if one believes in God and God exists, one will go to heaven, and if one does not believe in God and God exists, one will go to hell (or forgo heaven). Hence we have an infinite payoff if one believes and God exists, and an infinite cost if one does not believe and God exists. On the other hand, if God does not exist, it is assumed there will be no payoff if one does not believe (one does not face any penalty for not believing, nor any benefit in not believing). Also assumed is that there is a small cost in believing when God does not exist (i.e., forgoing any hedonistic pleasures in order to reap non-existent infinite gains). Pascal argues that, given the infinite payoffs and costs, no matter how remote the probability is that God exists, if the probability is greater than zero, it is more reasonable to believe in God.

Let us draw a parallel between Pascal's hell and Hobbes's state of nature (and assume here that life in the state of nature is as bad as Hobbes claims it to be). Pascal's wager works insofar as, given his assumptions about infinite payoffs and costs, believing will always yield the highest utility. Hobbes, however, is not dealing with infinite payoffs

and costs. Whether adherence is better than violation depends on what the values are of the relevant benefits and costs associated with each act. Thus, if the benefits of non-compliance are great, and the probabilities of being caught are fixed and small, then it will not necessarily be the case that the expected utility of adherence is greater than that of non-compliance. It is therefore not clear that it is most rational to adhere to covenants in all cases.

People regularly take risks—even life-threatening risks—when they drive on the highway, board an airplane, give birth to children, ski, golf, and so forth. But it is rarely the case that such acts are viewed as irrational. Are these cases any different from cases of covenant violation? Consider golfing. The probability of being struck dead by a golf ball on the golf course is significantly low. The potential cost of golfing (i.e., death) is thus significantly large, but highly improbable. The benefit of golfing, however, is very high for someone who enjoys golf. The same can be said of certain cases of covenant violation. The probability of being caught violating one's covenant might be significantly low (say equal to being struck dead on the golf course) but, again, if one is caught, the potential cost of violating one's covenant (i.e., life in the state of nature or legal penalty) is significantly large. On the other hand, the benefit of violating one's covenant is very high if one stands to gain substantial rewards by violating. If we take seriously Hobbes's contention that it is never rational to violate a second-party state-of-nature or civil-society covenant, then it seems that it is never rational to play golf, fly, have children, and so on. The principle that we ought not to perform a given action if there exists any possibility—no matter how remote

that might be—of negative consequences is surely absurd, but Hobbes's argument relies on it.

To What Extent is Hobbes Concerned With Morality?

There seems to be a difference between the decision to steal from one's neighbour and the decision to play golf. Whether one should steal from one's neighbour is a moral question, and it is in virtue of this that we often view it differently from the decision to play golf. But if Hobbes is right, then there should be no difference: the question of whether one should play golf should be regarded as similar in kind to the question of whether one should steal. Answering both questions according to Hobbes is simply a matter of calculating the risks, benefits, and costs of the given acts. If morality is to be understood merely as a matter of expected-utility maximization, it seems that one might often be led to act in ways that might intuitively appear wrong.

This is precisely what the Foole has in mind: he claims that there are cases where one can maximize one's expected utility by acting contrary to morality (i.e., by breaking covenants). Hobbes attempts to avoid this awkward consequence by arguing that it is never rational to violate one's second-party or civil-society covenant since the risk is never worth taking. As I have tried to argue, however, Hobbes fails to show that it is *always* in one's interest to adhere to one's covenant. The Foole therefore wins the argument.

From this I draw two conclusions. First, while morality may generally be a good policy, it is not always so, and when it is not advantageous it cannot be reckoned a good by Hobbes or anyone who argues that morality requires rationality, and rationality requires

advantage. Second, on that conception of morality, immorality is not something to be ashamed of. Concepts such as fidelity, veracity, honesty and charity, along with attitudes like guilt, remorse and shame are all swallowed up by a cost-benefit analysis. An accountant who fiddles the books and gets caught is on a moral par with a climber who takes on a difficult pitch and falls. As long as rationality is understood to be acting in pursuit of one's advantage, the prospect of reconciling a traditional sense of morality and rationality looks bleak.

Notes

[1] David Gauthier, "Bargaining Out Way Into Morality: A Do-It-Yourself Primer," *Philosophic Exchange,* vol. 2, no. 5 (1979), pp. 14-27.

[2] Thomas Hobbes, *Leviathan,* (1651), ch. 13, paras. 1-8.

[3] Hobbes, ch. 14, paras. 7, 20.

[4] Hobbes, ch. 13, para. 9.

[5] Hobbes, ch. 13, para. 14.

[6] Hobbes, ch. 15, para. 4.

[7] Hobbes, ch. 15, para. 4.

[8] Hobbes, ch. 15, para. 4.

[9] Gregory Kavka, *Hobbesian Moral and Political Philosophy,* (Princeton, NJ: Princeton University Press, 1986), pp. 137-140.

[10] Kavka, p. 138.

[11] Hobbes, ch. 13, para. 13.

[12] Hobbes, ch. 15, para. 5.

[13] David Gauthier, *Morals By Agreement,* (Oxford: Oxford University Press, 1986).

[14] Hobbes, ch. 14, para. 7.

[15] Hobbes, ch. 15, para. 6.

[16] Hobbes, ch. 15, para. 5.

[17] Kavka, p. 139.

[18] Hobbes, ch. 14, para. 4.

[19] Kavka, p. 139.

[20] Kavka, p. 139.

[21] Hobbes, ch. 15, para. 5.

[22] Hobbes, ch. 13, para. 9.

[23] Kavka, pp. 141-142.

[24] Hobbes, ch. 14, para. 5.

[25] Kavka, pp. 347-8.

[26] Manfred Milinski, "Predator Inspection: Cooperation or 'Safety in Numbers'?", *Animal Behaviour,* vol. 43 (1992), pp. 679-80.

[27] Robert Axelrod, "The Emergence of Cooperation Among Egoists," *American Political Science Review,* vol. 75, no. 2 (1981), p. 324.

[28] Axelrod, pp. 326-331.

[29] Kavka, pp. 351-352.

[30] Kavka, pp. 254-255.

[31] Stephen Darwall, "Kantian Practical Reason Defended," *Ethics,* vol. 96 (October 1985), p. 99.

[32] David Gauthier, *Morals By Agreement*, p. 162.

[33] Michael Resnik, *Choices: An Introduction to Decision Theory,* (Minneapolis: University of Minnesota Press, 1987), pp. 93-100.

CHAPTER 2

HUME AND THE KNAVE

I have argued thus far that Hobbes's minimalist assumptions about moral motivation do not allow him to reply effectively to the Foole. I have tried to show that, given the assumptions (1) that human beings are motivated by self-interest, and (2) that human beings are rational, no compelling argument for being moral can be given. If I am right, then to get an effective reply we will have to go beyond Hobbes, and alter our set of assumptions about human moral motivation. In the following chapter, I will explore the moral theory of Hume, whose "sensible knave" presents a structurally similar problem to that posed by Hobbes's "Foole," although it is embedded in a moral system that has elements that are quite alien to Hobbes's system. Hume rejects the egoist psychology that Hobbes endorses. Instead, he holds that the moral sentiments play a significant role in our motivation to be moral. Where Hobbes makes no assumptions about benevolence and altruism, Hume holds these to be central to morality. In spite of this radical divergence between the two theories of morality, there remains a striking similarity between Hobbes's moral theory and Hume's theory of justice, specifically that they are both founded on interest. Given this similarity, I will focus on Hume's theory of justice. I will argue that Hume's appeal to moral sentiment yields a more robust reply to the knave than does Hobbes's appeal to reason in replying to the Foole, but that it too is ultimately unsuccessful.

1. Hume's Theory of Morals and Justice

Moral Distinctions

In order to understand Hume's theory of justice, we must begin by locating the place of justice in his moral theory in general. According to Hume, moral distinctions are not derived from reason, but rather by moral sentiment.[1] Moral sentiments are feelings of pleasure or pain which constitute our moral approbation or disapprobation. The distinction between virtues and vices, or good and bad, is simply a matter of how we feel. That which causes pleasure is deemed virtuous, while that which causes pain, vicious.[2]

For Hume, our assessments of virtue and vice rest on motives: "'Tis evident, that when we praise any actions, we regard only the motives that produced them, and consider the actions as signs or indications of certain principles in the mind and temper."[3] Motives are indications of character, and character is what determines the quality of actions as virtuous or vicious. As Hume says, "If any action be either virtuous or vicious, 'tis only as a sign of some quality or character."[4] Virtuous actions thus derive their merit from virtuous character, and virtuous character from virtuous motives. Likewise, vicious actions derive their demerit ultimately from vicious motives.

To make this clear, imagine that we have observed an agent perform an action. The object of our moral approval or disapproval with respect to this action will be the agent's character. In making our moral assessment, we look at the effects that the agent's character has brought about; these will be either pleasing or displeasing. If the action affects us directly we will be directly aware of this pleasure or pain. If, on the other hand, the action was directed at someone else, we will become aware of the pleasure or pain of that person through the mechanism of sympathy. The principle of

sympathy grants us awareness of the feelings of others; as Hume says, "by means of this lively notion I am interested in them; take part with them; and feel a sympathetic motion in my breast, conformable to whatever I imagine in his."[5] We imagine ourselves in the position of others and, in effect, share in their pleasure or pain. If the effects of the character are pleasurable, we will have feelings of approbation; if painful, we will have feelings of disapprobation, which constitute our moral distinctions. These moral evaluations are expressed as "virtue" and "vice."

Natural and Artificial Virtues

Hume draws a sharp distinction between the natural and artificial virtues. Natural virtues are those, such as benevolence and humanity, to which we are naturally disposed, and of which we naturally approve.[6] Artificial virtues, on the other hand, are those, such as justice, to which we are disposed and of which we approve only by some artifice.[7] While there are natural inclinations and motives to perform naturally virtuous actions, there is no natural motive in the case of the artificial virtues.

Justice is arguably the chief artificial virtue and for Hume has to do primarily with property and ownership. Being just thus can be understood as being honest and respectful towards property and ownership rights. Honesty in this respect, Hume claims, is something to which we are neither naturally disposed nor naturally approving of. It is only once justice is established as an artificial virtue that we approve of just acts and disapprove of unjust acts. Thus two questions that are central to understanding Hume's theory of justice are (1) How are rules of justice established? and (2) How do rules of justice obtain their moral status? I will consider both of these issues in turn.

The Origin and Moral Status of Justice

The rules of justice, according to Hume, arise artificially and are founded on interest. Where there is limited generosity and scarcity of resources, the rules of justice are required to bestow stability of ownership. Stability of ownership is necessary for social cooperation.[8] We enter into conventional agreements with each other to refrain from violating one another's property.[9] Unlike Hobbesian contracts, however, which are characterized by bargaining and are founded on explicit promise, Hume's covenants are tacit agreements (and agreements are for Hume conventions), where one trusts that the other will cooperate. Hume explicitly denies that promises have a place in agreements relating to justice and morality. He holds that promising itself is based on a convention and thus cannot be the basis of conventions.[10]

A well-known example of convention is Hume's example of rowing.[11] Two men pull the oars of a boat by common convention for common interest, without any promise or contract. Each prefers the outcome if both row to the outcome if neither rows (and the boat does not move at all), or only one rows (and the boat moves in circles). Personal conformity to the convention of rowing is each person's most preferred response to nonconformity. Participation in such Humean conventions is thus conditional on the participation of others.

For Hume, the rules of justice are conventions which, given the scarcity of resources and limited benevolence among men, we accept as being in our collective and mutual interests.[12] The rules of justice are necessary to establish and secure ownership of possessions. It is the utility of justice—specifically in creating stability in society—

that gives justice the status of a virtue, and it is from this that our moral obligation to justice arises. The mechanism by which it does this is sympathy.

Hume claims that sympathy is the source of the moral approval of justice and moral disapproval of injustice; indeed, "... sympathy produces our sentiment of morality in all the artificial virtues."[13] He argues that when we become aware of an act of injustice, no matter how remote from us it may be, we develop feelings of disapprobation from that act, which are subsequently directed towards the agent responsible for the act. We approve of that which causes pleasure and disapprove of that which causes pain. Recall that expressions of moral approval and disapproval take the form of "virtue" and "vice." Justice is therefore "a moral virtue, merely because it has that tendency to the good of mankind; and, indeed, is nothing but an artificial invention for that purpose."[14]

The moral obligation to act in accordance with virtue and to abstain from acting in accordance with vice is, according to Hume, derivative from these notions of virtue and vice.[15] As he says, "when any action or quality of the mind pleases us after a certain manner, we say it is virtuous; and when the neglect or non-performance of it displeases us after a like manner, we say that we lie under an obligation to perform it."[16] Thus through sympathy comes the status of justice as a virtue and injustice a vice; consequently we regard the performance of just acts as our moral obligation.

2. The Sensible Knave's Objection

That everyone will accept the rules of justice is not guaranteed. As Hume notes,

> ... a sensible knave, in particular incidents, may think that an act of iniquity or infidelity will make a considerable addition to his fortune, without causing any

> considerable breach in the social union and confederacy. That *honesty is the best policy,* may be a good general rule, but is liable to many exceptions; and he, it may perhaps be thought, conducts himself with most wisdom, who observes the general rule, and takes advantage of all the exceptions.[17]

It is important to recognize the similarity between the knave's objection here and that raised by Hobbes's Foole. Both ask why one cannot violate the rules of justice if it appears to be in one's interest to do so and if one can do so with impunity.[18] Hobbes's reply to the Foole relies on appeal to prudence and, I have argued, is ultimately unsuccessful. Hume's reply to the knave strikingly parallels that of Hobbes with regard to its appeal to prudence; but just as Hobbes, as we have seen, adds a second layer to his argument by appealing to reason, viz., arguing that it is absurd to break a covenant, Hume adds a second layer to his argument by appealing to moral sentiment. Before turning to Hume's appeal to moral sentiment, let us first look at his appeal to prudence in replying to the knave.

3. Hume's Reply

Hume's Appeal to Prudence

According to Hume, the rules of justice command inflexible adherence; but as he notes, not every individual act will prove immediately beneficial to one who performs it. This leads to the further recognition that not every man will always desire to abide by the rules of justice. As he says, "'tis easily conceiv'd how a man may impoverish himself by a signal instance of integrity, and have reason to wish, that with regard to that single act, the laws of justice were for a moment suspended in the universe."[19]

Hume replies that one who is tempted to violate the rules of justice fails to see the benefits of the system as a whole. In accepting the rules of justice, one might forgo the satisfaction of some immediate interests in order to acquire long-term benefit. This benefit is reaped through membership in society. In Hume's view, the benefits gained through such membership outbalance any benefits possibly attained through non-membership. For him, eschewing membership of society is a mistake. As he claims, "A perfect solitude is, perhaps, the greatest punishment we can suffer."[20]

Hume presents two arguments in support of his claim that one ought to adhere inflexibly to the rules of justice. The first argument is that that violation of the rules of justice threatens the system as a whole:

> The happiness and prosperity of mankind, arising from the social virtue of benevolence and its subdivisions, may be compared to a wall, built by many hands, which still rises by each stone that is heaped upon it, and receives increase proportional to the diligence and care of each workman. The same happiness, raised by the social virtue of justice and its subdivisions, may be compared to the building of a vault, where each individual stone would, of itself, fall to the ground; nor is the whole fabric supported but by the mutual assistance and combination of its corresponding parts.[21]

Hume argues that "without justice, society must immediately dissolve, and every one must fall into that savage and solitary condition, which is infinitely worse than the worst situation that can possibly be suppos'd in society."[22] If the rules of justice establish property and possession, and are necessary for social cooperation, then violation of the

rules of justice will lead to the collapse of social cooperation. This would be a complete reply to the knave if it were true, for then there would be a coincidence between the rules of justice and the interests of individuals such that no one could get any advantage from disobedience. Unfortunately for Hume, however, it is not true. Society is not so fragile that occasional acts of disobedience will bring about its destruction, and hence his first argument fails.

Hume's second and more substantial argument for why we should be just is that in violating the rules of justice, the knave risks being found out by others, which will prevent him from entering into covenants with others and from reaping the benefits of membership of society. As he argues,

> Such a one [who obeys the rules of justice] has, besides, the frequent satisfaction of seeing knaves, with all their pretended cunning and abilities, betrayed by their own maxims; and while they purpose to cheat with moderation and secrecy, a tempting incident occurs, nature is frail, and they give into the snare; whence they can never extricate themselves, without a total loss of reputation, and the forfeiture of all future trust with mankind.[23]

For Hume, the loss of trust of others is too serious a consequence of unjust acts and ought to deter one from so acting. But Hume here overstates his case. A single violation will not necessarily or typically yield "a *total* loss of reputation" or "the forfeiture of *all* future trust." As we have seen in the case of the Foole, even if it did, the payoff of a violation may be so high that it might be worth the gamble. Furthermore, when the risks are not so high, as they typically are not, it is easier to justify disobedience on cost-

benefit grounds. Thus, Hume's second argument fails to show that the sensible knave should not selectively violate the rules of justice.

Hume's Appeal to Moral Sentiment

Hobbes has to rely on external sanctions to answer the challenge raised by the Foole. Since human beings are egoistic creatures (i.e., psychologically they cannot act other than in their own interest), Fooles cannot feel badly about their actions. Hume, however, is not so handicapped and can have recourse to internal sanctions. He holds a conception of man quite different to that of Hobbes: according to him, we are social beings, capable of feelings of sympathy, and are not driven purely by egoist concerns.

The essence of Hume's appeal to sentiment in his reply is that the knave, in acting unjustly, can be expected to feel badly about himself. As Hume notes, "Inward peace of mind, consciousness of integrity, a satisfactory review of our own conduct; these are circumstances, very requisite to happiness, and will be cherished and cultivated by every honest man, which feels the importance of them."[24] Hume argues that the knave cannot achieve happiness and contentment through the unjust acts that he performs. The dishonest man will not be able to have the "inward peace of mind" that Hume claims is necessary for his happiness.

Hume claims there is a close connection between virtue and personal happiness and between vice and personal misery. This is the result of the workings of the indirect passions. Hume divides the indirect passions into two pairs: love and hate, on the one hand, and pride and humility on the other.[25] Both sets of passions originate from some cause such as perceiving an action where a person is kind or cruel to another. The pleasure or pain that the action produces in the observer (through the mechanism of

sympathy) also produces in him independent pleasures or pains which manifest themselves as love or hate, or pride or humility. If it is someone other than ourselves who is producing the pleasure or pain, we feel the indirect passion of love or hate. If it is ourselves, we feel pride or humility. Feelings of pride will aid in securing happiness while feelings of humility will hinder happiness.

Hume points out that the happiness attained through feelings of pride generated by the indirect passions depends not only on how a person views himself, but also on how others view him.[26] Human beings have a tendency to measure their self worth according to how others regard them. He says:

> By our continual and earnest pursuit of character, a name, a reputation in the world, we bring our own deportment and conduct frequently in review, and consider how they appear in the eyes of those who approach and regard us. This constant habit of surveying ourselves, as it were, in reflection, keeps alive all the sentiments of right and wrong, and begets, in noble natures, a certain reverence for themselves as well as others, which is the surest guardian of every virtue.[27]

One's sense of character (i.e., whether it be virtuous or vicious) is thus dependent on the manner in which others regard one. If others regard a person as having a virtuous character, then that person will likely view himself in a similar light, and subsequently have feelings of pride. Likewise, if others share the view that he has a vicious character, it can be expected that he too will view himself in this manner, and feel humility (or shame or guilt).

It seems, then, that determining whether to adhere to the rules of justice cannot merely be reduced to a question of whether one will be found out for one's violations, and so disqualified from society and the making of agreements. The knave, in violating a rule of justice, risks compromising his measure of self worth, and thus his happiness. Hume seems to think that forgoing this happiness is not worth the material gains attainable through violation. As he says, "... for they [knaves] are, in the end, the greatest dupes, and have sacrificed the invaluable enjoyment of character ... for the acquisition of worthless toys and gewgaws."[28] In the evaluation which follows, I will argue first that not all violations will produce feelings of guilt in one who violates and, second, that feelings of guilt will not in all cases override material gains.

4. Why Hume's Answer Fails

Not All Violations Produce Guilt

Hume is committed to the view that the rules of justice command strict and inflexible adherence, but it is not obvious that his appeal to moral sentiment supports such absolute adherence to these strict rules. To make this clear, it might be useful to distinguish between two different "knaves." I will here use the term "knave" to describe any individual who accepts, generally, the rules of justice, but makes exceptions in his own case. The important point here is that a knave is one who wilfully breaks inflexible rules of justice. The first sort of knave is one who breaks the rules of justice for his own self-interest. The second is one who breaks the rules of justice for the interests of others. Hume's appeal to moral sentiment, I will argue, fails to warrant the assertion that either knave ought to (or will) feel badly about himself for acting unjustly.

Consider first the knave that breaks the rules of justice to further his own self-interest. We likely will agree that he ought to feel badly about himself if he, for example, stole a large sum of money from an honest man. The principle of sympathy here will yield a feeling of disapprobation for such an act. We can sympathize with and feel badly for him from whom the money is stolen. We conclude that the knave should (and perhaps typically will) feel badly. Now consider the case where the knave steals a loaf of bread from a large grocery chain to feed himself. In such a case, the loss to the store is negligible, whereas the gain to the knave is substantial. It is here harder to condemn such a violation of justice where there is such substantial gain for minimal loss. Consequently the knave can weigh the consequences of his actions and will likely not lose peace of mind, knowing that his actions will reap great benefit for himself and minimal loss for the store. Others might likewise view the knave's actions in this case as not worthy of generating feelings of disapprobation. It is therefore not clear that sympathy is effective in commanding adherence to inflexible rules when self-interested gain is disproportionate to the harm done through the violation of rules of justice.

The same goes for cases where the knave breaks a rule of justice to further the interests of others. Consider the second knave, who steals a loaf of bread from the large chain to feed his starving wife and child, or even to feed a stranger. The knave here can compare the loss of the store to the gain of those he helps and will likely not lose peace of mind for his violation of the rules of justice. Again, others too might not disapprove of the knave's character in this case.

In both of these cases there appears to be a conflict of virtues. On the one hand, the knave's feelings of benevolence for others or himself will lead him to violate the rules of justice. On the other hand, adherence to justice will lead the knave to act

contrary to benevolence. Given that justice requires strict and inflexible adherence, it seems that Hume is committed to the view that justice will trump other virtues. In cases where one has the option of being just or benevolent, for example, one must choose being just over being benevolent. Hume does not provide us with an argument for why this ought to be the case, and for many it will seem counterintuitive.

Guilt Will Not Always Override Gains

Even where violations produce guilt, it is not clear that violations are not in the knave's interest. Suppose, for example, that a truly sensible knave has made all relevant calculations and has correctly determined that violating an agreement will yield a utility so great that it is worth forgoing the "inward peace of mind" achieved only through honest behaviour. We may suppose that this loss is included in the knave's initial calculation and that the knave, on the basis of the projected gains, judges violation to be to his advantage. If this is the case, then it is not clear that he cannot reasonably (in the self-centred sense of rationality) violate the rules of justice in those cases where violation will be utility-maximizing. If guilt is viewed as just another pain (albeit of a special kind) it is reasonable to think that it sometimes must be able to be outweighed by a greater amount of pleasure. If so, a sensible knave may look at the balance-sheet and violate.

Gerald Postema, however, suggests a possible defence of Hume's position. This defence admits that a sensible knave might make substantial material gains from violating rules of justice, but claims that these come at an even greater cost. Postema argues that given that the knave's objection relies on the possibility of successful

deception, successful exploitation will come at the cost of cutting oneself off from others and, on Hume's psychology, this will be a substantial loss. As Postema says,

> To cut oneself off from others is to cut oneself off *from oneself,* for it is only in the mirror of the souls of others that one finds one's own self, one's "character." The pleasures and satisfactions of conversation and intercourse are essential to human life, because they are essential to a sense of one's continuity through a constantly changing external and internal world. But the knave's strategy requires for success just the sort of "privacy" that makes such intercourse impossible, for the central governing principle of one's life must be scrupulously kept from public knowledge. One's true perspective must be kept private. Thus, a truly successful strategy of deception effectively cuts oneself off from the community in which alone one can find the confirmation essential to one's own sense of self. The knave is cut off from enjoyment of a character, not because his character is *bad,* but because he *has no* character at all.[29]

Postema argues that since a knave relies on the perspective from which others view him in order to affirm his own character, and since such a perspective is a result of deception, the knave will lose his concept of self. For, "the only perspective from which he can assess his own behaviour, and uncover the shape of a self in it, is radically at odds with the policy he has privately set for himself. Or rather, on this view, one cannot coherently adopt such a policy, for it requires one simultaneously to adopt and to reject the only perspective from which he can view himself."[30]

It is not perfectly clear that we can only see ourselves as others see us, and that successful deception of others will result in an inaccurate perception of ourselves. Nor is it obvious that this would be such a terrible thing that no gains can possibly outweigh it. But even if we assume that the loss of character is a far greater cost than any potential gains available through exploitation, there is a way to avoid this cost. Postema suggests that the knave might be able to selectively adhere to justice and so maintain a sense of self.[31] The knave can, for example, adhere to the rules of justice when interacting only with those with whom he shares close relationships, such as his family and friends. This solution is not incompatible with Hume's view, for, according to him it is possible that the human need for self-affirmation by others can be attained through interaction only with close relatives.[32] Thus the knave can attain such affirmation by those closely related to him, and can "take advantage" of the exceptions to justice when interacting with those more remote from him. In so doing, the knave can avoid the loss of character that universal deception entails, while reaping those gains that he can through successful exploitation of those more remote from him.

Neither appeal to prudence nor moral sentiment, I have argued, can produce a satisfactory answer to the objection raised by the knave. Hume's claim that "if a man think that this reasoning [that honesty is the best policy is a good general rule, but is liable to many exceptions] much requires an answer, it will be a little difficult to find any which will to him appear satisfactory and convincing"[33] can be viewed as a concession for which he has no real answer.[34] If the prudential case for adhering to the rules of justice fails, Hume has nothing left to say to those who are not moved by sympathy to do so. Hume is, I think, right in conceding this. We cannot easily convince someone who cannot feel pain not to touch a hot stove. Likewise, we cannot convince one who lacks

sympathy to adhere to the rules of justice. Yet one need not lack sympathy to remain unmoved by considerations of justice. As I have argued, sympathy will not generate moral disapproval of all violations of the rules of justice, nor will it yield guilt that cannot be outbalanced by substantial gains. Hence it will not support inflexible adherence to such rules.

It thus seems that we cannot show that violating rules of justice is necessarily contrary to one's interest. If we can benefit materially from doing so, and do not suffer psychologically by way of guilt or loss of self, it cannot be contrary to reason to violate those rules. The only objection one could make is that doing so is unfair. If justice is a virtue that requires general obedience to the rules if benefits are to come about, one who takes the benefits but does not share in the burdens is a free-rider. But showing that an action is unfair does not show that it is contrary to our interest to perform and, hence, that we have any reason (in the self-centred sense) not to do it.

Notes

[1] David Hume, *Treatise of Human Nature,* ed. L.A. Selby-Bigge and P.H. Nidditch, (Oxford: Oxford University Press, 1978), p. 458.

[2] Hume, *Treatise,* p. 471.

[3] Hume, *Treatise,* p. 477.

[4] Hume, *Treatise,* p. 575.

[5] Hume, *Treatise*, p. 386.

[6] Hume, *Treatise,* p. 578.

[7] Hume, *Treatise,* p. 484.

[8] Hume, *Enquiries,* ed. L.A. Selby-Bigge and P.H. Nidditch, (Oxford: Oxford University Press, 1975), p. 192.

[9] David Gauthier ("David Hume, Contractarian") has argued that Hume's political theories are contractarian. This is a surprising claim given what we know of the theories of Hobbes and Hume. On Hume's account, there are no promises, no absolute sovereign, no egoist psychology. Hume even denies that his theories are contractarian. But Gauthier does not intend to contradict Hume's assertion that his theories are not contractarian. Instead, he wants to show that the underlying similarities between the theories of Hobbes and Hume point to a contractarian character of Hume's theories of property and justice, and government and obedience.

[10] Hume, *Enquiries,* p. 306.

[11] Hume, *Treatise,* p. 490.

[12] It is important to note the similarity between Hume's and Hobbes's theories here. According to both theories, human beings accept the restraints that justice (or morality) imposes in view of the fact that doing so will be to their long-term benefit. It is "enlightened self-interest" that causes us to accept the rules of justice on both accounts.

[13] Hume, *Treatise,* p. 577.

[14] Hume, *Treatise,* p. 577.

[15] Terence Penelhum, *Hume,* (London: MacMillan Press Ltd., 1975), p. 152.

[16] Hume, *Treatise,* p. 517.

[17] Hume, *Enquiries,* pp. 282-3.

[18] David Gauthier ("Three Against Justice") argues that the Foole is making a stronger claim than that made by the Knave. The Foole, who concerns himself solely with self-interested pursuit, in effect denies justice when he says "there is no such Justice." The Knave, on the other hand, accepts justice as "a good general rule," though holds it to be "liable to many exceptions." The difference here lies in the fact that one explicitly denies justice, whereas the other regards it as mere policy.

[19] Hume, *Treatise,* p. 497.

[20] Hume, *Treatise,* p. 363.

[21] Hume, *Enquiries,* p. 305.

[22] Hume, *Treatise,* p. 497.

[23] Hume, *Enquires,* p. 283.

[24] Hume, *Enquiries,* pp. 283-4.

[25] Hume, *Treatise,* pp. 277-280, 329-332.

[26] Paul Russell, *Freedom and Moral Sentiment: Hume's Way of Naturalizing Responsibility,* (New York and Oxford: Oxford University Press, 1995), pp. 157-8.

[27] Hume, *Enquiries,* p. 276.

[28] Hume, *Enquiries,* p. 283.

[29] Gerald J. Postema, "Hume's Reply to the Sensible Knave," *History of Philosophy Quarterly,* vol. 5 (1988), p. 35.

[30] Postema, p. 35.

[31] Postema, p. 36.

[32] Hume, *Treatise,* pp. 363, 591.

[33] Hume, *Enquiries,* p. 283.

[34] David Gauthier, "Three Against Justice: The Foole, The Knave, and The Lydian Shepherd," *Midwest Studies in Philosophy,* vol. 7 (1982), p. 22.

CHAPTER 3

GAUTHIER AND CONSTRAINED MAXIMIZATION

The Foole and the knave raise the fundamental question of why one should be moral. In the preceding chapters, I have argued that Hobbes and Hume fail to give effective replies to their respective amoral counterparts. Specifically, I have argued that neither Hobbes's appeal to external sanctions nor Hume's appeal to moral sentiments is able to support strict adherence to the rules of justice in all cases. Since it cannot be shown that it is *always* in one's interest to be moral, the Foole and the knave remain undefeated, and the question of why one should be moral, unanswered.

There does remain hope, however, for the moralist. David Gauthier agrees that Hobbes and Hume fail in their respective endeavors to reply to the Foole and the knave, but denies that this entails that one cannot derive morality from self-interest. He argues that we can render morality and (self-interested) rationality compatible by appealing to dispositions rather than directly to actions. Gauthier seeks to show that it is rational to cultivate a disposition to be just and that non-maximizing actions which flow from that disposition are rational.[1] In this chapter, I will outline Gauthier's view, but will argue that he too fails to reconcile morality and (self-interested) rationality.

1. Appeal to Dispositions

Recall that the Foole and the knave hold that rational persons ought to adhere to the rules of justice when it is in their interest to do so, but that they are liable to violate these rules when

it appears in their interest to do so, and are reasonable in so doing. According to Gauthier, the objections raised by the Foole and the knave cannot be answered so long as one maintains a direct link between "rationality" and self-interest, broadly understood as the direct maximization of expected utility.[2] If it is to one's benefit to violate a given rule of justice, and if rationality dictates that one act to promote one's interest, then it is not contrary to reason to violate the rules of justice in those cases where one stands to benefit from so violating.

To answer the objections that the Foole and the knave raise, Gauthier argues that it has to be shown that it is rational to adhere to one's covenants in those cases where violation appears to be to one's benefit. This, he argues, can be done by appealing to the rationality of just dispositions. According to Gauthier, the Foole and the knave, in violating their covenants when doing so is to their benefit, make themselves unfit for society. In cultivating a disposition to adhere to their covenants only when it is in their interest to do so, the Foole and the knave no longer remain as trustworthy candidates with whom to make agreements. Others will no longer want to make agreements with them, and so (as the argument goes) the Foole and the knave will not be able to reap the benefits that making agreements with others provides. Reaping these benefits is assumed to be a greater benefit than any benefits attainable through violation. The Foole and the knave therefore, according to Gauthier, mistake their true interest and disable themselves from attaining the benefits of cooperation. Gauthier acknowledges that both Hobbes and Hume allude to this point in their replies to the Foole and the knave, but argues that they do not properly develop it.[3] To do

this, Gauthier argues that one must appeal to the rationality of cultivating a just disposition. As he says,

> Only the person truly disposed to honesty and justice may expect fully to realize their benefits, for only such a person may rationally be admitted to those mutually beneficial arrangements—whether actual agreements or implicitly agreed practices—that rest on honesty and justice, on voluntary compliance. But such a person is not able, given her disposition, to take advantage of the 'exceptions'; she rightly judges such conduct irrational. The Foole and the sensible knave, seeing the benefits to be gained from the exceptions, from the advantageous breaches in honesty and compliance, but not seeing these benefits, do not acquire the disposition. Among knaves they are indeed held for sensible, but among us, if we be not corrupted by their smooth words, they are only fools.[4]

2. The Formal Argument

Gauthier's argument runs as follows:

P1: Having a just disposition, i.e., a disposition to keep one's agreements, given sufficient security that others will too, will afford one a utility equal to or greater than that afforded by an unjust disposition.[5]

P2: A disposition is rational if and only if it affords one a utility at least equal to that afforded by any alternative disposition.[6]

C1: Therefore, rational persons will cultivate a just disposition.

P3: Actions which flow from a rational disposition are themselves rational.[7]

P4: Some actions which flow from a just disposition will be non-maximizing.[8]

C2: Therefore, non-maximizing actions which flow from a just disposition that it is rational to have are rational.

I will first examine Gauthier's argument for C1, and will then turn to his argument for C2.

Straightforward and Constrained Maximization

The Foole and the knave act solely to maximize the benefits that flow directly from their actions. They think it rational to adhere only to those agreements that promote their interests and to violate agreements when violation will produce more benefit than will adherence. The Foole and the knave are each what Gauthier calls a "straightforward maximizer," which he describes as "a person who seeks to maximize his utility given the strategies of those with whom he interacts."[9] It is worth nothing that the straightforward maximizer is not a simple-minded person who will opt for immediate advantages without any thought of long-term consequences. Geoffrey Sayre-McCord refers to straightforward maximizers as "enlightened egoists."[10] As such, they will carefully take into account the full consequences of what they do before they do it, and a large part of this will be the possibility for future co-operation: whether violations will make further agreements with this person impossible, whether further agreements with this person will be useful, whether the possibility of agreements with others will be impaired by loss of reputation, and so on. The

essential characteristic of the straightforward maximizer is that he will not automatically keep agreements even when there is suitable assurance that others will do the same, but will calculate whether it is, all things considered, advantageous to keep them.

By contrast, a constrained maximizer will automatically keep agreements (given assurances that others will do so). Gauthier defines a constrained maximizer as:

> ... a person who seeks in some situations to maximize her utility, given not the strategies but the utilities of those with whom she interacts ... We shall therefore identify a constrained maximizer thus: (i) someone who is conditionally disposed to base her actions on a joint strategy or practice should the utility she expects were everyone so to base his action be no less that what she would expect were everyone to employ individual strategies, and approach what she would expect from the co-operative outcome determined by minimax relative concession; (ii) someone who actually acts on this conditional disposition should her expected utility be greater than what she would expect were everyone to employ individual strategies. Or in other words, a constrained maximizer is ready to co-operate in ways that, it followed by all, would yield outcomes that she would find beneficial and not unfair, and she does co-operate should she expect an actual practice or activity to be beneficial. In determining the latter she must take into account the possibility that some persons will fail, or refuse, to act co-operatively.[11]

The constrained maximizer, unlike the straightforward maximizer, tends to adhere to his covenants without calculating whether violating them will promote his overall interest. The constrained maximizer possesses a just disposition. As Gauthier says,

> Constrained maximization thus links the idea of morals by agreement to actual moral practice. We suppose that some moral principles may be understood as representing joint strategies prescribed to each person as part of the ongoing co-operative arrangements that constitute society. These principles require each person to refrain from the direct pursuit of her maximum utility, in order to achieve mutually advantageous and reasonably fair outcomes. Actual moral principles are not in general those to which we should have agreed in a fully rational bargain, but it is reasonable to adhere to them in so far as they offer a reasonable approximation to ideal principles.[12]

Thus, the constrained maximizer cooperates when he expects others to cooperate and where co-operation will not yield a utility significantly less than will defection, and defects only when he anticipates that others will do the same.

Which Strategy is Rational?

The rationality of possessing a given disposition will depend on what strategies others use, and on one's and others' abilities to predict which strategy each will adopt, where "a disposition is rational if and only if an actor holding it can expect his choices to yield no less

utility than the choices he would make were he to hold any alternative disposition."[13] To determine which strategy will be most beneficial (and thus rational), let us recall the relevant payoffs of Prisoner's Dilemma-type cases:

	Player 2 violates	Player 2 adheres
Player 1 violates	3 / 3	1 / 4
Player 1 adheres	4 / 1	2 / 2

Consider a society composed of a mix of straightforward and constrained maximizers. Each type of strategist will violate in all but two cases. First, two constrained maximizers interact and are able to identify each other as constrained maximizers. In such a case, both will cooperate. Second, a constrained maximizer mistakes a straightforward maximizer for a constrained maximizer. In such a case, the constrained maximizer will adhere to the agreement, and the straightforward maximizer will defect.[14] Thus, all possible interaction between individuals will proceed as follows:

(1) Constrained maximizers will cooperate with other constrained maximizers, so long as they are able to identify each other as such.

(2) Constrained maximizers will defect when interacting with straightforward maximizers, so long as they are able to identify them as such.

(3) Constrained maximizers will adhere to those agreements with straightforward maximizers if they mistake them for fellow constrained maximizers.

(4) Straightforward maximizers will defect when interacting with constrained maximizers when they are mistaken for constrained maximizers.

(5) Straightforward maximizers will defect when interacting with other straightforward maximizers when it is in their interest to do so.

The relevant payoff scheme for Prisoner's Dilemma-type cases is then as follows:

Case (1): Interaction between two constrained maximizers will give a payoff of 2 / 2. Each player receives the second-best possible payoff by mutual adherence.

Case (2): Interaction between a constrained maximizer who identifies his opponent as a straightforward maximizer will give a payoff of 3 / 3. Each player receives the third-best outcome by mutual defection.

Case (3): Interaction between a constrained maximizer who mistakes his straightforward maximizing opponent as a constrained maximizer will give a payoff of 4 / 1. The constrained maximizer will receive the lowest possible payoff and the straightforward maximizer will receive the highest possible payoff.

Case (4): Interaction between a straightforward maximizer who is mistaken for a constrained maximizer by his constrained maximizing opponent will give a payoff of 1 / 4. The straightforward maximizer will receive the highest possible payoff and the constrained maximizer will receive the lowest possible payoff.

Case (5): Interaction between two straightforward maximizers will give a payoff of 3 / 3. Each player will receive the third best outcome.

The best possible outcome is thus successful exploitation. If the straightforward maximizer is able to fool the constrained maximizer into thinking that he is a constrained maximizer, then he will be able to reap substantial rewards. Likewise, the worst possible scenario is being deceived, as what happens when the constrained maximizer interacts with a straightforward maximizer, but falsely believes him to be a constrained maximizer. The second best outcome is mutual adherence, and the third best is mutual defection.

It seems that dominance shows us that straightforward maximization will afford one a greater utility than constrained maximization. For when neither party cooperates, both the straightforward and the constrained maximizer will benefit equally. Each will receive the third best payoff. Yet straightforward maximizers are able to gain through successful exploitation. Consider an encounter between a straightforward maximizer and a constrained maximizer. If the constrained maximizer adheres to the agreement while the straightforward maximizer defects, the straightforward maximizer will receive a payoff greater than he would had he adhered. The straightforward maximizer is thus able to reap advantages unavailable to the constrained maximizer through unilateral defection. Although the constrained maximizer has available to him opportunities for gain through mutual adherence, mutual adherence yields a utility less than does unilateral defection. It is thus difficult to see the rationale behind choosing to be a constrained maximizer. If one can maximize one's utility by violating, then why would one choose to adhere?

Gauthier argues that the dominance argument works only insofar as one's character has no effect on one's opportunity for cooperation.[15] He argues that in cases where one's character can remain relatively unknown by others, straightforward maximizers will tend to do better than constrained maximizers. For they can then continue to make agreements with others, and can take advantage of adherence of the constrained maximizers. The straightforward maximizer is able to benefit from successful exploitation only if it is assumed that his unjust disposition will have no effect on his opportunities for exploitation. But, Gauthier argues, when one's disposition to violate becomes known by others, rational persons will not enter into agreements with the untrustworthy person.[16] Thus, the straightforward maximizer's opportunities to maximize his utility through defection will be markedly decreased. As he says,

> Only those disposed to keep their agreements are rationally acceptable as parties to agreements. Constrained maximizers are able to make beneficial agreements with their fellows that the straightforward maximizer cannot, not because the latter would be unwilling to agree, but because they would not be admitted as parties to agreement given their disposition to violate. Straightforward maximizers are disposed to take advantage of their fellows should the opportunity arise; knowing this, their fellows would prevent such opportunity arising. With the same opportunities, straightforward maximizers would necessarily obtain greater benefits. A dominance argument establishes this. But because they differ in their dispositions,

straightforward and constrained maximizers also differ in their opportunities, to the benefit of the latter.[17]

Gauthier's argument here, however, requires further qualification in order to be entirely convincing. For, given that (1) straightforward maximizers will do better so long as they are given the same opportunities as constrained maximizers, (2) whether one has the same opportunity that a constrained maximizer does depends on whether one will be admitted into cooperative agreements with others, and (3) whether one is able to partake in agreements with others depends on whether it appears that one can be trusted, it seems that so long as one is able to maintain the illusion of having a just disposition, one will be able to partake in agreements with others while reaping the benefits of defection.

If people were transparent (i.e., their characters were always accurately detectable by others), then constrained maximization would be the preferred strategy.[18] For in this case constrained maximizers would be able to identify straightforward maximizers, and consequently exclude them from agreements. If people were, on the other hand, opaque (i.e., their characters remained hidden to others), then straightforward maximizers would do better.[19] For straightforward maximizers would then be able to continue to make agreements with others and gain through successful exploitation. Gauthier argues that neither of these is realistic, and claims that people are translucent. As he says, "we may appeal to … *translucency,* supposing that persons are neither transparent nor opaque, so that their disposition to co-operate or not may be ascertained by others, not with certainty, but as more than mere guesswork."[20] In other words, if our characters are translucent, it is

assumed that one has a better chance of correctly identifying another's character than they do of being mistaken. There are degrees of translucency and how much of a better chance one has depends on how translucent a person is.

If it is the case that our characters are translucent, and that one who is suspected to have an untrustworthy disposition will be excluded from cooperative agreements, then the straightforward maximizer (so the argument goes) will not reap all projected benefits of both cooperation and defection. The straightforward maximizer will forgo many benefits of cooperation and reap only those gains that he can through successful exploitation. But as his untrustworthy character becomes more widely known, opportunities for exploitation will diminish. As the opportunities for interaction available to the straightforward maximizer decrease, the constrained maximizer's chances of being exploited will decrease. Thus, even if it is the case that the potential benefits from defection are substantially greater than those of cooperation, if opportunities to gain benefits from defection are few, the straightforward maximizer's projected utility will likely decrease to below that of the constrained maximizer. Thus, Gauthier argues that to maximize projected utility, rational persons will become constrained maximizers or, as C1 in our formulation of Gauthier's argument has it, develop just dispositions.

3. Why Gauthier's Solution Fails

Gauthier reaches the conclusion that it is rational to dispose oneself to morality by arguing that rational persons will choose constrained maximization over any alternative disposition based on the higher expected utility that such a disposition will yield. Since constrained

maximization leads one to act consistently with morality, rational persons will therefore choose to dispose themselves to act morally. In the evaluation that follows, I will first examine Gauthier's claim that constrained maximization is a strategy that rational persons ought to adopt, i.e., his argument for C1. I will then turn to his claim that if constrained maximization is rational, all actions which flow from this disposition are themselves rational, i.e., his argument for C2.[21]

C1: Constrained Maximization is Not Rational

Gauthier argues that constrained maximization is a rational disposition to adopt. Since a disposition is rational if and only if it affords one a utility equal to or greater than that afforded by another disposition, Gauthier must think that constrained maximization will afford one a utility equal to or greater than any other. This is questionable. I will give two arguments to this effect.

First, it is not clear that constrained maximization will yield a higher utility than will straightforward maximization for all agents. Gauthier defends the rationality of constrained maximization based on the average amount of utility that both it and straightforward maximization will yield.[22] Since constrained maximization, he argues, will yield on average a greater utility than straightforward maximization, constrained maximization is rational. That constrained maximization affords one a utility at least equal to that afforded by straightforward maximization, however, depends on the degree to which an agent's character is translucent. If an agent disposed to constrained maximization is correctly able to identify those with whom he interacts, and others are correctly able to determine his

character, then he will be able to reap those benefits only attainable through cooperation. He will also be able to continue to enjoy these benefits given the trustworthiness of his character. These benefits, Gauthier argues, are not attainable to one who is suspected to be disposed to straightforward maximization. Thus, it is argued that constrained maximization is utility maximizing.

It should be noted, however, that people will vary in the degree to which their characters are translucent. To assume otherwise is unrealistic. It seems, then, that constrained maximization will yield varying utilities to different agents and, as the degree to which agents are translucent decreases, so will the gap be narrowed between the utilities afforded to an agent through constrained and straightforward maximization. As Gauthier argues,

> If persons are translucent, then constrained maximizers (CMs) will sometimes fail to recognize each other, and will then interact non-co-operatively even if co-operation would have been mutually beneficial. CMs will sometimes fail to identify straightforward maximizers (SMs) and will then act co-operatively; if the SMs correctly identify the CMs they will be able to take advantage of them. Translucent CMs must expect to do less well in interaction than would transparent CMs; translucent SMs must expect to do better than would transparent SMs. Although it would be rational to choose to be a CM were one transparent, it need not be rational if one is only translucent.[23]

Straightforward maximizers therefore tend to do better as they become less transparent. Conversely, constrained maximizers do worse as their transparency decreases. Thus, if agents are able to conceal their characters such that their translucency is minimal, then it seems that a straightforward maximizer will sometimes attain a higher utility than will a constrained maximizer. It follows that constrained maximization will not be rational for all agents.

Second, there are other dispositions that it might be more rational to adopt than those that Gauthier considers.[24] Gauthier seems to ignore the possibility that dispositions other than straightforward and constrained maximization are available to agents. He argues that constrained maximizers will do better than straightforward maximizers. However, this does not entail that constrained maximization is the strategy that rational persons ought to adopt. David Copp, for example, suggests the disposition of "reserved maximization" as an alternative to constrained maximization, where one so disposed will act much like the constrained maximizer in most circumstances, but will violate an agreement in cases where the constrained maximizer will adhere, such as when he stands to gain substantially and the likelihood of detection is significantly low.[25] While a constrained maximizer will refrain from taking the contents of a found wallet, the reserved maximizer will take the money if the sum is substantial and the possibility of being found out is low. Since one will almost always act like a constrained maximizer, one will gain the benefits of co-operation. But since one will not absolutely close the door on a "big score," one will also have some of the benefits of a straightforward maximizer. There is nothing incoherent or odd about this disposition. Doctors are taught that when they hear hoof beats, think horses; but also to be

on the look-out for zebras. Likewise, the rational agent may develop the disposition to comply when he makes agreements, and generally to do so even when that is non-maximizing, but to keep an eye open for special opportunities.

C2: Actions that Flow From a Rational Disposition are Not Rational

Even if we grant that constrained maximization is rational, it is not clear that those actions that flow from it are themselves rational. It seems that Gauthier relies on a transitive relation between dispositions and actions which flow from that disposition. He contends that those actions which flow from a disposition that it is rational to have are rational. But this is not obvious.

Consider, for example, deterrent threats. It might be rational for a country to make a threat and dispose itself to carry it out if such a disposition is likely to yield a sufficient level of security against attack.[26] But what happens in the case of failed deterrent threats? Suppose that country A threatens country B that it will activate a doomsday device should country B attack. Country B nevertheless launches a first strike. Is it rational for country A to follow through with its original threat and activate the doomsday device? Gauthier must claim that it is, given his conviction that actions expressing a disposition that it is rational to possess are rational. Country A has disposed itself to carry through with threats given the high expected utility (i.e., security against attacks) that such a disposition can be expected to yield. The disposition to carry through with threats is therefore rational. Therefore, those actions which flow from this disposition are themselves rational. It must therefore be rational for country A to activate the doomsday device. But, given that activation of the

doomsday device will result in massive levels of destruction, it is not obvious that this is so. Such an implication might lead one to reject Gauthier's claim that those actions which flow from a rational disposition are rational.

In light of this, one might argue (as Kavka has[27]) that constrained maximization should be viewed as an instance of a rational disposition to perform irrational acts. In other words, one can claim that there are cases (as in the case of disposing oneself to constrained maximization) where it is rational to dispose oneself to perform irrational acts. In so doing, one can maintain the rationality of disposing oneself to carry out threats while conceding that some acts which flow from such a disposition are irrational, as in the case of activating the doomsday device. However, Gauthier explicitly denies that constrained maximization is such an instance, and insists that those acts which flow from it are rational.[28] As he says,

> ... it may be rational for a perfect actor [i.e., an actor who lacks no weakness or imperfection in her reasoning] to dispose herself to threat enforcement, and if it is, then it is rational for her to carry out a failed threat. Equally, it may be rational for a perfect actor to dispose herself to threat resistance, and if it is, then it is rational for her to resist despite the cost to herself. Deterrence, we have argued elsewhere, may be a rational policy, and non-maximizing deterrent choices are then rational.[29]

Gauthier thus cheerfully accepts the awkward consequences that his endorsement of constrained maximization entails. He must do this in order to maintain his view that those actions which flow from a disposition that it is rational to have are rational. But his

acceptance of these consequences is not sufficient to persuade one who is skeptical of the rationality of these actions.

I thus conclude that the move from C1 to C2 is problematic. But even if it were not, Gauthier's attempt to reconcile morality and rationality still fails because of his failure to establish C1. The real weakness in Gauthier's argument for C1 is his need to rely on the transparency of others' characters and intentions. Gauthier seems to recognize how crucial detection of intentions is. He says, "The ability to detect the disposition of others must be well developed in a rational CM. Failure to develop this ability, or neglect of its exercise, will preclude one from benefiting from constrained maximization. And it can then appear that constraint is irrational. But what is actually irrational is the failure to cultivate or exercise the ability to detect others' sincerity or insincerity." Given the low hope one must have of being able to develop those abilities of detection, this seems a damming admission.

Notes

[1] David Gauthier, *Morals By Agreement,* (Oxford: Oxford University Press, 1986), pp. 162-165.

[2] Gauthier, p. 162..

[3] Gauthier, pp. 162-165, 182.

[4] Gauthier, p. 182.

[5] Gauthier, pp. 174-179.

[6] Gauthier, pp. 182-183.

[7] Gauthier, p. 186.

[8] Gauthier, p. 169.

[9] Gauthier, p. 167.

[10] Geoffrey Sayre-McCord, "Deception and Reasons to Be Moral," *Contractarianism and Rational Choice: Essays on David Gauthier's Morals By Agreement,* ed. Peter Vallentyne, (Cambridge: Cambridge University Press, 1991), p. 185.

[11] Gauthier, pp. 167-168.

[12] Gauthier, p. 168.

[13] Gauthier, pp. 182-183.

[14] Geoffrey Sayre-McCord, p.189.

[15] Gauthier, p. 173.

[16] Gauthier, p. 172.

[17] Gauthier, p. 173.

[18] Gauthier, p. 174.

[19] Gauthier, p. 174.

[20] Gauthier, p. 174.

[21] Gauthier, pp. 174-179.

[22] David Copp, "Contractarianism and Moral skepticism," *Contractarianism and Rational Choice: Essays on David Gauthier's Morals By Agreement,* ed. Peter Vallentyne, p. 220.

[23] Gauthier, p. 174.

[24] David Copp, Review of *Morals By Agreement, Philosophical Review,* vol. 98, no. 3 (July 1989), p. 413.

[25] David Copp, "Contractarianism and Moral Skepticism," pp. 220-221.

[26] David Copp, "Contractarianism and Moral Skepticism," p. 206.

[27] Gregory Kavka, Review of *Morals By Agreement, Mind,* vol. 96 (1987), pp. 120-121.

[28] Gauthier, p. 186.

[29] Gauthier, p. 186.

CONCLUSION

In this essay I have looked at the attempts of Hobbes, Hume and Gauthier to reconcile morality and rationality. All three, I have argued, are unable to do so satisfactorily.

I have argued that rationality in the self-centred sense requires that one be a straightforward maximizer. Straightforward maximization requires that one break the rules of morality when others adhere to them. Breaking those agreements to which others adhere are "offensive violations." None of Hobbes, Hume or Gauthier endorses such violations. It follows that they do not provide a theory of morality (or justice) which comports with rationality.

There is also reason to think that any attempt of reconciliation must fail. Offensive violators that take advantage of the benefits of the moral system without incurring the cost are "free-riders." Free-riding is a paradigm case of immoral behaviour. Thus, if I am correct in claiming that rationality requires free-riding, rationality in the self-centred sense and morality cannot be reconciled.

The Foole and the knave thus win the argument. But they are not heroes, and conceding defeat to them does not entail that we ought to perform those actions that they recommend. What the Foole and the knave show is that certain violations of the rules of morality are necessary in order to remain consistent with a commitment to rationality in the self-centred sense. It thus seems that we are left with a choice between second-best options: either (1) embrace rationality in the self-centred sense and turn our backs on morality, (2) find another sense of rationality such as the universalistic sense that might more readily be

reconciled with morality, or (3) give up on rationality altogether and embrace morality on a non-rational basis.

BIBLIOGRAPHY

Axelrod, Robert. "The Emergence of Cooperation among Egoists." *American Political Science Review,* vol. 75, no. 2 (1981), pp. 306-318.

Ayer, Alfred Jules. *Hume.* Oxford: Oxford University Press, 1980.

Baier, Annette. *A Progress of Sentiments: Reflections on Hume's Treatise.* Cambridge, Mass.: Harvard University Press, 1991.

Baier, Kurt. *The Moral Point of View: A Rational Basis of Ethics.* New York: Random House Inc., 1965.

Baldwin, Jason. "Hume's Knave and the Interests of Justice." *Journal of History of Philosophy,* vol. 42, no. 3 (2004), pp. 277-296.

Baron, Marcia. "Hume's Noble Lie: An Account of His Artificial Virtues." *Canadian Journal of Philosophy,* vol. 12, no. 3 (1982), pp. 539-556.

Barry, Brian. *Theories of Justice.* Berkeley and Los Angeles: University of California Press, 1989.

Copp, David. Review of David Gauthier's *Morals By Agreement. Philosophical Review,* vol. 98, no. 3 (July 1989), pp. 411-414.

_____. "Contractarianism and Moral Skepticism." *Contractarianism and Rational Choice: Essays on David Gauthier's Morals by Agreement.* Ed. Peter Vallentyne. Cambridge: Cambridge University Press, 1991, pp. 196-228.

Darwall, Stephen. *Impartial Reason.* Ithaca: Cornell University Press, 1983.

_____. "Kantian Practical Reason Defended." *Ethics*, vol. 96 (October 1985), pp. 89-99.

Foot, Philippa. "Moral Beliefs." Philippa Foot, *Virtues and Vices.* Berkeley and Los Angeles: University of California Press, 1978, pp. 110-131.

Falk, W.D. "Morality, Self, and Others." *Morality and the Language of Conduct.* Ed. Hector-Neri Castaneda and George Nakhnikian. Detroit: Wayne State University Press, 1965, pp. 25-67.

Gauthier, David. *Practical Reasoning.* Oxford: Clarendon Press, 1963.

_____. "Morality and Advantage." *Morality and Rational Self-Interest.* Ed. David Gauthier. Englewood Cliffs, NJ: Prentice Hall, Inc.,1970, pp. 166-180.

_____. "Bargaining Our Way Into Morality: A Do-It-Yourself Primer." *Philosophic Exchange,* vol. 2, no. 5 (1979), pp. 14-27.

_____. "Three Against Justice: The Foole, The Sensible Knave, and the Lydian Shepherd." *Midwest Studies in Philosophy,* vol. 7 (1982), pp. 11-29.

_____. *Morals By Agreement.* Oxford: Oxford University Press, 1986.

_____. "David Hume, Contractarian." *Moral Dealings: Contract, Ethics, and Reason.* Ed. David Gauthier. Ithaca: Cornell University Press, 1990, pp. 45-76.

_____. "Justice and Natural Endowment: Toward a Critique of Rawls's Ideological Framework." *Moral Dealings: Contract, Ethics, and Reason.* Ed. David Gauthier. Ithaca: Cornell University Press, 1990, pp.150-170.

Hampton, Jean. *Hobbes and the Social Contract Tradition.* Cambridge: Cambridge University Press, 1986.

_____. "The Hobbesian Side of Hume." *British Journal for the History of Philosophy,* vol. 8, no. 1 (March 2000), pp.167-195.

Harrison, Jonathan. *Hume's Moral Epistemology.* Oxford: Clarendon Press, 1976.

_____. *Hume's Theory of Justice.* (Oxford: Clarendon Press,1981).

Hobbes, Thomas. *Leviathan.* 1651.

Hume, David. *Enquiries.* Ed. L.A. Selby-Bigge and P.H. Nidditch. Oxford: Oxford University Press, 1975.

_____. *A Treatise of Human Nature.* Ed. L.A. Selby-Bigge and P.H. Nidditch. Oxford: Oxford University Press, 1978.

_____. *Essays: Moral, Political, and Literary.* Ed. Eugene F. Miller. Indianapolis: Liberty Fund, 1987.

Kant, Immanuel. *Foundations of the Metaphysics of Morals.* 1785.

_____. *The Critique of Practical Reason.* 1788.

Kavka, Gregory. *Hobbesian Moral and Political Philosophy.* Princeton, NJ: Princeton University Press, 1986.

_____. Review of David Gauthier's *Morals By Agreement. Mind,* vol. 96 (1987), pp. 117-121.

Mackie, J.L. *Hume's Moral Theory.* London: Routledge, 1980.

_____. *Ethics: Inventing Right and Wrong.* London: Penguin Books, 1990.

Macnabb, D.G.C. *David Hume: His Theory of Knowledge and Morality.* Oxford: Alden Press Ltd., 1966.

Milinski, Manfred. "Predator Inspection: Cooperation or 'Safety in Numbers'?" *Animal Behaviour,* vol. 43 (1992), pp. 679-80.

Nagel, Thomas. *The Possibility of Altruism.* Oxford: Oxford University Press, 1970.

Nozick, Robert. *Anarchy, State, and Utopia.* New York: Basic Books, Inc., 1976.

O'Day, Ken. "Hume's Distinction between the Natural and Artificial Virtues." *Hume Studies,* vol. 20, no. 1 (April 1994), pp. 121-41.

Penelhum, Terence. *Hume.* London: MacMillan Press Ltd., 1975.

Plato, *Gorgias.*

_____. *Republic.* Book I.

Postema, Gerald J. "Hume's Reply to the Sensible Knave." *History of Philosophy Quarterly,* vol. 5 (1988), pp.23-40.

Rawls, John. *A Theory of Justice.* Cambridge, Mass.: Harvard University Press, 1971.

Resnik, Michael. *Choices: An Introduction to Decision Theory.* Minneapolis: University of Minnesota Press, 1987.

Russell, Paul. *Freedom and Moral Sentiment: Hume's Way of Naturalizing Responsibility.* New York and Oxford: Oxford University Press, 1995.

Sayre-McCord, Geoffrey, "Deception and Reasons to Be Moral." *Contractarianism and Rational Choice: Essays on David Gauthier's Morals by Agreement.* Ed. Peter Vallentyne. Cambridge: Cambridge University Press, 1991, pp. 181-195.

Schmidtz, David. *Rational Choice and Moral Agency.* Princeton, NJ: Princeton University Press, 1995.

Skyrms, Brian. *Evolution of the Social Contract.* Cambridge: Cambridge University Press, 1996.

Smith, Holly. "Deriving Morality from Rationality." *Contractarianism and Rational Choice: Essays on David Gauthier's Morals by Agreement.* Ed. Peter Vallentyne. Cambridge: Cambridge University Press, 1991, pp. 229-253.

Vallentyne, Peter, ed. *Contractarianism and Rational Choice: Essays on David Gauthier's Morals by Agreement.* (Cambridge: Cambridge University Press, 1991).

Velleman, J. David. "The Possibility of Practical Reason." *Ethics,* vol. 106 (July 1996), pp. 694-726.