Fritz- Anton Fritzson
FPR 303
University of Lund

Is a fundamental justification of morality possible?

Tutor: Wlodek Rabinowicz

Introduction

Does morality need justification? And if yes, what kind of justification will do the proper job? Peter Danielson distinguishes between *fundamental* and *non-fundamental* justification. A fundamental justification is a justification of a realm that does not appeal to any of the concepts of that realm.[1] In the case of morality this means a justification formulated in *a*moral or *non*-moral terms.[2] A non-fundamental justification of morality, on the other hand, assumes some moral premise(s), and therefore assumes what is to be proven. Thus, a non-fundamental justification is really no justification at all. Or as Danielson puts it, non-fundamental justification *"begs the central question of ethical theory"*. He continues:

> I realize that many philosophers think that begging this question is necessary. It may be that ethics is not possible unless one assumes the autonomy of the field. Or it may be that some moral premise is deeply true of human beings. Perhaps, but notice that both of these methodological moves make strong claims and should be seen to do so. I am inclined to make do with weaker claims, to seek a fundamental justification by reducing morality to something simpler and clearer. An obvious candidate is amoral instrumental rationality.[3]

The idea of fundamental justification has great philosophical appeal. But that this kind of justification is theoretically superior to non-fundamental justification would, of course, be uninteresting if fundamental justification were impossible. The best way to argue for fundamental justification is therefore to develop a positive account of such a justification and thus show that ethics *is* possible without assuming the autonomy of the field. This is the path that Danielson takes.[4] Another, probably the most well-known, writer who takes this route is David Gauthier. In his book *Morals by Agreement*, Gauthier is interested in just the kind of justification of morality that Danielson mentions above as an "obvious candidate", namely a justification formulated in terms of non-moral instrumental rationality. Even if Gauthier does not use the term fundamental justification, it is clear that he is trying to provide morality with such a justification. This is, for example, indicated by the claims that morality "*can be*

[1] Danielson (1992) p 19. Danielson is inspired by Robert Nozick on this point. See Nozick (1974) pp 6-9.

[2] I prefer to use the term "non-moral" instead of "amoral" because of the negative connotations that the latter term brings.

[3] Danielson (1992) pp 19-20. Another term for this kind of justification would therefore be "reductive justification".

[4] See Danielson (1992).

generated as a rational constraint from the non-moral premises of rational choice."[5] And, *"…we claim to generate morality as a set of rational principles for choice. We are committed to showing why an individual, reasoning from non-moral premises, would accept the constraints of morality on his choices."*[6]

I intend to follow Gauthier and Danielson in their general aim of finding a fundamental justification of morality based on instrumental rationality. The strategy is to seek to establish a connection between the rational and the moral or to "*reduce (some instances of) the question: why be moral? to questions about rationality*."[7] I will try to provide a rough sketch of what a successful justification of this type would look like. The starting point for my discussions will often be Gauthier's theory and the numerous criticisms launched against it. Since my intention is not to form a complete contractarian moral theory, I will touch only upon those issues that are directly relevant to fundamental justification. Contractarianism in general and Gauthier's theory in particular may have many other problems that I cannot discuss in detail here.

Why fundamental justification?

As I see it, there are at least three, more or less conclusive, reasons to pursue a fundamental justification of morality. I have already mentioned one such reason, that fundamental justification has great philosophical appeal. Danielson says, simply, that the easiest way to argue for fundamental justification is to consider the alternatives. Further, that anything short of a fundamental justification begs the central question of ethical theory.[8] Robert Nozick says, about the political realm, that a fundamental *explanation* of this realm, to fully explain the political it in terms of the non-political, "*stands as the most desirable theoretical alternative, to be abandoned only if known to be impossible*."[9] I am here exploring the consequences of saying, about moral justification, what Nozick says about political explanation.

The second reason is that there is, hopefully, no need for us to assume anything metaphysical, theological, transcendental, or the like, in our justification of morality. Neither do we have to rely on people's moral intuitions. This, I believe, should be seen as a great advantage. Our project is to justify morality, as a set of interpersonal rules for behaviour, in

[5] Gauthier (1986) p 4

[6] Gauthier (1986) p 5.

[7] Danielson (1992) p 28.

[8] Danielson (1992) p 19.

[9] Nozick (1974) p 6.

terms of something like an agreement, rationally acceptable from the point of view of each individual.[10]

The third reason is about motivation. Every satisfactory ethical theory should be able to answer the question “why be moral?” in a non-questionbegging way. We shouldn’t be happy with just being able to reach the ones who *want* to be moral. A fundamental justification wants to justify moral constrains to naturally unconstrained agents. In answering this question a fundamental justification must appeal to something to which a person must already be committed, individual interest. In this way a fundamental justification can account for motivation.

This line of reasoning in turn relies on the very plausible assumption that morality is a human-made institution, justified only to the extent that it furthers the interests of those bound by it. To generalise this a bit, we could point out that people have interests, and it is not hard to imagine that people often 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”[11]

The idea is that all parties to a conflict would be worse off without such rules. Without moral rules we could only resort to violence and end up in the Hobbesian war of “every man, against every man.”[12] This would be disadvantageous to all. Therefore, all have an interest in having moral rules. These assumptions, that morality is connected to our interests, in some way or another, and that morality is needed to solve conflicts of interest among people, are in my opinion very plausible. I think they should be acceptable, not only to contractarians, but to most moral theorists and to ordinary people as well, for that matter. Morality is not something that is “from above” and completely separate from the interests of human beings. Morality simply has to have something to do with our interests. Or as David Hume put it: *“What theory of morals can ever serve any usefull purpose, unless it can show that all the duties it*

[10] At least from the point of view of those whose preferences makes them welcome participants in society.
[11] Baier (1958) p 190.
[12] Hobbes (1651) p 84.

recommends are also the true interest of each individual?"[13] Hume's words should not, however, be taken literally, since morality must be able to override individual interest in order to be able to solve conflicts of interest. That morality must be able to overrule the very interests upon which it rests may seem like a paradox. It is not, however. The rational choice contractarian argues that, *"...rational constraints on the pursuit of interest have themselves a foundation in the interest they constrain."*[14] Morality overrules advantage but the acceptance of morality is itself advantageous. Even though this idea does not give rise to any paradox, it does give rise to some problems. I will discuss some of these below. But first I consider some external objections to the project of fundamental justification.

Contractarianism

The idea to give morality a justification in terms of rationality is closely linked to *contractarianism*. A contractarian justification of morality proceeds by showing that under certain specified conditions, rational agents would agree, or have agreed, to certain principles. More generally, the contractarian puts rational agents in a preferred situation and argues that their choice of some principle, practice, institution or social structure, in that situation, provides a justification of this principle, practice, institution or social structure.[15] What would the "preferred situation" look like in order to provide morality with a justification? Different contractarians answer this question differently and it may here be useful to sort out some important differences.

One important distinction, related to the one between fundamental and non-fundamental justification, is the distinction between *weak* and *strong* contractarianism.[16] Weak contractarians are working *within* morality, beginning with prior moral constraints and deriving the principles of morality from there. John Rawls, for example, uses the theory of rational choice to derive moral principles from a *morally loaded* choice situation.[17] Rawls sets up his original position to screen off morally irrelevant features such as the parties' talents and capacities. The "veil of ignorance" in Rawls' theory severely restricts the knowledge of the parties as to circumstances and capacities. Weak contractarians, by definition, cannot provide morality with a fundamental justification. They are not aiming to justify the moral realm

[13] Hume (1751) p 280.
[14] Gauthier (1986) p 2.
[15] Danielson (1992) p 25 and Vallentyne in Vallentyne (1991) p 3.
[16] Danielson (1992) pp 25-26.
[17] Vallentyne in Vallentyne (1991) p 2.

without appealing to any of the concepts of that same realm. Rawls' theory is therefore expository rather than justificatory.

Strong contractarians, on the other hand, argue from premises of non-moral individual rational choice. In this camp we find Gauthier and Thomas Hobbes[18] among others. Strong contractarians claim that any plausible moral theory should be able to reach normative conclusions without introducing prior moral assumptions. I therefore choose to work within the strong variant of the theory.

The distinction between weak and strong contractarianism should not, however, be confused with the distinction between *actual* and *hypothetical* contractarianism. Both Gauthier and Rawls, like most other contemporary contractarians, base their theories on hypothetical agreement. Rawls' theory is, however, in one sense "more hypothetical" than Gauthier's. Even though the latter sees agreement as hypothetical, in supposing a pre-moral context for the adoption of moral rules and practices, the parties to the agreement in Gauthier's theory are real, determinate individuals, characterised by their capacities, situations, and concerns.[19] Unlike Rawls, Gauthier imposes no veil of ignorance. The contracting parties have full knowledge of their capacities and interests. This point is important for the question whether the resulting morality has any motivational force. If the parties to the "agreement" in the contract situation were not sufficiently similar to the real people in this world, it is unclear why the moral principles generated by contractarianism would apply to us real people in this world. Why should anyone care to follow principles derived from a contract situation with no obvious connection with our own? Gauthier, I take it, has a better answer, since in his contract situation the parties to the agreement are much more like real people than the parties behind Rawls' veil of ignorance.

Intuitions in theory

One objection to our project might be that the objective of the moral philosopher is to develop a theory that "fits the world", a theory that makes sense of our moral phenomenology, or systematises our considered moral judgements into a coherent whole. One such idea is the method of "reflective equilibrium". According to this method, used by Rawls, we should start with our considered moral judgements and then build a theory that fits with these pre-theoretical judgements. In other words, the theorist of reflective equilibrium allows *initial*

[18] It is important to note the distinction between Hobbes' moral theory and his political theory. Here we only have his moral contractarianism in mind. I think that Hobbes' political theory is a wrong application of his moral theory.

[19] Gauthier (1986) p 9.

weight to our considered moral judgements. This method is not open to the strong contractarian who wants to provide morality with a fundamental justification. Gauthier 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.[20]

It must be pointed out here that a rational-choice justification of morality is not committed to the claim that it does not rely on *any* intuitions whatsoever. I think that a fundamental justification also must rely on some kind of intuitions. But in order for it to be fundamental it cannot rely on any *moral* intuitions. If it could be shown to be dependent on moral intuitions, it is no longer a fundamental justification. But what if our moral intuitions are stronger than our intuitions about rational choice?[21] They may very well be. Most people probably don't even have any intuitions about rational choice. This objection however begs the question against fundamental justification. The objector seems to presuppose that intuitions (moral and non-moral alike) should be weighed against each other and the "strength" of our individual intuitions determines which ones will "win". This is however exactly what fundamental justification denies. Allowing initial weight to *moral* intuitions (regardless of strength) in ethical theory would put us *within* the realm that we are trying to give an independent justification of, and thus would give us no justification at all. The argument of fundamental justification is *not* that some particular intuitions about rational choice are "stronger" than any of the moral intuitions we might have. It is rather that *moral* intuitions should be kept separate from other kinds of intuitions and that only the latter should be allowed in constructing ethical theory. My theoretical intuitions tell me that it would surely be better if we could build a moral theory without appealing to any moral intuitions.

A common view seems to be that if one can't do without intuitions altogether, one can appeal to as many as one likes. I don't think this way of arguing is very convincing. One major problem with appealing to intuitions in theory is that when intuitions conflict we have no further method for resolution. But there are further problems.

[20] Gauthier (1986) p 269.
[21] I thank Wlodek Rabinowicz for this objection.

To appeal to intuitions 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 person asking for reasons to be moral would be unmoved by our attempts. As long as this "sceptic" refuses to accept moral facts, it is open for him or her to say "so what?" I don't think that a person asking for reasons to be moral is asking us to point to some observational facts or some right- or wrong-making "properties" (natural or non-natural). What we are looking for then is not an epistemic justification. Someone asking for reasons to be moral is rather, I think, asking if and how, the ends to which the individual in question already is committed, can be promoted by adhering to morality. We are being asked for a *deliberative* justification of morality. To provide such a thing, we must find reasons to be moral that every rational person must accept. I think we would have to appeal to something to which the individual in question already is committed, something that is motivationally efficacious for each individual. In order for a person to have a reason for action that person must be able to start from something, for 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.

The relevance of rational choice to morality

We have now said something about what kind of justification we are aiming at but what exactly is it a justification of? What does "morality" stand for in this context? I have already said that our project is to justify morality as a set of interpersonal rules for behaviour. I will later characterize morality as a set of constraints on action. One way to interpret this is to say that we are trying to justify the commonly accepted moral code in this or any given society. To do this one would have to claim that the accepted moral code is rational (or something similar to this). This is, however, far removed from our project. The morality of fundamental justification may not resemble the moral code, in any present or historical society. Fundamental justification offers a revisionist account of morality. Gauthier 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.[22]

This identification of moral principles as a subset of rational choice principles can be questioned, however. The first interpretation of this relevance objection agrees that a fundamental justification of rational principles can be given, but it questions that these be called moral principles. David Copp writes:

> 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 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.[23]

Despite Copp's remarks that it is not merely verbal, I think that his objection is trivial. It is indeed true that the contractarian does not have anything to say to the sceptic about the rational credentials of morality, *if* morality is taken to mean *the commonly accepted* morality, or morality *as it is traditionally understood*. This is, however, not a big problem for the contractarian, since he is not aiming to justify the commonly accepted morality, as such. Gauthier himself writes in a later text that deliberative justification "*ignores morality, and seemingly replaces it.*"[24] It seems to me that the objection fails to take into consideration that the contractarian account of moral principles is a revisionist one. What the contractarian is trying to do is rather to provide a justification of an alternative account of morality, and it is assumed that this alternative account overlaps to a large degree with the commonly accepted morality.

Only that part of the commonly accepted morality that overlaps with our rational contractarian morality can be given a justification. And the contractarian *has indeed* something to say to the sceptic about the rational credentials of *that part of* morality. If the contractarian project is successful, that is. The "rest" of morality, however, remains unsupported. What the relevance objection finally boils down to is whether the rational choice

[22] Gauthier (1986) p 2.
[23] Copp in Vallentyne (1991) p 208.
[24] Gauthier in Vallentyne (1991) p 20. The same essay can also be found in Darwall (2003).

contractarian should be allowed to put the label "morality" on the rational principles that he generates through the theory. This is a verbal issue and, I think, trivial. It would not be devastating for the contractarian project if we gave up the label "morality", as the objection suggests, but it would, however, be very confusing, especially given the highly plausible assumption that the overlap between the contractarian morality and the commonly accepted morality is significant.

The second, more substantial, version of the relevance objection, challenges us to prove that the morality we want to give a fundamental justification is the *true* morality. This version of the objection comes from Holly Smith, she writes:

> 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.[25]

The interesting part of the contractarian argument is that it claims to show that we have self-interested reasons to be moral. Smith is right that the success of such an argument would not show that morality is true or correct in any *other* sense, apart from showing that it is sanctioned by each individual's rationality. It is unclear, however, what kind of justification Smith has in mind that would show morality to be true or correct. If there is such a justification of morality, this needs to be shown and the burden of proof, obviously lies with the one who claims that there is one. Showing that we have self-interested reasons to do "what morality, *if it were true* (or correct), would demand" is enough.

This version of the relevance objection can be turned against itself. Even if it is claimed that morality can be shown to be true (or correct) in any other sense than the one the contractarian has in mind, it must also be shown that this "true morality" would survive conflict with the contractarian morality. It is, to me, very unclear why anyone should accept

[25] Smith in Vallentyne (1991) pp 249-50.

moral principles in so far as they are not also rational principles. A moral principle, lacking rational support, not only hangs unsupported but is opposed by what is rationally required. 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 why one should be moral. Here I simply deny the importance of such separate claims and stick with the "real" claims of each individual's rationality.

Note on the assumption of mutual unconcern

Fundamental justification, as we have seen, requires that we derive morality strictly as rational principles for choice without introducing prior moral assumptions. The contractarian argues that moral constraints can be justified (from the point of view of the individual) regardless of what desires the individual has (provided that their preferences are such that mutual advantage from cooperation is possible). What it is rational for an agent to agree to depends on his or her preferences.

Both Rawls and Gauthier, however, make an assumption of mutual unconcern – that agents take no interest in each other's interests. They do not assume that actual persons are mutually unconcerned. It is only in the contract situation that mutual unconcern is assumed. But what place can an assumption of mutual unconcern have? The contractarian position is that (almost) no matter what people's preferences are like, it is rational for them to agree to (and to comply with) constraints on the pursuit of their interests. An assumption of mutual unconcern therefore, at least initially, seems out of place.

Is this a problem for fundamental justification? It is open to Rawls to argue that other-regarding interests are morally irrelevant and therefore should not be allowed any influence in agreement. This defence of the assumption is, however, not open to Gauthier, who wants to generate morality without assumptions about what is morally relevant. Since the parties to agreement in Gauthier's theory are real determinate individuals, characterised by their capacities and concerns, any assumption about people's preferences may seem out of place.

The idea of the assumption is to show that, (almost) no matter what our preferences are like, there are rationally acceptable constraints on conduct and these constraints does not depend on our mutual affectivity. Gauthier writes: "*…we agree with Kant that moral constraints must apply in the absence of other-directed interests, that indeed they must apply* *whatever preferences individuals happen to have*."[26] And further that one is not "*able to escape morality by professing a lack of moral feeling and concern, or a lack of some other particular interest or attitude, because morality assumes no such affective basis*."[27] For this purpose *no* assumption is needed. An assumption of mutual unconcern is not needed, and an assumption of mutual concern is also not needed. This is one assumption less (not one more) than the theorist who wants to base morality on affectivity has to make.

Peter Vallentyne argues that the idea of the *existence* of rational constraints must be kept separate from the *contents* of the particular constraints.[28] In establishing the existence of rational constraints on the pursuit of self-interest (given that mutual benefit from cooperation is possible), the contractarian does not depend on any sympathetic concern for others. The existence of constraints, specifically, does not depend on (altruistic) preferences. *Even if* we were purely self-interested it would be rational to agree to principles constraining the pursuit of self-interest. In determining the contents of the particular, rationally justified, constraints, however, it is inappropriate to ignore any of one's considered preferences. This gives the impression that the idea is that of a worst-case- scenario. This however, is not so, since the worst case would not be one in which people were mutually unconcerned, it would be one where people were strongly, negatively, concerned with others. If this were the case, the contractarian project, which depends on the prospect of mutual benefit, would be impossible.

Rationality, it seems to me, requires that all relevant information should be allowed and that all assumptions used be realistic. How realistic this assumption is, is in part an empirical question. Taken literally, the assumption is obviously false, but if we reflect on the issue, we may come to realize that the assumption can be interpreted realistically, as approximatively true. People surely take an interest in each others interests, on a small scale; we care about those close to us. But it is not totally unrealistic to assume that we are generally unconcerned, neither positively nor negatively, with the well- being of utter strangers. In this case, the assumption may be used as a simplifying assumption. Given this, I don't think that this issue is a big problem for fundamental justification. I now move on to discuss an important feature of our project; rationality.

[26] Gauthier (1986) p 100.
[27] Gauthier (1986) p 103.
[28] Vallentyne in Vallentyne (1991) p 73.

What is rationality?

As we have noted above, the theory of rational choice treats practical rationality as strictly instrumental. On this view, rationality is an instrument for achieving one's ends, whatever those ends might be. An agent acts rationally insofar as she acts effectively to achieve her ends, given her beliefs. This idea was expressed by David Hume, when he said that *"Reason is, and ought only to be the slave of the passions."*[29] The rational choice contractarian attempts to show that morality can be instrumentally efficient in this sense. This view, that we henceforth call *instrumentalism*, is closely connected to our task of finding a fundamental justification of morality. Danielson writes that "*the quest for a fundamental justification of morality properly begins with instrumental rationality, for two reasons. A justification must be embedded in a normative theory and the premises of a fundamental justification must be non-moral*."[30] Rational choice theory provides this non-moral but still normative framework that we need. The theory is *normative*[31] in that it tells us what we ought to do (in order to achieve our aims) and it is *non-moral* in that it doesn't presuppose anything moral; it takes as its starting point the non-moral ends of agents. Ends provide reasons for pursuing means and that these ends are non-moral guarantees that the justification will be fundamental. Instrumentalism thus provides the motivation needed for our theory to be practical. Without this motivation, we would have an explanation, not a justification.[32] I can't think of anything else that is both normative and non-moral in this required sense.

The instrumental theory, then, does not tell us what our aims ought to be, the theory is not concerned with *the ends* of action; that we leave to the individual's preferences. Again, we are in agreement with David Hume, when he wrote that it is *"not contrary to reason to prefer the destruction of the whole world to the scratching of my finger."*[33] But we do need, I think, as necessary conditions for rational preference, some conditions for coherent and considered preference. In order for an agent to act rationally he must order his ends in some kind of ranking that is transitive and perhaps in accordance with some other requirements as well. But there is no need to go into detail here, since the instrumentalist has no problem with these additional requirements as long as they don't address the *content* of the particular preferences.

Instrumental rationality then, as we have seen, is the view that ends provide reasons for pursuing means. Rationality is a means to some end, individual interest, which provides

[29] Hume (1739) p 415.
[30] Danielson (1992) p 61.
[31] The normativity of rationality has been questioned. I will discuss this below.
[32] Danielson (1992) p 20.
[33] Hume (1739) p 296.

the basis for rational choice. Here it is also important to point out what we mean by "individual interest." On this point misunderstandings are very common. When we say "interests" we do *not* mean interests *in* the self but interests *of* the self, held by oneself as subject. That is what provides the basis for rational choice and action.[34] Another way of putting it is that the individual will act so as to achieve what he *sees* to be his aims, whether these are good for himself or for someone else. The important point is that it is *he* who sees them to be good, and that he does so in a sense that motivates him to act. 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 don't need to, include altruistic concerns.

Gauthier has a specific variant of the instrumental theory that he calls the *maximizing* conception of rationality. The maximizing conception identifies rationality with individual utility- maximization.[35] Choosing rationally is to select the action that yields the outcome with greatest expected utility where utility is a measure of individual preferences. I will henceforth make no difference between the maximizing conception and the instrumental theory broadly conceived. But it is worth pointing out that one might question the maximizing conception while still staying within instrumentalism.[36]

Gauthier contrasts this conception of rationality with what he calls the *universalistic* conception of rationality. The universalistic conception, deriving from Kant, is committed to the view that what makes it rational to satisfy an interest is independent of whose interest it is. The universalistically rational person thus seeks to satisfy all interests.[37] Connecting morality with rationality is therefore more easily accomplished by proponents of the universalistic conception of practical reason.

Preferring universalistic rationality, however, is the same as giving up on fundamental justification. The universalistic conception already includes the moral dimension of choice that rational choice contractarianism seeks to generate. A strong reason to prefer instrumentalism over universalism is therefore that proponents of universalistic rationality cannot provide morality with a fundamental justification. This is a reason, however, only if fundamental justification is accepted. Therefore I shall see whether we can say something in favour of instrumentalism on independent grounds. Since instrumentalism is a fundamental feature of our project of finding a fundamental justification of morality I will devote a

[34] Gauthier (1986) p 7.
[35] Gauthier (1986) pp 6-8.
[36] See Danielson (1992) pp 28-29, 61-62.
[37] Gauthier (1986) p 7.

considerable amount of space to this issue. But before we turn to this, I would like to say something about what place *value* has in rational choice theory. This is relevant for some of the arguments for instrumental rationality that will follow.

Value and rational choice theory

There are theories of value implicit in our project of finding a fundamental justification of morality based on instrumental rationality; *subjectivism* (as opposed to objectivism) and *relativism* (as opposed to absolutism). The two positions are often confused. To conceive of value as subjective is to conceive of values as dependent on affective relationships, that is, to see value as created or determined through preference. To conceive of value as relative is, according to Gauthier, to think that "*each person has his own good (or bad), and that the goods of different persons are not parts of a single, overall good.*"[38] Opposed to (the individualistic form of) relativism that Gauthier advocates is the absolute (or universal) conception of value that holds that value is the same for all persons. Gauthier points out that subjectivism and relativism are logically independent so that each position can be held without the other, but he argues that they go well together and that other combinations are "under pressure."

The most important conflict here is surely the one between subjective and objective conceptions of value. This fundamental question is far too large and complex to be given anything like a satisfying treatment in this essay. But we should say something about why one should accept the view that there are no objective values. We could say that subjectivism is implicit in our project of grounding morality on the theory of rational choice. But that would not convince someone questioning our project. Can we state independent arguments against objectivism? Gauthier follows Gilbert Harman and finds objective value explanatory redundant. He writes: "*Objective value, like phlogiston, is an unnecessary part of our explanatory apparatus, and as such is to be shaved from the face of the universe by Ockham's razor.*"[39] I will treat John L. Mackie's argument from "queerness" below as an argument for instrumentalism.[40] We turn now to arguments and objections to instrumental rationality.

[38] Gauthier (1986) p 50.
[39] Gauthier (1986) p 56.
[40] Mackie has several other arguments against objectivism about values. Mackie defends an error theory. See Mackie (1977) pp 15-49.

The appeal to the social sciences

As Robert Shaver points out, many defenders of instrumental rationality do no more than simply assert their belief.[41] Gauthier for example says about the maximizing principle that *"there is simply nothing else for practical rationality to be.*"[42] Though, some have *argued* for the instrumental theory. Shaver considers several arguments; two of them come from Gauthier. The first one appeals to the social sciences. Gauthier argues that the maximizing conception is almost universally accepted in social sciences, economics, decision theory and game theory. Therefore the onus of proof falls on those who defend universalistic rationality.[43] This argument is not entirely convincing. As Shaver points out, it is unclear why the social sciences should be thought of as providing the correct account of practical rationality. It seems at the first look as this argument is no more than an appeal to authority. Gauthier suggests further that defenders of universalism would not defend their conception of rationality unless this was necessary in order for them to be able to argue for their specific ethical theories.[44] Still, I think, that this argument as such does not do what it was set up to do, namely to put the onus of proof on the defender of universalism. This additional suggestion, however, hints at a fundamental wisdom. Practical rationality is not confined to ethics and, as I see it, it would be strange if ethics required a specific conception of rationality entirely foreign to the sciences. This, however, is an entirely different argument that is only hinted at in Gauthier. But even this argument is far from conclusive.

The appeal to weakness

Gauthier's second argument is that the maximizing conception possesses the virtue of weakness, that the maximizing conception is weaker than the universalistic conception. He writes:

> Any consideration affording one a reason for acting on the maximizing conception also affords one such a reason on the universalistic conception. But the converse does not hold. 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."[45]

[41] Shaver (1999) p 40.
[42] Gauthier in Vallentyne (1991) p 20.
[43] Gauthier (1986) p 8.
[44] Gauthier (1986) p 8.
[45] Gauthier (1986) p 8.

Shaver argues against this that deontological theories as well as hedonism could be made weaker than the instrumental theory in this sense. A deontologist might claim that we never have reason to break promises; the instrumentalist, on the other hand, might well find a reason for promise breaking. Hedonism leads to the same result, if x in unconnected to pleasure, my desire for x provides me with a reason on instrumentalism while not on the hedonist theory.[46] Therefore, hedonism and some deontological theories should be preferable to instrumentalism on the grounds of weakness. Shaver admits that there is a sense in which Gauthier is right. If one argues for universalistic rationality by first arguing for the instrumental theory and then by invoking extra considerations arrives at the universalistic theory, then, of course, instrumentalism is weaker than universalism. But this, as Shaver points out, presupposes that one arrives at the universalistic theory by a specific argument. Shaver also denies that any reason one has according to the instrumental theory also is a reason on the universalistic theory. He says that the instrumental theory might give me a reason to go to a movie tonight while the universalistic theory might give me no such reason.

True. The argument from weakness, in Shaver's interpretation of this argument, therefore tells us nothing. However, I think that this is a misinterpretation of the argument. Gauthier invites this misunderstanding by stating two arguments at once, one weak, the one to which Shaver is replying, and one stronger. The weaker argument, that the instrumental theory is weaker because it affords one fewer reasons to act than any other theory is not only false but weird too. There is no reason whatsoever that a theory that afforded one fewer reasons to act is preferable to any other. Weakness in this sense is not a virtue. It is unclear whether Gauthier actually intended this argument to be interpreted in this way.

There is, however, a much more plausible way of interpreting the argument from weakness. This argument is hinted at in the very passage Shaver is quoting (same as above), namely that the impersonality or impartiality that is required by the universalistic theory needs defence. It says that the instrumental theory is weaker, not in that it affords one fewer reasons to act than other theories, but in that it presupposes less than other theories. This argument should not be confused with the argument from fundamental justification that I mentioned above. The argument from weakness does not presuppose fundamental justification. Again, the argument from fundamental justification says that the universalistic conception of practical reason incorporates a moral dimension and is therefore question begging if used to

[46] Shaver (1999) p 47.

justify morality. The argument from weakness (in its plausible interpretation), on the other hand, says that the universalistic conception presupposes that the interests of all are the basis of rational choice and that this is an unsupported assumption that the instrumental theory does not have to make. In this latter sense, weakness *is* a virtue.

No deontological or hedonistic theories could be made weaker than instrumentalism in this latter sense. Deontological theories and hedonism need to assume things such as that promise breaking is always wrong or that conduciveness to pleasure is the only thing that affords one a reason to act. In the same way the universalistic conception of rationality needs to say that the interests of others as such provide reasons (as well as one's own interests). The universalistic conception, as we noted above, is committed to the view that what makes it rational to satisfy an interest is independent of whose interest is it. And, again, this assumption, of the impersonality or impartiality of reason, needs defence.

The appeal to motivation

Another way to express the view we are defending, that instrumentalism exhausts practical rationality, is to put it in terms of *hypothetical* and *categorical imperatives*. An instrumentalist like John L. Mackie holds that no categorical imperative is objectively valid.[47] Shaver presents Mackie's argument "from queerness" (against objective values) as an argument for instrumentalism. Mackie himself says that: "*So far as ethics is concerned, my thesis that there are no objective values is specifically the denial that any such categorically imperative element is objectively valid*."[48] A categorical imperative "*would express a reason for acting which was unconditional in the sense of not being contingent upon any present desire of the agent to whose satisfaction the recommended action would contribute as a means*."[49] The objective values then, whose existence Mackie denies, "*would be action-directing absolutely, not contingently (in the way indicated) upon the agent's desires and inclinations*."[50] Mackie describes his view thus:

> Let us suppose that we could make explicit the reasoning that supports some evaluative conclusion, where this conclusion has some action-guiding force that is not contingent upon desires or purposes or chosen ends. Then what I am saying is that somewhere in the input to this argument – there will be something which cannot be

[47] Mackie (1977) p 29.
[48] Mackie (1977) p 29.
[49] Mackie (1977) p 29.
[50] Mackie (1977) p 29.

> objectively validated – some premise which is not capable of being simply true, or some form of argument which is not valid as a matter of general logic, whose authority or cogency in not objective, but is constituted by our choosing or deciding to think in a certain way.[51]

Mackie's denial of objective values supports an argument for instrumentalism. Mackie argues that categorical imperatives are "queer" because knowledge of them

> provides the knower with both a direction and an overriding motive; something's being good both tells the person who knows this to pursue it and makes him pursue it. An objective good would be sought by anyone who was acquainted with it, not because any contingent fact that this person, or every person, is so constituted that he desires this end, but just because the end has to-be-pursuedness somehow built into it.[52]

Mackie thinks there are no such objects of knowledge. Shaver argues against this saying that "*one problem is that the defender of categorical imperatives need not claim, with Kant, that knowledge of them motivates regardless of one's desires*."[53] This line of reasoning in turn leads the defender of categorical imperatives to give up an important feature of ethical theory. Separating morality and motivation is, I think, deeply problematic. Especially if we keep in mind what I said above about the practicality or morals. That morality is to help us to solve conflicts of interest.

Shaver argues further that not even hypothetical imperatives are (as Mackie thinks) capable of being "simply true", or "valid as a matter of general logic". If Shaver is right in this it would be a problem for the instrumentalist. Shaver asks us to consider a piece of instrumental reasoning:

> (1) I desire to leave the room. (Suppose I have no conflicting desires.)
> (2) I know that walking through the door is the best way of leaving the room.
> (C) Therefore it is reasonable – I ought – to walk trough the door.[54]

[51] Mackie (1977) p 30.
[52] Mackie (1977) p 40.
[53] Shaver (1999) pp 44-45.
[54] Shaver (1999) p 45.

Shaver then argues that I might know (1), (2) and (C) but still fail to be motivated to walk through the door. He claims that I might suffer from "weakness of will". Shaver admits that in this case I have a reason or justification for walking through the door but I can still fail to be motivated to action. He concludes that if this reason or justification offered by this instrumental or hypothetical imperative is not defeated by noting that I can fail to be motivated, then the reasons and justifications given by categorical imperatives cannot be defeated by noting that it is possible to fail to be motivated.[55] This argument relies on an implausible idea of "weakness of will." The instrumentalist could reasonably deny that there is any such thing as weakness of will (in this sense).[56] If you desire to leave the room and you know that the best way of doing so is to walk through the door (and you have no conflicting desires) you are necessarily motivated to walk through the door.

If asked how (C) follows from (1) and (2) the instrumentalist could answer by simply repeating the premises: I ought to walk through the door because; I desire to leave the room and; I know that walking through the door is the best way of leaving the room. Or, as Shaver points out, we might add a premise:

(3) I ought to do what best satisfies my desires.[57]

But neither of these answers would, as Shaver also points out, add anything to the initial argument. We might instead say that (1) and (2) causes me to go trough the door and since I have no mistaken beliefs my doing so is rational and cannot be criticised. To this Shaver answers that on this view "*any action I perform without mistaken representations is reasonable.*"[58] Again he needs to appeal to weakness of will in order to argue that (1) and (2) might be true and yet fail, through weakness of will, to cause me to go through the door. "*Here contrary to the instrumental theory, not going through the door turns out to be reasonable.*"[59] Here Shaver seems to think that instrumentalism leads to some kind of contradiction. This, however, is not so unless we accept that weakness of will in this sense is possible here. Shaver gives us no reason why we should accept the weakness of will idea. I think we should reject it. Shaver gives one further argument in favour of doubting that (C) follows from (1) and (2). This argument appeals to our moral intuitions:

[55] Shaver (1999) p 45.

[56] This should not be confused with the "weakness of will" one might suffer from when one fails to be motivated to do what one thinks is morally right.

[57] Shaver (1999) p 53.

[58] Shaver (1999) p 57.

[59] Shaver (1999) p 57.

> Suppose that I have no desire to help the person writhing in agony before my eyes; my desire is to leave the room. Many will deny that I ought to walk trough the door rather than help. Many will no longer approve of concluding (C) from (1) and (2) or endorse (3).[60]

Shaver argues that claims of instrumental reasonableness will not always win approval when they conflict with moral approvals. *"Indeed, one might even try to argue that they win approval only when they do not conflict with moral ... approvals."*[61] Surely, one might try to argue that, but we can reasonably ask what bearing moral intuitions have on determining the correct conception of rationality. Even if we would admit that moral intuitions should be allowed in constructing ethical theory (something we have denied above), we would still, I think, have to deny such appeal when it comes to determining the correct conception of rationality. Rationality as such is not specific to ethics and it would be unreasonable to say that the correct account of practical rationality relies on moral intuitions. Moral intuitions are, I believe, irrelevant when it comes to determining the correct conception of rationality. This, I think, should be agreed to even by those who (unlike me) admits moral intuitions in other contexts. Shaver says further that

> …there is no reason for thinking that instrumental rationality exhausts practical rationality. For we sometimes suppose that the desires of others can provide me with reasons directly. This latter thought does not need a special and dubious defence, to be conducted while the instrumentalist, secure with a firmly founded theory, looks on with a critical eye. Both accounts of rationality rest on equal footing – both are backed by appeal to the sort of inferences we are willing to admit.[62]

It may be true that we sometimes *suppose* that the desires of others provide us with reasons. But this does not prove that the desires of others actually do provide us with reasons. It is also unclear how "directly" is to be interpreted here, do the desires of others provide us with reasons through some other kind of reasoning different from instrumental reasoning or through no reasoning at all? Shaver's idea that "*both accounts of rationality rest on equal*

[60] Shaver (1999) p 58.
[61] Shaver (1999) p 58.
[62] Shaver (1999) p 58.

footing" suggests that there is a special kind of reasoning for cases when the desires of others provide us with reasons apart from the instrumental reasoning employed when it is our own desires that provide us with reasons. If this is what he means, it is strange for him to claim that this special kind of reasoning does not need any special defence. As I see it, these two kinds of reasoning do *not* rest on equal footing. And there are, as we have seen, indeed reasons to think that instrumental rationality exhausts practical rationality.

Reflections on the normativity of rationality

As I said above, the idea to give morality a fundamental justification depends on instrumental rationality. It depends, specifically, on the idea that rational choice theory provides a justificatory framework that is, at once, non-moral and normative. The alleged normativity of rationality has been questioned, however. A discussion of this issue is relevant for my thesis since I want to answer the question: "Why should I do what morality requires", with something like: "Because rationality requires you to do so", but then, there is the natural follow-up question: "Why should I do what rationality requires?" How, if at all, can this second question be answered?

We all agree that rationality requires various things of us. Specifically, the rational choice contractarian thinks that rationality requires us to be moral. There is, however, a further question about the status of these requirements; are the requirements of rationality normative, and if so, in what sense? John Broome says that they are automatically normative *in one sense*. "*Rationality is a system of requirements or rules. It therefore sets up a notion of correctness: following the rules is correct according to the rules. That by itself makes it normative in one sense*."[63] In this sense, he continues, the requirements can be compared to those of conventions or of Catholicism. Convention, for example, requires us to shake hands with our right hand and Catholicism, Broome says, requires us to abstain from meat on Fridays.

Broome, however, does not want to say that such requirements are "normative" in another, further sense. The term "normative", according to Broome, has to do with "ought" or "reasons", while "requirement" does not mean anything normative, in this sense. With this separation between the requirements (of rationality), at the one hand, and normativity at the other, he asks what he calls "the normative question"; is rationality normative (in his preferred sense of that term)? Can the normative question be answered, given this separation between requirements and normativity?

[63] Broome (2005) p 2.

Let us return again to the question, “Why should I do what rationality requires?” How can we answer this question? One might answer: “because you ought to be rational.” But what sense can we make of such an answer? Especially, what does the “ought to” mean here? As I see it, there are two possible interpretations of “you ought to be rational”; that you ought *morally* to be rational, and that you ought *rationally* to be rational. Will any of these answers do? With the first answer, it is obvious that we would be reasoning in a circle. “You ought morally to be rational” is equivalent to saying that “morality requires you to be rational” and then, we would be back where we started; “But why should I do what morality requires?” This appeal to morality would beg the question, and hence, not give us a fundamental justification of morality.

Is the second variant, that you ought rationally to be rational, more promising? This answer seems just as dissatisfying as “you ought to be moral because morality requires it.” And it is, exactly, the dissatisfaction with that kind of answers that motivates the thesis of this essay. It seems that we have no answer to the “normative question”. At least not any answer in addition to the one Broome himself gave, that the requirements of rationality are normative in the same way as the requirements of conventions and Catholicism are normative. Given that you are a Catholic, you ought not to eat meat on Fridays and given convention; you ought to shake hands with your right hand. Why does this not do as an answer to the normative question? Obviously we may ask; “But why should I be a Catholic?” or “Why should I follow conventions”. These questions makes full sense, reasons can be given for and against being a Catholic etc. Can we (reasonably) ask the same kind of questions about rationality?

Given the instrumental interpretation of rationality that I defended above, we can reduce the question “why be rational?” to something like “why should I do what best achieves my own aims (given my own preferences)?” Is this, second question, meaningful in the same way that “why should I be a Catholic” is a meaningful question? For sure, it is an open question; it is not an analytical truth. But what reasons could ever be given for and against being rational? Can someone asking that question be persuaded by any “rational” argument? Well, we could say that you ought to be rational for *instrumental* reasons but this is what rationality *is* and not a further reason to *be* rational.

If objective values existed, or if at least some categorical imperatives were objectively valid, then some of these might speak in favour of being a Catholic, or of being rational etc. But as I argued above, value is subjective and relative, and no categorical imperatives are objectively valid. Given that I am right about this, what reasons could ever be given for the conclusion that I ought to be rational? It seems to me that all we can say is that *given* that you *are* rational*, then* you ought (rationally) to be moral. This normativity is hypothetical. *In this sense* there is no difference between the normativity of rational requirements and the normativity of conventions or of Catholicism. But if rationality is normative only in the same way as conventions or Catholicism are normative, then why is it preferable to provide a fundamental justification of morality in terms of rationality rather than in terms of, for example, Catholicism? There are, at least, two reasons for this. I will return to this shortly.

Broome, however, seems to require more than this, as an answer to his normative question. He seems to require the normativity of rationality to be categorical. Broome's own conclusion about the normative question, therefore too, is sceptical.[64] He can see no conclusive reason to say that rationality is normative (in his sense). But he says that if it is, that would most likely be for "instrumental reasons". This claim, however, is confusing since he gives "instrumental" a radically different meaning than mine.[65]

As I see it, the rational choice contractarian does not need to assume that rationality is normative in any other, stronger, sense than, for example, conventions are normative. In fact the contractarian views morality *as* convention. When Gauthier says things like, "*Rational choice provides an exemplar of normative theory*",[66] we do not need to attribute to him a stronger view. What I want to do is, to give a fundamental justification *of morality*, *not* a fundamental justification of *rationality*. A fundamental justification of rationality, cannot, I believe, be given. The claim of fundamental justification is that the moral realm can be reduced to, or fully understood in, terms of instrumental rationality. It is *not,* that the rational realm can be reduced to, or fully understood in, terms of anything else.[67]

Importantly, this does not mean that moral justification and rational, or deliberative, justification rest on equal footing. Gauthier says, about deliberative justification and its relation to moral justification, that "*we need not suppose that this deliberative justification is itself to be understood foundationally. All that we need suppose is that moral justification*

[64] Broome (2005) p 14.

[65] Broome (2005) p 9. "*…if rationality is indeed normative, that seems to be because of what we can achieve by being rational. It seems likely to be for instrumental reasons, as I shall put it. Since some philosophers give a different meaning from mine to the word 'instrumental', I need to be clear about mine. I am not suggesting the requirements of rationality to be normative because satisfying them is a way of satisfying our desires. I am suggesting they might be normative because satisfying them is a way of achieving some of the things we ought to achieve. I ignore the possibility of total scepticism about normativity; I take it for granted there are some things we ought to do, some things we ought to hope for, some things we ought to believe, some things we ought not to do, not to hope for, not believe, and so on*."

[66] Gauthier (1986) p 4.

[67] Maybe the rational realm could be reduced to psychology, or biology? And ultimately to physics? But this is no part of my argument. (And such a reduction would not amount to a justification of any kind, merely an explanation.)

does not plausibly survive conflict with it."[68] There is no need to provide rationality with a fundamental justification. All we need to do is to show that deliberative justification is, somehow, more fundamental than moral justification. But exactly how is the rational realm more "basic", or more fundamental, than the moral realm?

As I noted above, there are, at least, two reasons for this. Firstly, no one thinks that the rational realm can be reduced to, or fully understood in, terms of the moral realm. Take an example, like the one I used above to illustrate instrumental rationality. I ought to walk through the door because (1) I desire to leave the room and (2) I know that walking through the door is the best way of leaving the room. We all agree that this is an example of rationality (the controversy was only whether instrumental reasoning *exhausts* practical rationality). But no one seriously maintains that this, necessarily, has something to do with morality. Rational principles cannot, plausibly, be viewed as a subset of moral principles. The other way around, surely, seems more appropriate.

A second reason has already been mentioned above. Questions like "why should I be a Catholic?" or "why should I follow conventions?" or for that matter "why should I be moral?" can reasonably be asked, and reasonable answers, that is, reasons for and against, can be given. The same cannot, it seems, be said of the question "why be rational?" To give reasons, for or against, being rational seems to presuppose rationality. Attempts to persuade someone with *arguments* seem to presuppose that this person is rational (in some sense, at least). One could argue, for example, that one should be rational because this is what God has commanded. But then, of course, we would be asked "but why should I do what God has commanded?" which in turn is met with "Otherwise you will be punished!" Is asking "but why should I not be punished?" a reasonable response to this? Given instrumentalism about rationality, this latter question seems out of place. It could be met with "Because you wouldn't want that, given your own preferences!" After that there is simply nothing else to add. We are not able to reach a person questioning that with any rational argument whatsoever. This is the rock bottom.

Yet another reason, to view the rational realm, or the deliberative mode of justification, as more fundamental, is that the class of rational beings is, I assume, larger than the class of moral beings. Or more correctly, the class of people who accept deliberative justification is, presumably, larger than the class of people who are prepared to accept moral

[68] Gauthier in Vallentyne (1992) p 20.

justification (and, surely, larger than the class of Catholics, anyway). Therefore, rationality seems to be a more appropriate foundation for morality.

Summary

I have defended an instrumental conception of practical rationality. On this view, rationality is a means to some end, individual interest, which provides the basis for rational choice. An agent acts rationally insofar as she acts effectively to achieve her ends, whatever those ends might be. The universalistic conception of rationality, on the other hand, incorporates a moral dimension and is therefore incompatible with fundamental justification. If we see the theoretical advantages with fundamental justification, we should reject universalism in favour of instrumentalism. This was the argument from fundamental justification. This argument, however, may not be universally convincing since it presupposes, in some sense, the goal of fundamental justification. Therefore, I discussed further, independent, arguments.

From the argument from the social sciences I concluded that it should be seen as a theoretical advantage to have a conception of rationality that is fit not only for ethics but for the sciences as well. Our conception of rationality should best be seen as more fundamental than ethical theory itself and therefore our conception of rationality should not rest on moral considerations. Furthermore, the instrumental view goes well together with the idea of moral motivation. Proponents of categorical imperatives need either to explain how these motivate us to act, or give up the tight connection between morals and motivation that is offered by the instrumental theory. This last move is especially unattractive against the background of plausible assumptions about the practical and conflict-solving nature of morality.

I argued further, that rational choice theory provides a non-moral but still normative framework needed for our justification to be fundamental. Rationality is normative in that it tells us what we ought to do in order to achieve our aims, and it is non-moral in that it doesn't presuppose anything moral in its premises.

I conclude from this that, in order to get a fundamental justification of morality off the ground, we need a conception of rationality that is strictly instrumental and, most importantly, does not incorporate any of the moral content that we wish to have emerging in our conclusions. The universalistic conception fails on these grounds, while the instrumental one does not. Therefore, I stick with the instrumental theory.

From rationality to morality

The structure of contractarianism

A person asking "Why be moral?" is really asking two separate, but closely related, questions. The first is whether it is rational to agree on certain principles and the second is whether it is rational to actually comply with these principles. Most people agree that it would be rational to agree on certain principles (at least provided that they expect others to comply) but the problem is that it would often be rational to defect, that is, fail to carry out that what was agreed. This, the second of the sceptics' questions, is known as *the compliance problem*. These two questions can be pictured as a two-level-structure and one should distinguish between the two "levels" of a contractarian fundamental justification. 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"*.[69] Hobbes imagines a pre-social, pre-moral state known as the "state of nature". We 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 freedom to do whatever he can to preserve himself and 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)...*" [70] The fundamental law of nature tells us to seek peace and follow it wherever it may be found, and when it may not, by "right of nature", to defend ourselves by all the means we can. 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 men, as he would allow other men against himself."*[71] (This is also the effect of the Lockean proviso that we will introduce below.)

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, there would be *"no reason*

[69] Hobbes (1651) p 86.

[70] Hobbes (1651) p 87. The word "right" here is confusing; perhaps it would be better to speak of an unlimited "freedom". Or say that in the state of nature there, simply, are *no* rights whatsoever.

[71] Hobbes (1651) p 87.

for any one, to divest himself of his: for that were to expose himself to prey […] *rather than to dispose himself for peace."*[72]

It is in the introduction of the third law that things start 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. He, therefore, introduces the third law , which is "*that men perform their covenants made*". And in this law of nature, he continues, "*consisteth the fountain and original of JUSTICE.*"[73] 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"*[74] 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 (I later refer to such situations as "tight corners"). 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.*"[75] As we will see, Hobbes and Gauthier answer the Foole rather differently, however.

Hobbes' answer involves the notion of a "sovereign." An institution, or agency, is empowered to intervene and thus change the incentive structure by imposing sanctions. This feature makes Hobbes' solution *political* rather than *moral*. Actually, it is not a solution at all, properly speaking, since it changes the problem into another one. Morality involves constraint. Constraints tell us to sometimes act against our own interests. On Gauthier's view, these constraints are themselves justified by appeal to those interests. Hobbes' political solution, on the other hand, involves no constraints. The sovereign makes compliance directly rational. Therefore, the political solution makes morality unnecessary.[76] There are further problems with the political solution, 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.*"[77] Gauthier also says that if the free market acts as an invisible hand,

[72] Hobbes (1651) p 87.
[73] Hobbes (1651) p 95.
[74] Hobbes (1651) p 96.
[75] Gauthier (1986) p 161.
[76] Gauthier (1986) pp 163-164.
[77] Gauthier (1986) p 164.

the sovereign acts as a very "visible foot", *"directing, by well-placed kicks, the efforts of each to the same social end."*[78]

The need for constraint

Before I discuss Gauthier's solution to the compliance problem, I want to say something about the need for morality. As I said above, the political solution makes morality unnecessary. The political solution is, however, both rationally and morally unattractive and should be rejected. Morality is needed in order to solve some of the problems the sovereign was supposed to solve. There are, however, other, more attractive, ways in which morality is unnecessary. Morality involves constraints, and whenever optimal outcomes can be achieved without constraints, morality is not needed. Adam Smith's idea of the market as an invisible hand is an example. In the (perfectly competitive) market, no constraints are needed to achieve optimal outcomes (in the Paretian sense of that term). Gauthier's speaks of the market as a "morally free zone." This is misleading, however. The perfectly competitive market is, as Gauthier also realizes, an idealization. This is still slightly misleading. Real-world markets are not morally free zones. The market is more like a morally free zones *framed by morals*.

Gauthier's argument is that in a perfectly competitive market, mutual advantage is assured by the unconstrained activity of each individual, so that there is no place, rationally, for constraint.[79] "*The market exemplifies an ideal of interaction among persons who, taking no interest in each other's interests, need only follow the dictates of their own individual interests to participate effectively in a venture for mutual advantage*."[80]

To say that morality is unnecessary is *not*, Gauthier stresses, to "*denigrate the value of morality, which makes possible an artificial harmony where natural harmony is not to be had. Market and morals share the non-coercive reconciliation of individual interest with mutual benefit."*[81] If the market acts as an *in*visible hand, morality works as a *visible* hand, constraining each, for the good of all. But where is this "natural harmony" to be found and when do we need morality to achieve "artificial harmony"?

The compliance problem is often depicted as a so-called prisoner's dilemma. I will not go into detail here since most readers, I believe, are familiar with the problem. Gauthier's

[78] Gauthier (1986) p 163.
[79] Gauthier (1986) p 13. Gauthier also argues that there is no need, morally, for constraint. Since in the market each person enjoys the same freedom in her choices that she would have in isolation from her fellows, "*and since the market outcome reflects the exercise of each person's freedom, there is no basis for finding any partiality in the market's operations*."
[80] Gauthier (1986) p 13.
[81] Gauthier (1986) p 14.

solution, that I will discuss shortly, is a solution to a special class of prisoner's dilemmas, namely the one-shot games. In order to achieve an optimal outcome in a one-shot prisoner's dilemma, constraint is needed. In (indefinitely[82]) iterated prisoner's dilemmas, however, we don't need to use morality to achieve cooperation. Peter Danielson argues that since iterated games can be solved by straightforwardly rational agents, they are not morally significant.[83] The straightforwardly rational principle, to be followed in iterated prisoner's dilemmas, is called "tit for tat." This principle says, "Cooperate initially and then match the other player's previous action."[84] Danielson says:

> As Axelrod has shown, TFT [tit for tat] effectively induces widespread co-operation. Therefore it is an impartial, mutually beneficial principle. None the less, TFT is not a moral principle because in the iterated Prisoner's Dilemma it is straightforwardly in an agent's interests. Given the expectation of future interactions, and other agents' responsiveness, each of the choices required by TFT is directly maximizing. Since straightforward maximization suffices here, there is no need for a new kind of principle, namely a moral principle constraining an agent's self-interest.[85]

Thomas Hobbes' sovereign, Adam Smith's invisible hand and Robert Axelrod's tit for tat, therefore, have that in common, that they are examples of how constraints can be made redundant. The sovereign is, as I said, rationally (as well as morally) unattractive. To what extent the perfectly competitive market can be realized in the real world, and to what extent real-world social problems are best depicted as iterated, or one-shot, games are both empirical questions. Danielson calls all these solutions, that do without constraints, *institutional*, and he says that "*…by relieving morality of some burdens, institutions that promote co-operation make morality's job easier. Morality need not support the entire social world by chains of obligation*."[86] I wanted to bring this up, only in order to put it aside, and I continue now to discuss situations where constraints *are* needed.

[82] There is a further question about how one should act in a finite set of games. I leave that aside.
[83] Danielson (1992) p 45.
[84] Danielson (1992) p 46. The Tit for tat principle was originally formulated by Robert Axelrod.
[85] Danielson (1992) p 46. The rationality of TFT is dependent on other's dispositions. In an unlikely world full of defectors, TFT would not be rational.
[86] Danielson (1992) p 49.

Gauthier's solution

I return now to the compliance problem and the objection of the Foole. Gauthier thinks the rational person adopts a *disposition* to co-operate. The rational agent has reasons, ultimately traceable to his or her preferences; to adopt a disposition, or principle, to behave cooperatively and to keep commitments. This is so, even when the keeping of commitments is not directly utility-maximizing. This disposition, which Gauthier calls *constrained maximization,* makes one an eligible partner in beneficial co-operation, and so it is itself advantageous. The ability to keep commitments has advantages, whatever ends an agent might have. It is not always advantageous to cooperate though. Only when cooperating with others who are similarly disposed can advantage be had by both parties. If one were to cooperate unconditionally (with all others), one would be open to exploitation; this would make one a "sucker", and allow others to "free ride". Constrained maximization is thus conditional; it tells us to cooperate only with those who are similarly disposed. That is, given that one's disposition can be known, or sufficiently suspected.[87]

Not to comply with one's rationally made agreements, on the other hand, *by appealing directly to utility-maximization*, is itself also a disposition and one that it is disadvantageous to have because it excludes one from participating in highly beneficial co-operative arrangements. This, the latter disposition, Gauthier calls *straightforward maximization*.

Here is what Gauthier thinks we need to say to the Foole. We need to say that it is rational to abide by one's covenants, even when such performance is not directly to one's benefit, given that the *disposition* to perform *is* to one's benefit. This means, in particular, that individual actions are somehow made rational, insofar as they are manifestations of a rational disposition. Gauthier's argument identifies practical rationality with utility-maximization *at the level of dispositions to choose*.[88] According to this idea, the rationality of the disposition somehow carries over, or transfers, to the individual actions in which the disposition manifests itself. This idea has been subject to much criticism.

The problem arises in so called "tight-corner" situations. The "forward-looking" nature of the theory of rational choice judges possible courses of action on the basis of their consequences. Commitments, on the other hand, are "backward-looking" and require one to act on the basis of dispositions or principles. A tight-corner situation is one is which keeping a

[87] If human beings are "opaque" to their fellow beings, there is a big problem for this approach. Then it no longer has to be true that being disposed to constrained maximization is advantageous for an agent. Whether we are opaque or not is an empirical question, and one I intend to pass on.

[88] Gauthier (1986) p 187.

commitment one has made happens to cross one's interests. How can, for example, the keeping of a promise, be rational under such circumstances? Rational choice theory seems to treat all commitments as either redundant (if the same action required by the commitment would be recommend by forward-looking rationality anyway) or irrational (if the keeping of the commitment would be contrary to rationality.)

The problem is analogous to that of rule utilitarianism. The breaking of a promise might do more good (in terms of maximization of overall happiness) than the keeping of the promise. How then can promise-keeping under these circumstances be good?

I do not claim to have a conclusive answer to this ancient question. Nor can I pretend to be able to respond to all of those critics of Gauthier's solution. My strategy here is merely to say something very general, inspired by Michael Thomson's interesting essay "*Two Forms of Practical Generality*."

Transfer principles

Michael Thomson calls Gauthier's theory of rationality a "two-level theory". A two-level theory, he suggests, is marked by two central propositions, (1) *a standard of appraisal* and (2) *a transfer principle*.[89] Gauthier's standard of appraisal is the agent's own utility as a measure of individual preference satisfaction. And his transfer principle is something like: "a rational disposition makes the actions manifesting it rational." Thomson argues that a similar structure can be found in rule utilitarianism and also in Rawls' principle of fairness.[90] The standard of appraisal is used to "*govern the attribution of the relevant type of goodness to a practice or disposition*."[91] In the case of Gauthier, the "relevant type of goodness" is the rationality of the disposition. How, then, can this rationality be "transferred" from the standard of appraisal to the individual choices or actions?

A transfer principle specifies a *transparency principle* or *mediating element*. A disposition is one example of such an element. Other examples are "*a rule, a principle, a set of principles for the general regulation of behaviour, a practical identity, an intention, a plan, a plan of life, a course of action, a motive, a maxim, or the like*."[92] Thomson says further that a transfer principle will also refer to some *relation of expression* that individual actions may bear to occupants of the mediating category, such as "*executing, falling under, manifesting,*

[89] Thomson in Morris & Ripstein (2001) p 129.
[90] Thomson in Morris & Ripstein (2001) pp 133-135.
[91] Thomson in Morris & Ripstein (2001) p 129.
[92] Thomson in Morris & Ripstein (2001) p 130.

realizing, acting on, according with, being part of, etc."[93] The transfer principle will also refer to "*the particular normative quality that the doctrine represents as "transferred" from occupants of the mediating category to the individual actions that "express" them.*[94] And this might be "*rationality, moral goodness, moral rightness, fairness, reasonableness…*" etc.

The notion of a transparency principle or mediating element may seem to be a little obscure. I wish to help understanding, by means of an analogy. We might think of the standard of appraisal as a source of light and the transparency principle, or mediating element, as a medium through which the light is supposed to pass, in order to fall on the individual actions. Thomson says, that a proposition that expresses transparency tells us something like this:

> If an occupant of a mediating category [in Gauthier's case, a disposition] has the appropriate normative property, then an individual action that bears the expression relation [in Gauthier's case, a manifestation] to it *thereby* also acquires that normative property (or another suitably associated property) [in Gauthier's case, rationality]. Various qualifications might be admitted into such a thought without affecting its standing as a transfer principle.[95]

So, Gauthier's transfer principle says something like this: "a disposition (the mediating element), if rational (the relative normative property), makes the actions manifesting it (the expression relation) rational (the transferred property)." This, however, is merely a clarification and a generalisation of the position and not a conclusive reason why we should accept, what I called, Gauthier's solution. I don't intend the above discussion to be conclusive. But hopefully Thomson's remarks can, at least, help to shed some new light on the discussion.

Despite its many problematic aspects, I myself am inclined to accept Gauthier's solution. I do this on the basis of that there is something deeply disturbing about the idea that rational agents, defined so as to maximize their utility, still fail to bring about the outcome that is better for each. There simply has to be something wrong with a theory that tells, each rational agent to defect in prisoner's dilemma (and related) situations, even though everyone acting like this brings about a suboptimal result. This intuition of mine should not, however,

[93] Thomson in Morris & Ripstein (2001) p 130.
[94] Thomson in Morris & Ripstein (2001) p 130.
[95] Thomson in Morris & Ripstein (2001) p 130.

be confused with the moral intuition that straightforward maximization leads to morally unacceptable results. Appeal to the moral intuition would be inappropriate here. It is rather the intuition that straightforward maximization leads to *rationally* unacceptable results that I appeal to here. The fact that I cannot argue conclusively for this may be that I'm just not clever enough.

I move on now to one other issue, more closely related to fundamental justification. I will discuss whether it is possible to give a justification of rights and duties, without sneaking in metaphysical assumptions, or relying on moral intuitions.

Rights and the Lockean proviso

As noted above, contractarians view morality as a kind of agreement among rational beings. And further that these rational beings "bargain" their way into morality. If this is so, the bargaining has to start somewhere. This "somewhere" is referred to as "the initial bargaining position". There may be several ways of specifying the initial bargaining position and the question that we are primarily concerned with here is which initial position(s) are compatible with fundamental justification, and which are not. Gauthier argues that the initial bargaining position is the non-cooperative outcome *constrained by the Lockean proviso.*

The Gauthierian interpretation of the proviso differs significantly from that of Locke himself. Firstly, Locke's theory has some theological undertones that are entirely absent in Gauthier. Secondly, Locke assumes that each person begins with an exclusive right to his body and its powers, and introduces the proviso to constrain a particular activity, namely, the acquisition or appropriation of external objects. For Gauthier, on the other hand, the proviso has a much wider and more basic role to play in moral theory. What Gauthier wants to do is, instead, to start in a Hobbesian state of nature in which there are no rights whatsoever, and then use the proviso to derive what Locke simply assumes. Another way of putting it is that his project is to "*convert the predatory natural condition described by Hobbes into the productive natural condition supposed by Locke.*"[96]

What, then, does the Lockean proviso say? In Locke's own formulation it is that one should leave enough and as good for others.[97] In Gauthier's formulation it is that the proviso prohibits *worsening* the situation of others *except when this in necessary to avoid worsening one's own position*. Rational utility-maximizers would not accept a stronger proviso. Worsening, and equally bettering, one's situation makes sense only in relation to some *base*

[96] Gauthier (1986) p 208.
[97] Locke ch v § 33 p 291.

point. This base point, Gauthier points out, cannot be the initial bargaining position itself because the base point is needed to determine the initial position.[98] Remember that, according to Gauthier, the initial position is the non-cooperative outcome constrained by the Lockean proviso. What then is the base point? Gauthier's answer is that the base point for determining how I affect you in terms of "worsenings" and "betterings", is determined by the outcome that you would expect in my absence.[99] Gauthier uses an example to illustrate this.

If I am drowning in a river and you pass by on the river bank, leaving me to drown, then you fail to better my position but you do not worsen it; I would still have drowned in your absence. In technical language, if the outcome is the same as it would have been in your absence, then you have not worsened my situation. But if, on the other hand, you come along and push me into the water, and then ignore my cries for help and leave me to drown, then you worsen my situation. I wouldn't have drowned in your absence; the outcome is worse for me than it would have been if you weren't there, hence you worsen my situation.[100] The proviso thus prohibits bettering one's situation through interaction that worsens the situation of another, unless abstaining from doing so would worsen one's own situation. Each is then free to better his or her own situation as he or she chooses, provided that he does not thereby worsen the situation of another (unless this is necessary to avoid worsening one's own situation).

To require that one must better the situation of others, would be to require that one gives *free rides*. And to allow that, in order to better one's own situation, one may worsen that of others would be to allow one to be a *parasite*.[101] This, Gauthier claims, expresses the underlying idea of not taking advantage[102] and thus of justice. Justice is "*the disposition not to take advantage of one's fellows, not to seek free goods or to impose uncompensated costs, provided that one supposes others similarly disposed*."[103]

The "ontological" status of the proviso

Can the Lockean proviso be used as the basis of a fundamental justification of moral rights? Or does Gauthier presuppose something moral in his derivation of moral rights from a pre-moral contract situation, and thus fails to live up to the requirements of a fundamental

[98] Gauthier (1986) p 203.

[99] "In my absence", should here, I think, be interpreted as; in the absence of me, but in the presence of all others. A moral objection, against this idea, would be that I may do whatever I want with you, given that someone else *would* have done so, in my absence.

[100] Gauthier (1986) pp 204-205.

[101] Gauthier (1986) p 206.

[102] Gauthier (1986) p 205.

[103] Gauthier (1986) p 113.

justification? If so, the Lockean proviso can be ruled out as the basis of moral rights by considerations of fundamental justification alone. As I see it, the proviso itself, being a moral constraint on interactions, also needs to be *derived* or *agreed to,* in some way or another, if we want to stay within fundamental justification. Does Gauthier do this? There are passages in *Morals by Agreement* that suggest that he does not. He writes, for example, that

> …the idea of morals by agreement may mislead, if it is supposed that rights must be the product or outcome of agreement. […] Rights provide the starting point for, and not the outcome of, agreement. They are what each person brings to the bargaining table, not what she takes from it.[104]

This suggests that Gauthier, like Locke himself, believes in some "natural rights." After all, the principle is called "Lockean" by Gauthier, and this suggests that he regards the proviso as expressing some "natural" rights. This would not fit very well with fundamental justification. However, Gauthier explicitly denies that the rights afforded by the proviso are natural in this sense. "*We must however recognize that these rights are not inherent in human nature.* […] *they do not afford each individual an inherent moral status in relation to her fellows*."[105] To clear away some confusion, regarding the ontological status of the proviso, we must explain the underlying idea that rights depend on the prospect of mutual advantage. In the state of nature, where people interact non-cooperatively, rights have no place. In the state of nature, nothing is forbidden or morally wrong. It is only a prospect for mutual advantage that brings rights into play.

> The moral claims that each of us makes on others, and that are expressed in our rights, depend neither on our affection for each other, nor on our rational or purposive capacities, as if these commanded inherent respect, but on our actual or potential partnership in activities that bring mutual benefit.[106]

This, I believe clears away some of the confusion. The rights afforded by the proviso in its Gauthierian interpretation are not "natural" in any metaphysical or intuitionist way, the proviso is not an a priori truth. If we want to derive morals strictly from rational choice, we

[104] Gauthier (1986) p 222.
[105] Gauthier (1986) p 222.
[106] Gauthier (1986) p 222.

cannot allow any of that into our premises. The proviso is rational only with the prospect of mutual advantage.

Extending the proviso retroactively?

The proviso, as we have seen, prohibits each from bettering his situation by worsening that of others. This, Gauthier claims, confirms each in the use of his own powers and affords to each the exclusive use of his own. The interaction constrained by the proviso "*generates a set of rights for each person, which he brings to the bargaining table*."[107] The proviso thus "*converts the unlimited liberties of Hobbesian nature into exclusive rights and duties*."[108] According to Gauthier, these rights are *presupposed* by any rational agreement giving rise to morality. Gauthier claims that "*without these rights, persons would not be rationally disposed, either to accept the prohibition on force and fraud needed for market competition, or to comply voluntarily with the joint strategies and practices needed for cooperation*."[109] Rational bargainers will not agree to, comply with, or expect others' compliance with bargains reached from a position other than one constrained by the proviso. If these rights are not respected, it would not be rational to voluntarily make and comply with any agreement.

The problem with bargains reached from an unconstrained position, from Gauthier's point of view, is that the effects of past predations are preserved. The initial bargaining position needs to be adjusted to negate all past forms of advantage-taking, "*all effects of taking advantage must be removed from the initial position*."[110] And as we have seen, Gauthier claims that this constraint is not itself object of any agreement among rational individuals. It is rather a precondition to agreement. Gauthier thus wants to extend the Lockean proviso retroactively, to cover activities that took place before any agreement. I will question whether this is consistent with fundamental justification.

Let's go back to the state of nature. What happens there? As we have seen, violence is not, in general, irrational there. What we will call predatory activity occurs frequently. This predatory activity is then met by defensive activity, and so on. The distribution resulting from this predatory/defensive interaction is called *the natural distribution*, or the non-cooperative outcome. This situation is suboptimal. Predation and defence are both wasteful activities, and the resources spent on predation and defence, could be better used. Now there is a basis for

[107] Gauthier (1986) p 227. "*He brings a right to his person, a right in the fruits of his labour, and a right to those goods, whose exclusive individual possession is mutually beneficial, that he has acquired either initially or through exchange*."

[108] Gauthier (1986) p 209.

[109] Gauthier (1986) p 227.

[110] Gauthier (1986) p 192.

agreement. Both parties (assuming a simple two-person-world[111]) stand to benefit from an agreement that relieved them of the necessity of unproductive predatory/defensive activities. The suboptimality of the natural distribution results from the fact that each imposes costs on each other. Investment in predation leads to counter-investment in defence. Agreement ends this unproductive situation. It yields an outcome where predation and defence are absent. Would this agreement be rationally acceptable to each party? Gauthier's answer is in the negative.

> …clearly an individual would be irrational if she were to dispose herself to comply, voluntarily, with an agreement reached in this way. Someone disposed to comply with agreements that left untouched the fruits of predation would simply invite others to engage in predatory and coercive activities as a prelude to bargaining.[112]

Even if agreement yields an optimal outcome, Gauthier argues that, the effects of predatory (and defensive) activity remain, since each brings to, and from, the bargaining table the fruits of these activities, as part of the overall outcome. An agreement reached "*from the natural distribution therefore does not elicit the rational, voluntary compliance of both (or all) parties, if the natural distribution is in part the result of coercion*."[113] If predatory activity is banned, which it is by the Lockean proviso, then, Gauthier thinks, one no longer has reason to behave in a way that would maintain the effects of predatory and coercive activities that took place before agreement.

Jan Narveson questions Gauthier's reasoning. Why, Narveson wonders, shouldn't both parties accept, and keep, an agreement the outcome of which is better for both? The possibility of this mutually advantageous outcome is *dependent on compliance*. And, as Narveson points out, if the argument is that compliance is impossible, then there is a problem for the principle of constrained maximization itself.[114] The person who carries out her part of the bargain first, lays herself open to non-compliance by the other party, who has now got what he wants. Why should the other party carry out her part of the agreement? Why should she not defect? But "defection" is precisely the attitude that constrained maximization rules out. Narveson asks:

[111] A more detailed version of this example is found in Buchanan (1975).
[112] Gauthier (1986) p 195.
[113] Gauthier (1986) p 195.
[114] Narveson in Vallentyne (1991) p 141.

> Why should things be any different when the Pareto superior situation [a situation which is better for at least some and no worse for anyone] is a move up from an arrangement originally secured by superior force alone? Why, that is, when the situation ex ante [the situation before agreement] was one in which the agreement in which “Morals by agreement” consists is *not* yet in place. Doesn’t that make it out of order to object to the initial depredations as “unfair”?[115]

As Narveson also notices, *at the time* the predatory activity took place it was not wrong, because nothing was wrong.[116] This is the problem from the view point of fundamental justification. Remember that in the state of nature (the situation ex ante) nothing is forbidden, morally wrong or unfair. It is first after the agreement, that such considerations have any bearing. The proviso is justified only when there is a prospect for agreement, remember. In stressing that the proviso must be agreed to, and not just a precondition for agreement, Narveson’s view is more Hobbesian than Gauthier’s. Writing in *Leviathan*, Hobbes said that

> …where no covenant hath preceded, there hath no right been transferred, and every man has right to every thing; and consequently, no action can be unjust. But when a covenant is made, then to break it is *unjust:* and the definition of INJUSTICE, is no other than *the not performance of covenant*. And whatsoever is not unjust, is *just*.”[117]

Does this clash with Gauthier’s the idea that the bargaining situation must be non-coercive? I think not. It is perfectly reasonable to say that the bargaining situation is, by definition, non-coercive, in the sense that there is no coercion going on *in* the bargaining itself. It would not be “bargaining” if coercion was involved in it. If violence, or threats of violence, are (re)introduced in the bargaining situation, it ceases to be a bargaining situation and the parties are on their way back to the suboptimal state of nature. According to Narveson, Gauthier is wrong in supposing that the proviso extends to cover activities that took place before any agreement. Gauthier is right in that bargaining requires that the baseline of non-worsening is respected, but only *from then on*. Gauthier’s retrospective Lockeanizing of his (otherwise) Hobbesian state of nature, is not legitimate, however.

On Narveson’s view, on the other hand, rights are not so much *brought to* the bargaining table. Rather, they arise with, and in the bargaining. When bargaining commences,

[115] Narveson in Vallentyne (1991) p 141.
[116] Narveson in Vallentyne (1991) p 144.
[117] Hobbes (1651) p 95.

the proviso sets in, constraining all future interaction, from that point on. The proviso is, then, if you like, the first item on the bargaining agenda. As I see it, bargaining requires that the proviso is respected, in the bargaining situation, but it is not legitimate to talk of rights until after agreement is reached.

The pure, or Hobbesian, state of nature, where nothing is agreed on, and there are no constraints whatsoever, therefore, must be our proper starting point. Why doesn't Gauthier agree with this? He does explicitly agree that adherence to the proviso is equivalent to Hobbes first law of nature[118], which tells us to seek peace and follow it wherever it may be found, and when it may not, then everyone, "*may seek, and use, all helps, and advantages of Warre*."[119] Gauthier also agrees that the proviso is rational, if and only if, we conceive the state of nature as "giving way to society." Without the hope of passing from the state of nature to society, without the prospect of agreement, "*there would be no morality, and the proviso would have no rationale*."[120] As I see it, if we want to stay within this, broadly Hobbesian framework, we must, somehow, find a basis for agreement on the proviso. But is the Lockean proviso, then, really, what rational beings would agree to in a pre-moral state? As I said above, what makes the state of nature suboptimal is the fact that each imposes costs on others by predatory activities. It therefore seems fairly reasonable that rational persons would want a prohibition on just such activities. And this is just what the proviso does. Narveson argues:

> Left absolutely to their own devices […] people will perform actions that lead to a condition that will make their lives immeasurably worse than if they were instead subject to restrictions: namely, restrictions on just the sort of actions that have that effect.[121]

Is this plausible, or even possible? In order for that to be possible, the social contract must be such that all the parties to it are better off after the agreement, than in the state of nature. In technical language, the state of nature must be suboptimal. Otherwise it would not be rational to abandon it. But might there not be persons who are better off in the state of nature, persons whose preferences are better served in the natural state, than in society? Sure, there might be such persons. For them, it would not be rational to respect the proviso. These people, thereby, decided to stay in the state of nature and they should be treated thereafter. There is no reason

[118] Gauthier (1986) p 193.
[119] Hobbes (1651) p 87.
[120] Gauthier (1986) p 193.
[121] Narveson (1988) p 136.

to respect the proviso in relation to such people. Narveson argues that morality is like a club – “the morality club.”

> Those who join have certain responsibilities and certain rights, and we, the people who run this club, offer a package that we think no remotely reasonable person could really refuse; but nevertheless, some might. All we are saying is that *our* package is such that it must appeal to the widest set of people any set of principles *could* appeal to. Anyone who doesn’t buy our package wouldn’t buy any package compatible with living among his fellows on terms that they could possibly accept.

Gauthier now agrees with this too. In an essay called, “*Uniting Separate Persons”,* he writes that, “*My defence of the rationality of morality must be limited to those persons whose overarching life-plans make them welcome participants in society*.”[122] The assumption here is that the vast majority are such persons.

This, I think, is part of the answer to the question whether the proviso is rational. A fully adequate fundamental justification of rights would also have to answer whether the proviso is uniquely rational. Might there not be a social contract based on another principle, incorporating violations of the proviso? This, second question would probably need an essay of its own. I, however, think that the Lockean proviso could be shown to hold as the basis of rights, given the remarks I made.

Conclusion

A fundamental justification, a justification of a realm that does not appeal to any of the concepts of that realm, has great philosophical appeal and should be abandoned only if known to be impossible. Anything short of fundamental justification is really no justification at all. In order to avoid begging the question, a justification of the moral realm has to be formulated in non-moral premises. This means, in particular, that we are not allowed to assume anything moral, or rely on our moral intuitions, in our conception of rationality, or in the formulation of the hypothetical contract situation. A fundamental justification requires us not to incorporate into our premises any of the moral content that we wish to have emerging in our conclusions.

Another way to characterize our project is to ask how morality can be justified, given that value is subjective (created or determined through preference) and relative (with each

[122] Gauthier (1993) p 189.

person having his own good, which is not part of a single, overall good) and that no categorical imperatives are objectively valid.

What then are the prospects for such an enterprise? Is a fundamental justification of morality possible? This is, admittedly, a very large issue and I do not claim to have a conclusive answer. The only way to arrive at a conclusive answer to this question is to present a fully developed positive account of such a justification. David Gauthier's theory of "Morals by Agreement" is an attempt to do just this. I have discussed a number of objections that might be put forward against such a theory; some of these may be seen as objections to the whole idea of fundamental justification. I think that none of these objections is decisive. And therefore I can see no conclusive reason for thinking fundamental justification *im*possible, or even *im*plausible.

I have tried to provide a very rough sketch, or a very general outline, of what a fundamental justification of morality might look like. The reason for doing this was to show what problems a fundamental justification might have and how, if at all, they could be solved. In order to embark on a fundamental justification, we needed to rely on a conception of rationality as strictly instrumental. This was because we needed something normative, which, at the same time, had to be non-moral. On this view, ends provide reasons for pursuing means, and that these ends are non-moral guarantees that the justification will be fundamental. Instrumentalism thus provides the motivation needed for our theory to be practical. Motivation has been an important concept in my arguments. I wanted a justification not merely an explanation. I could not think of anything else, other than instrumental rationality, which was both normative and non-moral in this required sense.

I discussed different objections against instrumentalism about rationality and concluded that none of these objections defeats it. I have also discussed in what manner we may say that rationality is normative. The normativity of rationality is hypothetical. Hopefully, this hypothetical normativity is sufficient for this kind of justification. It is by no means a necessary requirement that rationality itself be justified fundamentally. I argued that we can plausibly see the rational realm, the realm of deliberative justification, as more basic, or more fundamental, than the moral realm.

Morality involves constraints, and whenever mutual advantage can be achieved without constraint, morality is not needed. The perfectly competitive market is an example of a domain where straightforward maximization suffices to bring about optimal outcomes. Also in iterated games, there is no need for constraints on the individual's pursuit of his or her own interests. Where the market acts as an invisible hand, morality works as a visible hand, constraining each for the good of all. The question is whether constraints can be rationally justified. Is it rational to keep commitments in so- called tight corner situations? Does a rational disposition make the actions manifesting it rational?

Gauthier's answer to this, as we have seen, is positive. The rational person adopts a disposition to cooperate, even in situations where this is not directly maximizing, given that the disposition to comply is advantageous. This is highly controversial but I am inclined to accept Gauthier's solution to this problem on the basis of that there is something strange with the idea that rational agents, defined to maximize, still fails to bring about the outcome that would be better for all.

Finally, I discussed the Lockean proviso, and its role in the justification of rights. I argued that, if we want to stay within a broadly Hobbesian framework, we must, somehow, find a basis for *agreement* on the proviso. I considered whether it is plausible to view the proviso as rational, and not merely as having an intuitionist or metaphysical basis. In order for it to be possible to agree on something as general as the proviso, the social contract must be such that all the parties to it are better off after the agreement than in the state of nature. The state of nature must be suboptimal. I admitted that it might be the case, that there are people whose preferences are such that morality is not rational for them. Our defence of the rationality of morality must be limited to those who benefit by joining the "morality club." And, I assumed, this applies to the vast majority of our fellows.

A fundamental justification has problems. Many of these problems are of an empirical nature. Are agents opaque, so that we cannot know our interactees' dispositions? Are people strongly negatively concerned with each other's interests? etc. I have no answers to these questions, but the assumptions the contractarian makes here are, I believe, reasonable. That much is left for empirical investigation is something I consider an advantage. This gives the theory an empirical testability. Some other problems are of a more philosophical nature; these are (at least some of) the problems that I have discussed. And I can see no reason for thinking that these problems are insoluble. Any theory has problems, and I think that we should be optimistic about the prospects of providing morality with a fundamental justification.

Bibliography

Axelrod, Robert. (1984). *The Evolution of Cooperation*. Basic Books.

Baier, Kurt. (1958). *The Moral Point of View*. Cornell University Press.

Broome, John. (2005). *Is Rationality Normative?*

Buchanan, James. M. (1975). *The Limits of Liberty: Between Anarchy and Leviathan*. University of Chicago Press.

Danielson, Peter. (1992). *Artificial Morality: Virtuous Robots for Virtual Games.* Routledge.

Darwall, Stephen. (Ed.) (2003). *Contractarianism/Contractualism*. Blackwell.

Gauthier, David. (1986). *Morals by Agreement*. Oxford University Press.

Gauthier, David. & Sugden, Robert. (Eds.) *Rationality, Justice and the Social Contract: Themes from Moral by Agreement*. University of Michigan Press, 1993.

Hobbes, Thomas. (1651). *Leviathan.* 1996. Oxford university press.

Hume, David. (1739). *A Treatise of Human Nature*. 1978. Oxford University Press.

Hume, David. (1751). *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. 1975. Oxford University Press.

Locke, John. *Two Treatises of Government*. 2003. Cambridge University Press.

Mackie, John. L. (1977). *Ethics: Inventing Right And Wrong*. 1990. Penguin Books.

Morris, Christopher. W. & Ripstein, Arthur. (Eds.). (2001). *Practical Rationality and Preference: Essays for David Gauthier.* Cambridge University Press.

Narveson, Jan. (1988). *The Libertarian Idea.* 2001. Broadview Press.

Nozick, Robert. (1974). *Anarchy, State And Utopia*. 1995. Blackwell Publishers.

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

Scanlon. Thomas. (1998). *What We Owe to Each Other*. The Belknap Press of Harvard University Press.

Shaver, Robert. (1999). *Rational Egoism*. Cambridge University Press.