Fritz- Anton Fritzson
FPR 503
University of Lund

Deriving morality from rationality

-A defence of the rational-choice framework of David Gauthier's moral theory

Tutor: Magnus Jiborn

Introduction

In this essay I will argue that any plausible moral theory should be able to provide a good answer to the question "why should I be moral?" This is the question of *moral motivation*. I will further suggest that a contractarian moral theory, such as that of David Gauthier,[1] provides a plausible answer to this question. And that the ability to do this is due to the *rational-choice framework* of the contractarian theory. A moral theory that is able to provide an answer to the question of moral motivation is theoretically superior to other theories since it provides some *foundations* of morals.[2] Anyone who wants to claim superiority of some other type of moral theory, a theory not grounded in rational-choice, will either have to deny the importance that I give to moral motivation and perhaps also deny the possibility of rational foundations in moral philosophy or show how their own theory gives an equally good answer to the question of moral motivation.

The primary topic, and inspiration, for this essay is the contractarian moral theory of David Gauthier but the main ideas can be found as far back as in the writings of Plato[3] and Epicurus[4] and in full form in the writings of Thomas Hobbes[5]. It is important that we distinguish between two different types of theories referred to as "contractarian." Peter Danielson calls these different types *weak* and *strong* contractarianism respectively.[6] Weak contractarians, such as John Rawls, are working *within* morality; beginning with prior moral constraints and derive the principles of justice from there. Strong contractarians, such as Hobbes[7] and Gauthier, on the other hand, argues from premises of non-moral individual rational choice. I hope it will become clear as we go along why contractarianism in its weak form has to abandon almost everything that makes contractariansim attractive in the first place. I am here then, like Gauthier, working within the strong tradition.

First I will briefly say something about what I take to be our project and then what I take to be requirements of a reasonable moral theory. I will then argue that contractarianism lives up to these requirements and also that Gauthier's theory, with some modifications perhaps, answers the question of moral motivation and gives a *rational justification* of

[1] Proposed in his book Gauthier, D. (1986). *Morals by Agreement*. Oxford University Press
[2] Our approach is therefore foundationalist rather than coherentist. I assume that if foundationalism is possible it is superior to coherentism.
[3] The view presented by Glaucon in *The Republic*. Oxford University Press. 1998.
[4] Epicurus,"Letter to Menoeceus" in Cooper, D.E. (1998). *Ethics – The Classic Readings*. Blackwell Publishing
[5] Hobbes, T. (1651). *Leviathan.* 1996. Oxford university press
[6] Danielson, P. (1992) *Artificial Morality*. Routledge. pp 25- 26
[7] It is also important to note the distinction between Hobbes moral theory and his political theory. Here we have only his *moral contractarianism* in mind. I think that Hobbes political theory is the wrong application of his moral theory.

morality. To show this we need some clarity about what "rationality" is and also what we mean by "motivation" and "justification". We will explore some problematic aspects of the theory and an objection to the project of grounding morality in rational choice known as "the relevance objection".

We will *not*, however, discuss the particular *content*[8]of a rational morality except where prior knowledge of this is required for our arguments. We will also not discuss the rationality of keeping agreements[9] and we will not defend the theories of value that are implicit in the theory of rational choice.[10] We will instead concentrate on the *framework* of morality and that is, if we are right, the framework of *rational choice*. Our main arguments are from the perspective of moral motivation.

I follow Gauthier in that a plausible moral theory should be able to reach normative conclusions without introducing prior moral assumptions. He writes *"If the reader is tempted to object to some part of this view, on the ground that his moral intuitions are violated, then he should ask what weight such an objection can have, if morality is to fit within the domain of rational choice. We should emphasize the radical difference between our approach [...] from that of moral coherentists and defenders of "reflective equilibrium", who allow initial weight to our considered moral judgements"* [11] Here then, we allow no such initial weight. The main problem with appealing to moral intuitions in theory is that when intuitions conflict we have no further tools for reaching reasoned agreement. My *theoretical* intuitions tell me that it would surely be better if we could build a theory without appealing to any *moral* intuitions.

Our project

So if our project is to show that a good moral theory should be able to answer the question "why be moral?" we need to consider what kind of answer to this question would be satisfactory. We will call the person asking this question "the moral sceptic" despite the negative connotations that this brings. The sceptic is, as we see her here, neither an anti-social being nor an egoist. She is rather a fully rational person asking for reasons to do what

[8] In Gauthier's case this is the principle of minimax relative concession (see Gauthier. (1986). chapter V. "Co-operation: Bargaining and Justice", pp 113- 156) and the Lockean proviso (see ibid, chapter VII. "The initial bargaining position: Rights and the Proviso", pp 190- 232)
[9] Known as the "compliance problem" (see ibid, chapter VI. "Compliance: Maximization Constrained", pp 157-190)
[10] Gauthier claims that value-subjectivism and value-relativism are implicit in the theory of rational choice. He points to Gilbert Harman and John L. Mackie for a detailed defence of these value- theoretical positions. He follows Harman and finds objective value explanatory redundant. ibid, pp 55- 59
[11] ibid, p 269

morality requires. What kind of answer is the moral sceptic looking for then? To answer this we first need to distinguish between four different questions in moral philosophy, all with their own importance and appeal.

a.) What *is* morality
b.) Why *do* people act morally?
c.) Why *should* we act morally?
d.) What is/are the fundamental principles of morals?

I think that the sceptic is primarily interested in question C. So defending the possibility and importance of answering C is our main task here. (But in arguing for this we must, however, say quite a lot about A as well.) Evolutionary theories of morals seem to be focusing on question B. T.M. Scanlon on his account of moral motivation seems also to be interested primarily in question B. Scanlon writes *"I myself accept contractualism largely because the account it offers of moral motivation is phenomenologically more accurate then any other I know of"* [12] I will not dispute his claim here since we are not primarily concerned with what is "phenomenologically accurate" or, for that matter, evolutionary accurate. These other perspectives are very interesting in their own right but they do not help us much in answering the question of why we *should* be moral.

The person asking "Why?" is looking for a justification of morality. But what kind of justification and how are we to achieve one? Appealing to people's moral intuitions, like e.g. theorists of reflective equilibrium do, would not help us out here. That would be like answering: "most people don't think so!" and I don't think this would convince the sceptic. To appeal to intuitions in this way is to appeal to some *facts*. What about some other kinds of facts then? What if we could find the *true* morality or some "moral facts"? This would be an attempt to provide an *epistemic* justification of morality. Still, I think that the sceptic would be unmoved by our attempts. It would still be open for him or her to say "so what?" The sceptic is, I think, not asking us to point to some observational facts or some wrong-making "properties" (natural or non-natural). What we are looking for then is not an epistemic justification. The sceptic is rather asking for a *prudential* justification of morality. Providing such a thing would be to appeal to something to which the individual is already committed and show how these ends could be promoted by adhering to morality. Much of the remainder

[12] Scanlon. (1998). *What We Owe to Each Other*. The Belknap Press of Harvard University Press p 187

of the essay will be about if and how this is possible. But we need first to consider why we at all would like to show this.

Why moral motivation is important

Many have explicitly denied the importance of our project, T.M. Scanlon for example says that it would be *"misleading to say that we are looking for a way of justifying the morality of right and wrong to someone who does not care about it – an "amoralist" – because this suggests that what we are looking for is an argument that begins from something to which such a person must be already committed and shows that anyone who accepts this starting point must recognize the authority of the morality of right and wrong."* [13] We on the other hand *are* looking for a way of justifying morality and we *are* looking for an argument that begins from something to which a person must already be committed. But is this kind of motivation really important?

To answer this question we need to consider what morality is for. Is morality something that is "from above" and completely separate from the interests of human beings or has morality got something to do with our interests? What we mean by "interest" will hopefully become clear in our discussion of "rationality" below. But it is not hard to imagine that people have *different* interests and sometimes other people's interests come into conflict with our own. If these kinds of conflicts appear frequently we need *a rule*. The point of morality then is to solve conflicts of interest. Kurt Baier writes *"if the point of view of morality were that of self-interest, then there could never be moral solutions of conflicts of interest. However, when there are conflicts of interest, we always look for a "higher" point of view. . . . by 'the moral point of view' we mean a point of view which is a court of appeal for conflicts of interest"* [14] The idea is that both parties to the conflict would be worse off without such rules. Without moral rules we could only resort to violence and we will end up in the Hobbesian war of *"every man, against every man"* [15] and this would obviously be disadvantageous to all. Therefore all has an interest in having moral rules.

The assumption that morality is connected to our interests in some way or another, exactly how will be considered next, and the assumption that morality is needed to solve conflicts of interest among people are in my opinion very plausible ones. I think that these assumptions should be acceptable not only to contractarians but to most moral theorists and to

[13] ibid, p 148
[14] Baier, K. (1958). *The Moral Point of View*. Cornell University Press. p 190
[15] Hobbes. (1651). p 84

ordinary people as well for that matter. We agree then with Jan Narveson when he says *"a set of moral rules that those addressed have, simply, no possible interest in accepting, is a non-starter, a nonsense morality."* [16]

Reasons and motivation

We must also point out what we mean by "motivation" What we are looking for is something that can possibly be motivationally efficacious for each individual. What we are looking for are *reasons* to be moral, reasons that every rational person must accept. The person asking for such a reason is, I believe, asking for a "sound deliberative route" from her own "subjective motivational set." Borrowing these terms from Bernard Williams[17], what she is asking for are "internal reasons." For a person then, according to Williams, to have an (internal) reason for action is for that person to be able to start from something from which she already has some kind of motivation and through *deliberative reasoning* reach the conclusion that she has the reason in question. And this "something" to which we have to appeal cannot be something *external* (such as a divine authority or a "moral reality"). It must be a resource *within* each person. Williams concludes that there are no external reasons for action.

One could argue against this saying that is counter-intuitive that a man who treats his wife badly has no reason to treat her differently if there is nothing in his subjective motivational set that would be served by changing his ways[18]. There certainly is *something* counter-intuitive about this peculiar situation, but I think that the "counter-intuitiveness" stems from the fact that it is *unrealistic* that there is *nothing* in the man's motivational set that would be served by changing his behaviour towards the wife. The wife certainly has reason to leave him anyway, and perhaps even to call the police if it's a serious matter, if the man wouldn't want that he certainly has an *internal* reason to chance his ways.

I think this intuitive argument against reasons internalism is without merit. But what if we were persuaded by it? This would commit us to some difficult tasks. We would have to be able to explain the ontological status of external reasons; if reasons exist regardless of one's beliefs, desires and interests, then where do they come from? And we would also need to explain how we can get knowledge about these reasons. And further, and most importantly, explain how these reasons motivate us to action. If the externalist wants to show us how this is possible then he is committed to the view that facts could contain *within themselves* some

[16] Narveson, J. (2003) "The Contractarian Theory of Morals: FAQ." *Aufklärung und kritik*, Sonderheft 7/2003
[17] Williams, B. (1981) "Internal and External Reasons" in *Moral Luck*. Cambridge University Press
[18] Scanlon. (1998). pp 366- 367

kind of "normative authority". That is, that facts can be reason-giving (and motivating) all by themselves. Many attempts have been made to explain these things[19] and to refute them all would need another essay! Some of these attempts have included the idea that there are moral "properties" present in some actions and not in others and that we can *observe* these properties (perhaps with a special faculty) and that these observations can guide our actions. Or that some moral facts are just "self-evident."[20] What all this really amounts to is *intuitionism* about reasons and we have already expressed our worries about appealing to intuitions in moral theory. It is really difficult to say something about this without making a straw man out of our opponent. However, I think that we have established that the burden of proof lies on the reasons externalist and we follow Williams in that there are only internal reasons for action. We continue now with our positive account of finding (internal) reasons to be moral.

Deriving morality from rationality

All reasons to act are obviously not *moral* reasons. But we propose that moral reasons cannot be of a totally separate nature than all other reasons.[21] Moral reasons are simply "reasons" grounded in rationality and they motivate us in the same way that all other reasons do. There are, however important differences as hopefully will become clear below. For now it is enough to point out that these important differences are not, and cannot be, that moral reasons are something completely separate from other reasons. They are not "non-natural" or transcendental as some have claimed. Gauthier writes *"A person [...] considers what he can do, but initially draws no distinction between what he may and may not do"* He then asks *"How then does he come to acknowledge this distinction? How does a person come to recognize a moral dimension to choice?"* [22] We will return to this question below in our discussion of the relevance objection. We will argue that moral reasons must be reasons of rational- choice.

But why not just assume, like David Hume, that morality stems from our fellow-feelings and sympathy for one another or that our moral motivation is grounded in "sociability." Gauthier writes:

[19] See for example Shafer- Landau, R. (2003). *Moral Realism.* Oxford University Press. and Scanlon. (1998).
[20] Shafer- Landau. (2003). pp 247- 250
[21] Narveson, J. (1988). *The Libertarian Idea.* 2001. Broadview Press. p 126
[22] Gauthier. (1988). p 9

"One is not [...] able to escape morality by professing a lack of moral feeling or concern, or a lack of some other particular interest or attitude, because morality assumes no such affective basis. Hume believed the source of morality to lie in sympathetic transmission of our feelings from one person to another. But Kant, rightly, insisted that morality cannot depend on such particular psychological phenomena, however benevolent and humane their effect, and however universally they may be found"[23]

The argument here is from motivation. If morality is grounded on sympathy, we can not convince a person who lacks the particular feelings required to act morally. Moreover, different people have different feelings and attitudes as well as differently strong ones. We need to say two things about this, first that neither sociability nor "fellow-feeling" are the *foundations* of morals, and second that even though they are not the foundations these things are certainly not incompatible with Gauthier's account of morality. Living "in unity with our fellows" is certainly an important feature of human life for most of us but it is not the basis of moral rights and duties. "Unity" and sociability are, so to say only a bonus feature of morality and not its foundation. Gauthier writes:

"The contractarian need not claim that actual persons take no interest in their fellows; indeed, we suppose that some degree of sociability is characteristic of human beings. But the contractarian sees sociability as enriching human life; for him, it becomes a source of exploitation if it induces persons to acquiesce in institutions and practices that but for their-feelings would be costly to them. Feminist thought have surely made this, perhaps the core form of human exploitation, clear to us. Thus the contractarian insists that a society could not command the willing allegiance of a rational person if, without appealing to her feelings for others, it afforded her no expectation of net benefit"[24]

As noted above we want to argue for the importance of being able to answer the moral sceptic. Is it then possible to ground moral duties on something as instable as sympathy? We conclude, like Gauthier that it isn't; we must find something more stable and continuous. And since not everybody are "believers" in the same God, the general happiness, objective value or share the same moral intuitions, appealing to such things won't get us the result we want. So shortly, Kant was right about that morality must have something to do with *reason*. He

[23] ibid, p 103
[24] ibid, p 11

was only wrong about what reason *is*. Below we will take a look at different conceptions of rationality.

Rationality and moral motivation

So if we want to ground morality in rationality one promising strategy would be to find premises that the "moral sceptic" must accept, and then show that certain moral statements follow from those premises. In Gauthier's case the premises are not factual statements, or definitions of moral terms, but rather principles of individual rationality. And since Gauthier, as we will see below, identify rationality with individual utility-maximization the question "why be rational?" hardly makes any sense. Jan Narveson writes *"The "shouldness" of what is reasonable is not an accident, and not simply an additional factor pointing in favor of a given choice. That we should "follow reason" is true just because that's what reason talk is all about: what we should do and what reason tells us to do are not two different things"* [25] If then, the "should" of moral principles are not very different from the "should" of other non-moral rational principles that is, if these principles cannot be rejected and if moral principles can be shown to follow from them, then morality will have been provided with a suitable foundation. The moral sceptic will be hard put to reject such principles.

What is rationality?

We have already been talking a lot about "rationality" but what exactly is rationality? Gauthier makes an important distinction between two different conceptions of rationality, the *universalistic* and the *maximizing* conception. Gauthier identifies rationality with individual utility- maximization. Choosing rationally is to select the action that yields the outcome with greatest expected utility. And utility is a measure of individual preferences. This is the maximizing conception of rationality. Gauthier here follows David Hume in his often quoted: *"Reason is, and ought only to be the slave of the passions"* [26]

The universalistic conception, deriving from Kant and used by R.M Hare[27] among others, on the other hand is committed to the view that what makes it rational to satisfy an interest is independent of whose interest is it. The universalistically rational person thus seeks to satisfy all interests.[28] It is important to notice that the universalistic conception of reason already includes the moral dimension of impartiality that Gauthier seeks to generate.

[25] Narveson. (1988). p 126
[26] Hume, D. (1739) *A Treatise of Human Nature*, ii. iii. iii. 2003. Dover philosophical classics. p 295
[27] See Hare, R.M. (1981) *Moral Thinking*. Oxford University Press
[28] Gauthier. (1986). p 7

Connecting morality with rationality is therefore easily accomplished by proponents of the universalistic conception of practical reason.[29] Gauthier argues that the maximizing conception is almost universally accepted in the social sciences, economic-, decision- and in game theory. Therefore the onus of proof falls on those who defend universalistic rationality.[30]

My main argument against universalistic rationality is, again, that it runs into trouble with moral motivation. Gauthier writes *"on the universalistic conception all persons have in effect the same basis for rational choice – the interests of all – and this assumption, of the impersonality or impartiality of reason, demands defence"* [31] The question "why should I promote the interests of all?" seems to me to need an answer - while the question "why should I promote my own interests?" does not (and can perhaps not even be given a meaningful answer). Individual interest provides the basis for rational choice.

Here it is very important to point out what we mean by "interest" or "self- interest." On this point of the theory misunderstandings are very common. When we say "self- interest" we do *not* mean interests *in* the self but interests *of* the self, interests held by oneself as subject.[32] So our conception of rationality surely has nothing to do with egoism. Perhaps it is better to talk of "self-*perceived*- interest"[33] rather than "self- interest." We cannot, I think, have something like "purely selfless interests" for the interests we have must be *ours* to the extent that we are proper agents with motives. And these motives can, but must not, include altruistic ones. I think this is very plausible because it seems to me altogether unintelligible that things other than individual people can have interests. Inanimate objects and groups of people cannot have interests.

It is also important to point out that (even though our theory speaks of conditions for coherent and considered preference as conditions for rational preference) we do not address the *content* of the particular preference. We are not concerned with *the ends* of action; we leave that to the individual's preferences. The theory of rational choice treats practical reason as strictly instrumental.[34] Again we are in agreement with David Hume when he writes that it is *"not contrary to reason to prefer the destruction of the whole world to the scratching of my finger."* [35]

---

[29] See for example Kant and Hare
[30] Gauthier. (1986). p 8
[31] ibid, p 8
[32] ibid, p 7
[33] Lester, J. (2000). *Escape from Leviathan.* Macmillan Press. p 37
[34] Gauthier. (1986). p 25
[35] Hume. (1739). ii. iii. iii. p 296

The structure of contractariansim

A common objection to the idea of a rational morality is that it is very often advantageous to comply with the present moral in ones society even when this code differs from what we usually consider to be moral. Must we therefore say that the present moral code is rational? If we were to go along with this, contractarianism would lose much of its appeal. This objection is, however based on an equally common misunderstanding. The misunderstanding consists in the failure to recognize that we have rationality operating on two different levels here. The first is that we want principles that are rational and the second is that it should be rational to comply with these principles. Most agree that it would be rational to agree on certain rational moral principles. The problem is that it would often be rational to defect, that is fail to carry out that what was agreed. This is known as the compliance problem and this problem is subject to much dispute and much has been written about it. Gauthier claims to have an answer to it. We will not discuss his particular answer here but we stress the importance of distinguishing between the two "levels" of rational choice.

This two-level structure was already recognized by Hobbes in his "laws of nature". These "laws", he imagines, are *precepts of individual rationality, "found out by reason"*.[36] Hobbes imagines a pre-social, pre-moral state known as the "state of nature". We certainly don't need to say that the state of nature really has existed. It is often a very good way of understanding why we have something by imagining that we didn't have it. So if we want to know why we have morality, or why we want morality, it is a good way to imagine how it would be without morality.

In the state of nature there are no moral rules whatsoever, each person has the unlimited right to do whatever he can to preserve himself, but there are no obligations towards others, "*every man has the right to every thing; even to one another's body. And therefore, as long as this natural right of every man to every thing endureth, there can be no security to any man, (how strong or wise soever he be)...*" The word "right" here is confusing; perhaps it would be better to speak of an unlimited *freedom*. The idea is that rationality directs us to leave this horrible state. The fundamental law of nature tells us to seek peace and follow it where ever it may be found, and when it may not, by right of nature, to defend ourselves by all the means we can.[37] The second law, which Hobbes takes to be derivable from the first is that a man be willing, when others are so too, for the sake of peace, to lay down his right of nature (his freedom) to do all things *"and be contented with so much liberty against other*

[36] Hobbes. (1651). p 86
[37] ibid, p 87

men, as he would allow other men against himself." [38] Hobbes argues that as long as men do not lay down their right to all things, all men are in the condition of war. But if others did not lay down their right it would be *"no reason for any one, to divest himself of his: for that were to expose himself to pray […] rather than to dispose himself for peace."* [39]

It is in the introduction of the third law that things starts to become difficult. Even if it is advantageous to *make* agreements, or covenants, it does not follow that it is advantageous to *keep* these agreements. Hobbes himself, of course, recognized this and introduces therefore the third law that is *"that men perform their covenants made […] And in this law of nature"* he continues *"consisteth the fountain an original of JUSTICE."* [40] The problem is that even if each individual maximizes her expected utility in making an agreement she does not (always) maximize her expected utility in *complying* with this agreement. This opens for the objection of "the Foole." The Foole accepts the first two laws of nature, but questions the third. The Foole asks whether *"reason, which dictateth to every man his own good"* [41] could not sometimes call for non-compliance. He questions why one should keep one's covenants in situations where it would be advantageous to break them. Gauthier says *"The Foole challenges the heart of the connection between reason and morals that both Hobbes and we seek to establish – the rationality of accepting a moral constraint on the direct pursuit of one's greatest utility."* [42] As we will see, Hobbes and Gauthier solve this problem rather differently however.

Prisoner's dilemma

Let's take a practical example.[43] If I have an apple and you have a banana and I would rather have your banana and you would rather have my apple. A trade seems convenient. But most of all *both* you *and* I want to have *both* the apple *and* the banana. The best possible outcome for me (I get to have both fruits) would be the worst possible outcome for you (You having none of the fruits.) And so the best for you would be the worst for me. The second-best outcome would be the trade (I get your banana and you get my apple) and the second- worst outcome would be the status-quo (I get to keep my apple even though I would rather have

[38] ibid
[39] ibid
[40] ibid, p 95
[41] ibid, p 96
[42] Gauthier. (1986). p 161
[43] A similar example can be found in Narveson. (1988). p 137- 138 (He speaks of fry pans and dollars where I speak of apples and bananas)

your banana and you get to keep your banana even though you would rather have my apple) This kind of situation can be depicted as a so called "prisoner's dilemma." [44]

The dilemma is then whether I should choose the *co-operative strategy* (go along with the trade) or if I should choose the *non-co-operative strategy*. If the other person chooses to co-operate (letting me have the banana) I could either co-operate too (giving away my apple) or choose the non-co-operative strategy. The first option will bring us the second-best outcome and the latter option will bring my best outcome and hence the other person's worst. The other person, of course, reasons in the same way. The "traditional" "solution" to the dilemma is that the rational strategy is the non-co-operative since regardless of what the other person chooses this strategy will have me better off.

Hobbes's solution involves the notion of a "sovereign." (I.e. I get arrested if I run away with your banana without giving you my apple) This feature makes Hobbes's solution a *political* one rather than a *moral* one. In fact the political solution makes morality unnecessary.[45] There are further problems with the sovereign however. Sovereigns are *costly*! Gauthier says *"those subject to the Hobbesian sovereign do not, in fact, attain an optimal outcome; each pays a portion of the costs needed to enforce adherence to agreements, and these costs render the outcome sub-optimal."* [46] Gauthier also says that if the free market acts as an invisible hand, the sovereign acts as a very "visible foot", *"directing, by well-placed kicks, the efforts of each to the same social end."* [47]

Gauthier has a different solution. As I see it, Gauthier's reasoning starts with a presumption that there must be something awkward with what we called the traditional solution to the prisoner's dilemma. If rationality consists in maximizing one's utility it seems really strange that rational persons none the less fail to bring about the co-operative outcome that would be better for both. Instead of "defecting" in prisoner's dilemma-games, Gauthier thinks that the rational person adopts a *disposition* to co-operate. His solution involves seeing the rationality of *dispositions to choose* rather than rationality of individual choices. Gauthier's argument identifies practical rationality with utility-maximization at the level of dispositions to choose.[48] This is also what we need to say to the Foole. Remember the person questioning Hobbes third law of nature, that of compliance. We need to say to him that it is rational to perform one's covenant even when such performance is not directly to one's

[44] A more detailed explanation of the dilemma can be found in Gauthier. (1986). p 79- 82
[45] ibid, p 163- 164
[46] ibid, p 164
[47] ibid, p 163
[48] ibid, p 187

benefit. Given that the disposition to perform is to one's benefit. We must also say that the disposition to decide whether or not to comply with one's rationally made agreements by appealing directly to utility-maximization is itself disadvantageous because it excludes one from participating in highly beneficial co-operative arrangements. The disposition to keep one's agreements on the other hand makes one an eligible partner in beneficial co-operation, and so it is itself advantageous. That is, gives than one's disposition can be known, or sufficiently suspected.

This has all been quite sketchy and I do not intend to give the complete picture. Gauthier's solution is neither uncomplicated nor uncontroversial. We will not, however, further discuss the numerous objections to it here. Much has been written about this[49] and I really have nothing to add to those discussions. We will instead move on to consider an objection to our project for which we are now ready.

The relevance objection

Gauthier's project is in considerable part to derive morality from rationality, which is deriving morality from *non-moral* premises, as we have seen; this (if it is successful) has several theoretical advantages. Gauthier himself describes his project thus:

"...we shall exemplify normative theory by sketching the theory of rational choice. Indeed we shall do more. We shall develop a theory of morals as part of the theory of rational choice. We shall argue that the rational principles for making choices, or decisions among possible actions, include some that constrain the actor pursuing his own interest in an impartial way. These we identify as moral principles"[50]

Objections have been raised against this identification, however, and this is the "relevance objection". David Copp describes the objection thus:

"The issue here is not a verbal one, nor is it purely technical. It is whether the contractarian has anything to say to the sceptic about the rational credentials of morality; it is whether the topic is still morality. Perhaps Gauthier's argument succeeds in justifying certain

[49] see Sayre-McCord, G. (1991). "Deception and Reasons to be Moral" and Copp, D. (1991). "Contractarianism and Moral Scepticism" and Smith, H. (1991). "Deriving Morality from Rationality" and Danielson, P. (1991). "Closing the Compliance Dilemma: How it's Rational to be Moral in a Lamarckian World" all in Vallentyne, P. (Ed). *Contractarianism and Rational Choice.* 1991. Cambridge University Press

[50] Gauthier. (1986). p 2

requirements of rational choice, such as to maximize their opportunities for making advantageous agreements. Yet he still needs to show that these are moral requirements. This is the relevance objection"[51]

This is a serious objection to the contractarian project. What is under attack here is not Gauthier's *conclusions* (they may, or may not, be correct) but the *object* and the *starting premises* of his whole project, basically the idea of a rational morality and the idea of developing a moral theory within the framework of rational choice. We assume now that Gauthier has shown that it is rational both to adopt constrained maximization in many cases and then to carry it out. Would the success of this argument show, as Gauthier believes it does, that morality is founded on rationality, and hence that rationality provides a justificatory framework for moral behaviour and principles? Holly Smith argues that *"if the success of Gauthier's argument would provide a rational justification for morality, then we would be well repaid to tinker with the details of his argument in an attempt to salvage it from my previous criticisms. But if success would not provide such a justification, then such tinkering has little or no point"* [52] We agree with Smith that *if* success in Gauthier's argument would still not provide a justification of morality, then tinkering with the details of the arguments would, indeed, have little or no point. But we will instead argue that success in Gauthier's argument *would* provide a justification of morality and we will try to show that (some of the) the arguments Smith and Copp offer are mistaken. Smith puts the relevance objection slightly different:

"We may characterize what Gauthier has done as arguing that individual rationality, or self-interest, requires a person to dispose herself to perform certain cooperative acts, and then actually to perform those acts when the time comes. Suppose we assume that the acts in question are precisely the same ones as morality requires. Still, the success of this argument would not show that *morality* has been provided with a justification. It would show that we have self-interested reasons to do what morality, *if it were true* (or correct), would demand – but it would not show that morality *is* true (or correct). Such an argument would merely show an interesting coincidence between the purported claims of morality and the real claims of self-interest" [53]

[51] Copp. (1991). p 208
[52] Smith. (1991). p 249
[53] ibid, pp 249-50

Smith seems to presuppose that there are something like "the purported claims of morality" *that are separate from* "the real claims of self-interest" This needs, I believe, defence and here we simply deny the existence of such separate claims and stick with the *real* claims of each individuals reason. There are only internal reasons, remember. As we argued above there is no fundamental difference between moral reasons and other kinds of reasons in the respect that they are both grounded in rationality. But there are however important differences between them and a large part of the remainder of this essay will be about pointing them out and defending them. David Copp writes *"In order to answer the relevance objection, one would need a theory as to the nature of moral codes. Gauthier does not have such a theory..."* [54] In answering the relevance objection, then, we will sketch such a theory.

What is morality?

All rational principles are clearly not moral principles. There has to be something more there to make a principle a moral principle. But precisely what? Gauthier claims that the traditional conception of morality identifies any *impartial* constraint on self-interested behaviour as moral. Holly Smith argues that impartial constraint is *not* sufficient to show that a principle is moral. She writes: *"Consider a rule of etiquette requiring thank-you notes to be handwritten rather than typed. Such a rule is certainly an impartial constraint, but it does not thereby qualify as a moral principle"* [55] She then argues that what is lacking in this principle is *"something like appropriate deontic force"* [56] She suggests these three (additional) criteria for moral principles to qualify as moral principles:

a.) They must be overriding
b.) They must be categorical
c.) They must support attitudes of guilt and blame

In a footnote Smith suggests a fourth criterion; that of "appropriate content" she argues:

"Another example is the principle of malevolence, which prescribes any action maximizing the general unhappiness. This principle is both impartial and a constraint on self-interest

[54] Copp. (1991). p 208
[55] Smith. (1991). p 251
[56] ibid

(since it is often highly damaging to an agent's own interests to follow it), yet it hardly seems to be a moral principle. Other examples are supplied by club rules, legal codes, mafia codes of honour, professional codes, administrative regulations, etc. Of course, some of the prescriptions stemming from these sources will require acts that are morally right, but it does not follow that all such prescriptions coincide with morality, or that any of them *in itself* constitutes a moral prescription. The difficulty with the principle of malevolence is not the one I cite in the text – inappropriate deontic force – but rather inappropriate content"[57]

This criterion of "appropriate content", however, we will have to reject. The principle of malevolence may be both impartial and a constraint on self-interest but it will hardly pass the test as a rational principle as Smith herself points out. So this principle then constitutes no objection to Gauthier's account of moral principles. I think, on the other hand, that is very reasonable indeed to think of morality as analogous to club rules or legal codes. "Mafia codes of honour" on the other hand are seldom both rational and impartial.

Smith then asks *"whether Principle 2* [principle 2: If the agent rationally forms the intention to cooperate, then it is rational for her to carry out this intention (assuming she has acquired no new information and has not altered her values)] *has the kind of deontic force required of genuine moral principles. Is its recommendation overriding and categorical, and does it support attitudes of guilt and blame, etc.?"* [58] This we will try to show. We will deal with all her suggested criteria.

How contractarian principles are overriding and categorical

Even though we argue that the interests of men are the foundations of morals, morality have to be able to *override* individual interest. Why? Gauthier quotes Hume's question of *"what theory of morals can ever serve any useful purpose, unless it can show that all the duties it recommends are also the true interest of each individual"* and concludes that Hume seems to be mistaken because *"such a theory would be too useful. Were duty no more than interest, morals would be superfluous"* [59] But he also says that Hume's mistake in insisting that moral duties must be the true interest of each individual conceals a fundamental insight. *"Practical reason is linked to interest, or, as we shall come to say, to individual utility, and rational*

[57] ibid
[58] ibid, p 252 (By "Principle 2" she means the second of two principles that are implicit in Gauthier's theory. They are: (1.) under circumstances C, it is rational for an agent to adopt constrained maximization and form the intention to cooperate. (2.) If the agent rationally forms the intention to cooperate, then it is rational for her to carry out this intention (assuming she has acquired no new information and has not altered her values).
[59] Gauthier, (1986). p 1

constraints on the pursuit of interest have themselves a foundation in the interests they constrain" [60] This may seem as we are trying to catch the best of two worlds – and perhaps - we are! Remember Gauthier's answer to the Foole above. We shifted our focus from the rationality of particular actions to dispositions to choose. Even if our disposition forces us to make disadvantageous choices, the disposition itself is advantageous. The idea of a *moral* disposition somehow involves the idea of "overridingness." The moral disposition requires of an agent the she is able to let moral principles take precedence over other rational principles.

Holly Smith seems to agree that contractarian principles can be both overriding and categorical – *in one sense*. She writes *"They are overriding because they always outweigh the recommendations of self-interest to maximize one's own utility, and they are categorical for the same reason: they tell the agent what to do regardless of her desires at the moment of action."* But she continues *"On the other hand, there is a sense in which they are neither overriding nor categorical, since these prescriptions only arise because of the agent's prior attempts to satisfy her desires and maximize her self-interest by adopting constrained maximization. This is not the kind of independence from desires that Kant, for example, had in mind"* [61] Gauthier would surely agree with this which is indicated by this quote *"The Kantian ideal of a pure reason which is practical despite its utter indifference to passion is entirely foreign to our argument. The imperatives of reason remain assertoric. And they remain imperatives of individual reason."* [62] So we thereby avoid all resort to something "trancendental" which I consider a great theoretical advantage. It would be question begging to require the principles to be categorical in the stronger Kantian sense.

I think we have answered Smith's challenge to show that contractarian principles can be overriding and *sufficiently* "categorical" but we can do more than that! The contractarian doesn't just *assume* that moral principles "must be" overriding and categorical. We can even give independent reasons *why* they must be so. As we argued above it is very reasonable to consider morality to be a tool to resolve conflicts of interest. Morality overrides "inclination" in the sense that it is intended to settle conflicts of interest, which it could not do if it couldn't overrule one or the other or both parties. But it doesn't, as Narveson writes, *"override "inclination" in any sense that envisages a contrast between mere "sensuous" motives and any others one could imagine"* He continues *"The trancendental aspirations of rival religious groups, or for that matter rival metaphysicians, can give rise to wars or other disputes that*

[60] ibid, p 2
[61] Smith. (1991). p 252
[62] Gauthier. (1986). p 237

require conflict resolution just as much as any more "sensuous" motives. Indeed, quite a lot more if history is our guide: really major fracases, such as the world wars, the Napoleonic ones, the Thirty Years' War, and so on, are invariably fought with strong ideological motivation."[63] We conclude that principles generated by "morals by agreement" not only *can* be overriding and sufficiently categorical. We have also an explanation to why they must be so. If morality was unable to override self-interest it would not fill the desired function. Morality would be useless if it was unable to override individual advantage.

How a contractarian morality support attitudes of guilt and blame

Holly Smith thinks that it is…

"far less clear that that one can argue that prescriptions generated by Principle 2 appropriately support attitudes of blame, guilt and so forth. I myself see no reason why an agent should feel guilty (or blame others) for violating Principle 2, which is essentially a demand that one's actions and intentions show a certain form of consistency. Inconsistency in not usually the object of blame and guilt" [64]

Our argument here goes something like this: If people in general did not follow morality that would be a *dis*advantage to all. If this is a disadvantage to all this will lead people in general to *blame* people that break the moral rules. And when you yourself perform an act that you know will result in people blaming you, you feel *guilty*. Is this argument any good?

If it is true that it is rational to dispose oneself to keep agreements it is perhaps also rational to "wire" oneself in such a way as to be disposed to cooperate with others similarly disposed as well as disposing oneself to attempt to *persuade others* to adopt the cooperating disposition? "Wiring" oneself to persuading others could reasonably include attitudes of blame towards those who break the moral rules. Narveson writes:

"There is another, and crucial, sense in which morality is general which in fact can be derived from the two aspects of overridingness and generality that have been brought out so far. This is the aspect of "enforcement". Morality is (to be) enforced. Not in the sense in which the Law of the Land is enforced, with specially appointed enforcers, the "moral police", but rather by

[63] Narveson. (1988). p 128
[64] Smith. (1991). p 252

what I shall call "informal" means. Verbal means are pre-eminent among these: we shout at miscreants, we prod and natter and nag both ourselves and others."[65]

This view can be reinforced by the arguments found in Robert Sugden's "The Economics of Rights, Cooperation and Welfare" where he argues that "conventions" often tend to become "norms". He follows Hume in suggesting that "natural laws" can come to have moral force on us. He does not argue that we *ought* to behave according to "natural law" he argues only that we tend to believe that we ought to. And this is exactly what Holly Smith seems to doubt.

Sugden argues that "conventions" are maintained not only by the interest each individual has in keeping to them but also by the *expectations* of other people. To make this clear let's apply Sugden's thinking to our apple-banana- example from above. Suppose you *want* me to perform some action X, say, give you my apple. You also have a confident *expectation*, based on your experience on other people's behaviour in similar circumstances, that I *will* give you my apple. In the event I do something else, i.e. run away with your banana, leaving you worse off than you had expected to be. Then you would probably feel some resentment against me. And in order to explain this sense of resentment, Sugden suggests that it is enough that (1) you had expected me to do X (2) other people, in my situation, would have done X; (3) my not doing X has hurt you. In these circumstances resentment is a "primitive human response." He continues:

"It is another natural human response to feel uneasy about being the focus of another person's resentment. Because of this, our actions – and our evaluations of our actions – are influenced by other people's expectations of us."[66]

And further:

"We expect that our dealings with other people will be regulated by convention, but this expectation is more than a judgement of fact: we feel *entitled* to expect others to follow conventions when they deal with us, and we recognize that they are entitled to expect the same of us. In other words, conventions are often also norms…"[67]

[65] Narveson. (1988). p 125
[66] Sugden, R. (1986). *The Economics of Right, Cooperation and Welfare*. Palgrave Macmillan. 2004. p 154
[67] Ibid

But why think that other peoples expectations matter? Especially why think that the expectations of strangers, that it is unlikely that we will ever meet again, matter? Sugden uses a real-life example to persuade us of this.[68] Suppose you take a taxi ride. You reach your destination safely and you will probably never have any further dealings with this particular taxi-driver ever again. And even if you did he would be unlikely to remember you anyway. Why give him a tip? It is not in your interest to do so given these circumstances. Nevertheless people often do. And if they don't tip they often experience "sensations of unease and guilt." Sugden then suggests that the answer lies in the fact that we know that the taxi-driver wants us to tip him, we know that he expects us to tip him and we know that he knows that we know that he expects this.

So it's a matter of observation that we often are motivated by what we take to be other people's expectations about us. And this even when those other people are total strangers, and when there seems to be "no solid reason" for us to care about their opinions of us.[69] Further we also care about the opinions of third parties – *"people with no direct interest in the game, but who happen to observe it, or who are told about it afterwards."* [70] And also when we ourselves are "bystanders", witnesses to, but not ourselves involved in "games" we often react with resentment against those who breach established conventions. Again we can use our good-old example with the apple-banana- trade. Suppose you are not party to the trade but you happen to pass by and observe that one of the parties runs away with both fruits after having made a deal to trade. Even if we are not ourselves affected by the trade we have a tendency to *blame* such behaviour.

Sugden draws the conclusion that other peoples expectations of us *do* matter. They matter because we care what other people think of us. Sugden also says that the *"desire to keep the good will of others – not merely of our friends, but even of strangers"* seems to be a *"basic human desire"* and further that this desire is presumably the product of biological evolution. We are social animals and Sugden says that some in-built tendency to accommodate oneself to others *"must surely be an aid to survival for a social animal."* [71]

Earlier in this essay I argued *against* the ideas that the foundations of morals lie in our sympathetic feelings for each other or that they lay in a shared ideal of "sociability." This may seem to clash with our present argument. It might seem that, in admitting that persons do come to take an interest in their fellows, we undermine the rationale for requiring that moral

[68] Ibid, p 155
[69] Ibid, p 154
[70] Ibid, p 156
[71] Ibid

constraints have a "non-tuistic" basis. But, as Gauthier says *"…an affective capacity for morality does not give rise to moral constraints but presupposes their prior recognition; the desire to do one's duty cannot determine the content of duty."* [72] And if moral constraints underlie "tuistic" values, then they must have a basis independent of those values.

So we can therefore hold on to our rational, "non-tuistic" foundations of morality. We argue that persons *come to take* an interest in their fellows *after* they have recognized the rational basis for morality. They come to take an interest in participation because they recognize their mutual willingness not to take advantage of each other. Gauthier again *"In accepting moral constraints they do not express their concern for each other, but rather they bring about the conditions that foster such concern."* [73] He continues:

"A rational morality is contractarian. But this does not imply that it is of purely instrumental value to us. In relating morality to the provision of benefits that themselves involve no affective concern with others, we do not thereby impoverish the moral feelings of persons who have such concern. It is because we can give morality a rational basis that we can secure its affective hold" [74]

Answering the relevance objection

In answering the relevance objection then, the contractarian first reminds us what morals is for. As we have pointed out above morals should be able to resolve conflicts of interest. If there is no conflict, there is no place for morality. If we look at Holly Smith's example *"consider a rule of etiquette requiring thank-you notes to be handwritten rather than typed"* We agree with Smith's intuition that this hardly seems to be a *moral* principle. But the reason it isn't is that there is no real *conflict* to "solve" here. You cannot just pick any old principle you like and proclaim it a moral one, *"Nor,"* as Narveson says *"despite the "internal" aspects of morals, can we derive the content of morality only by looking within our own soul. There are others to worry about."* [75] Morals is for the governance of *everyone*.

I believe that we have shown, not only that contractarian moral principles are overriding, (sufficiently) categorical, and that they can support attitudes of guilt and blame, but we have also given these requirements independent support. That is, if moral principles did not have these mentioned features they would not perform the desired functions.

---

[72] Gauthier. (1986). p 338
[73] ibid
[74] ibid, p 339
[75] Narveson. (1988). p 125

Is there anything left of the relevance objection? It is still possible to object that even though we have shown that there are *rational*, principles that are also *overriding*, (sufficiently) *categorical*, and that they can support attitudes of *guilt and blame* but that we *still* have not shown that these are *moral* principles. This, I believe, reduces the relevance objection to a mere verbal matter. But perhaps not, maybe we need to say something more. If there are any "moral" principles that differ from our conception of moral principles, that is, principles other than our rational-overriding-categorical-principles it is very unclear why anyone should accept these principles. In short, if we picture moral principles as something else than principles of rational choice we cannot give reasons that all rational persons must accept to follow these principles and hence we cannot answer the question of moral motivation.

Conclusion

I have argued that any plausible moral theory should be able to provide a good answer to the question "why should I be moral?" and that a rational-choice based theory such as David Gauthier's contractarian moral theory provides a plausible answer to this question. Such an answer must take the form of a justification of morality. A justification that starts from premises that the "moral sceptic" must accept, and then shows that certain moral statements follow from those premises. This justification must be prudential since the premises are principles of individual instrumental rationality rather than factual statements about the world e.g. some "moral reality" or people's intuitions. If these principles cannot be rejected and if moral principles follow from them, then morality will have been provided with a justification that even the fully rational moral sceptic must accept. And since we accept a conception of rationality that identifies rational-choice with individual utility-maximization the question "why be rational?" doesn't makes any sense. If then, the "should" of moral principles are not very different from the "should" of other non-moral rational principles that would establish moral principles on a firm foundation. If on the other hand moral reasons would be of a wholly separate nature e.g. "non- natural" or "transcendental" then there would be no reason why a fully rational person could not ignore these reasons.

I think that this is the way a rational justification of morality will have to take. A great theoretical advantage with this kind of justification that we have been advocating is that it doesn't need to appeal to anything transcendental or metaphysical. Neither do we need to appeal to facts about people's moral intuitions nor such particular psychological phenomena as sympathy or ideals of sociability. The only thing we need to appeal to is the most widely

accepted and most plausible conception of rationality, the one that identifies rational- choice with individual utility- maximization, where "utility" is the measure of individual preferences. This provides us with a rational justification of morality that even the "moral sceptic" must accept.

We have also tried to answer the powerful "relevance objection", that says that even if there are rational principles for choice these are never *moral* principles. We have tried to answer this by sketching a theory of the nature of moral codes. We have argued that contractarian moral principles are overriding, (sufficiently) categorical, and that they can support attitudes of guilt and blame and we have provided independent support for these requirements. That is, if moral principles were not overriding they would not perform the functions we want them to. If moral principles would not be able to override individual self-interest of one or the other of two conflicting parties it would not help us to solve the conflict. And if it did not that would render morality meaningless.

If we have been successful in showing that there are rational principles that are also overriding, (sufficiently) categorical, and that they can support attitudes of guilt and blame, then we have "uncovered" the relevance objection and reduced it to a mere verbal matter. If there are any principles (epistemically true moral principles perhaps?) other than the kind of principles we have been picturing (rational-overriding-categorical-guilt/blame- supporting-principles) it is very unclear why anyone should accept these other principles.

These features of the theory (prudential justification, maximizing conception of rationality, instrumental conception of rationality, overridingness etc.) are crucial for our project in answering the important question of moral motivation. Rational- choice contractarianism is the only theory I know of that meet all of these criteria.