14

Categorically Rational Preferences and the Structure of Morality

Duncan MacIntosh

1. Introduction: The Reduction of Morality to Rationality

Infamously, David Gauthier (1986) has sought to reduce morality to rationality. Simplifying (and exaggerating) his position somewhat, he claims moral problems are partial conflicts of interest, game-theoretically depictable as the Prisoner's Dilemma. Here, one agent can do well only if another does poorly, but both can do fairly well by making and keeping agreements to comply with Pareto-optimal solutions to their conflict; in the PD, this consists in making and keeping agreements to co-operate. And co-operating seems to be the morally required action, since it consists in refraining from exploiting another agent for one's own gain. However, in the PD, it maximizes for each agent, no matter what the other does, to defect. In complying with optimal compromises, then, agents must refrain from maximizing their individual expected utilities, something *prima facie* irrational on the standard theory of rationality as maximization. So the other part of Gauthier's reduction is the claim that, if it is rational to adopt dispositions constraining one's tendencies to maximize (as it is in PDs), then it is rational to act from that constraint (act as per the compromise) with other agents inclined to do likewise. Gauthier applies maximization first to choice of disposition rather than of action; rationality then dictates acting out maximizing dispositions.

Both claims have been criticized. It has been doubted, first, whether all moral problems are really PDs, and, second, whether it is there rational to act "morally" – co-operatively. In this paper, I defend the reduction of morality to rationality.

I first introduce claims for which I have argued elsewhere: it *is* rational to co-operate in PDs. For it is instrumentally rationally obligatory to revise the preferences on which one maximizes when about to face a PD so that one would then find co-operating maximizing and so rational with others with (rationally) appropriate preferences.[1] This instances a more general feature of rational preferences: given preferences are rationally obligatory just in case having them maximizes on those held prior to them; and it maximizes on those one has when about to face a PD, to adopt different ones in going in to a PD.[2]

I then claim that PDs do not capture all moral problems, for there are ways to fail to be in a partial conflict (PD) with someone that are themselves morally problematic. To be fully moral, not only must one, when about to be in partial conflicts with fellow rational agents, so revise one's values that one is inclined to co-operate with such agents; sometimes, one's values must be such as to place one either in a pure co-ordination game, or in a partial conflict, both situations where, because of one's values, in order to increase one's own utility, one must increase that of another agent. To fully reduce morality to rationality, it must be shown irrational for agents to have preferences which would fail to put them in PDs or co-ordination games when morality requires them to be in such games. It would then follow that, for agents with fully rational preferences, all remaining moral problems really are PDs, and that since it is rational to be moral in PDs (because rational to acquire "moral" preferences when going in to PDs), morality does reduce to rationality.

I try to achieve this result by analyzing the conditions on having fully rational preferences. I claim one's preferences are fully rational just if they could have been arrived at by a series of preference choices of the sort described above (where at each stage in the history of one's preferences, one had ones maximizing on one's prior ones), from the first preferences one ever had, were they rational. And I argue that there are unsuspected constraints on rationally permissible first values, ones deriving from three facts: preferences are just reasons for choices; in choosing one's first preferences, the factors which normally make it merely agent-relative which preferences are rational either do not operate, or will not permit immoral preferences; and in choosing preferences to have in situations where others too have preferences, one rationally must co-ordinate one's own with theirs to ensure that everyone's satisfy the condition of being able to be reasons for choices (on pain of contradicting the assumption that other agents had certain preferences in the imagined situations). These constraints impose on rationality in application to first values, a character like that of the categorical rationality envisioned by Kant. And the only preferences rational on these tests are moral ones. Thereafter, in any later choices of values given first values, if all agents are rational, all will always arrive at moral values.

2. Instrumental Rationality and the Rationality of Preference-Revision; PD Co-operation Rationalized

Many philosophers doubt that it is rational to act from the constraint Gauthier advocates for the PD.[3] I have tried to solve this problem by arguing (1991a, 1991b, 1991c) that it is rational to revise one's preferences so as to value keeping agreements which resolve partial conflicts; then, since one's values have changed, even if one maximizes on one's (new) values (as one should as an instrumentally rational agent), one will keep the agreement if the other agent has similar values.

This is possible because – as I argue in my (1993) and (1992) – the structure of hypothetical or instrumental rationality allows it rationally to evaluate the ends instrumentally served. Thus, we may speak not just of the rationality of a choice of actions given one's preferences and beliefs, but of the rationality of one's preferences given that it is rational to do things – possibly including revising one's preferences – that would help cause the conditions targeted by one's preferences. Since sometimes other people will act to help cause the targets of one's preferences just if one were to change those preferences, it can sometimes be rationally required that one adopt different ends. For example, in a paradoxical choice situation (PCS) like the PD, others will help one to reduce one's jail time (the aim of one's original preferences) only if one stops caring only about that, and comes more to prefer to keep agreements.

More technically, hypothetical or instrumental rationality consists in maximizing (one's individual expected utility as defined) on one's preferences given one's beliefs. But rational agents would use this principle to criticize and revise the preferences on which they would maximize; they would ask whether them holding the preferences they now hold maximizes on those preferences. Where it would not, as in PCSs, the agents would revise their preferences in whatever ways would be maximizing on their original preferences.

3. Moral Problems and the Prisoner's Dilemma

So if agents are about to find themselves in PDs, it is rational for them to acquire values which would rationalize co-operation, and then to co-operate accordingly; and were all moral problems merely PDs, and if being moral simply consisted in co-operating in them, then morality would reduce to rationality.

But are all moral problems partial conflicts of interests? It might seem so; for surely if there is no conflict in our interests, there are no moral objections to us each acting on our interests – neither of us harms

the other in the sense
While if there is a tot
ally arbitrary to requ
to the other; so the
interests are either
flicting, since it seem
ter, and since it is
to rationality.

But what deter
ues, our circumst
are in a total confl
innocent; I have
prefer that they
if so, surely I can
we want? It may
we are in a total
your help in som
need not fear yo
weak. Surely th
our situation ca
conflict, but th
to my prefere
want be satisfi
guardian, I co
isfaction of m
your money
against the
women wh
others – the
something
partial con
profit mea
ference wi
not having
must offer
innocence
agents If
do anythi
But our
Someti
require
must be
should

the other in the sense of preventing him from advancing his interests. While if there is a total conflict in our interests, it would surely be morally arbitrary to require one of us to yield any advantage we may have to the other; so there can be no moral duty to do so. Thus, since our interests are either non-conflicting, totally conflicting, or partially conflicting, since it seems there can be a moral issue only if they are the latter, and since it is rational to compromise here, surely morality reduces to rationality.

But what determines whether we are in a partial conflict? Our values, our circumstances of choice, and our powers to act. Suppose we are in a total conflict because, while you have values *prima facie* morally innocent, I have ones defined this way: whatever your preferences, I prefer that they not be satisfied. Surely I may be malevolent here; and if so, surely I cannot claim it morally indifferent which of us gets what we want? It may be that I have morally problematic values. Or suppose we are in a total conflict because, while if I had to fear you, or needed your help in some project, I would find it reasonable to compromise, I need not fear you, or do not need your help, because I am strong, you weak. Surely the mere fact of our power asymmetry should not mean our situation cannot be morally problematic? And suppose we have no conflict, but the reason for this is that I raised you to be a willing slave to my preferences: whenever I want something, you want that my want be satisfied. (For example, you are the heir to a fortune, I am your guardian, I covet your inheritance, and I raise you into wanting the satisfaction of my preferences, and so cause you to want to give me all your money when you come of age; or think of feminist arguments against the self-oppressing patriarchal values of "real" women, women who value only or primarily the satisfaction of the values of others – their men and their family members.) Surely there could be something morally problematic about this way of there failing to be a partial conflict? Further, it is one thing for my rational actions not to profit me at your expense, for my acting on my goals not to be an interference with you acting on yours. But it often seems insufficient for one not having been immoral that one not have interfered; sometimes, one must offer help, even at one's own expense. This connects with moral innocence not necessarily resulting from power asymmetries between agents. It could happen that, given my powers and values, I need not do anything that would enhance your utility in order to enhance mine. But our intuitive morality requires agents to have some fellow-feeling. Sometimes my values must be such that, even if no action of yours is required in order that my values be advanced, some of your values must be advanced in order for some of mine to be advanced. That is, I should care, at least a little, about the fate of those affected by my actions and omissions – and this amounts to an obligation that one's values not fail to put one in a PD with another agent just because one's utility does not depend on his actions; rather, one's values may only fail to put one in a PD either because, given his values, your utility and his are independent, or because, given your values and his, your utility is dependent on his utility.

So we can have a moral problem not just if there is a partial conflict, but also if the reason we have, instead, a total conflict, is the malevolence of an agent, or the excessive power or weakness of one of the parties; or if the reason we have, instead, no conflict, is the unsavoury origin of the values of one party; or if, while we have no conflict, one of us has a duty to aid the other, to confer a benefit on someone even where our own utility is not (unlike in the partial conflict situation of the PD) even partially contingent on his actions – translating into values, we may have duties to have our utility be partly conditional on the utility of others.

So to fully reduce morality to rationality, we need somehow to show that there are rational constraints on the contents and origins of people's values. If we could show it was irrational to have malevolent values, and to have slavish ones, and to fail to have inclinations to help another where that would mean little cost to oneself, we could say it is irrational for people to be in total conflicts due to malevolence, or in non-conflicts due to value enslavement, or to find themselves without a kindness inclining them to aid the needy; and that, if, after our values have passed these tests, we are still sometimes in a partial conflict, it is rationally obligatory there to compromise as per Gauthier's innovation to the theory of rational choice given values, or mine to the theory of the rationality of values given values. Morality then *would* reduce to rationality.

So how can we assess the rationality of our values? We might apply my criterion of instrumentally rational preference-revision to the history of our each arriving at our current values. But we may each have made rational choices from values with which we were first endowed by nature or rearing. And unless something can rationally criticize *those*, our current values may be of the problematic sort, and yet still rationally permissible. I might have been born with or raised to malevolent or stingy values, or you to slavish ones; or events might have made it rational for us to have arrived at evil values, whatever ones we began with.

The only hope of reducing morality to rationality, then, is to find a way of rationally evaluating agents' first values. For it is from the vantage of these first values that agents choose their later values; and it is from the vantage of these later values that they choose their current values. So our first values determine, by way of a series of rational choices of values given prior values, the values we now hold (and determine their rationality). The reduction project requires it to be irrational to have first values morally problematic (assuming everyone else is to have rational first values), and irrational, assuming everyone else is rational, to move to values morally problematic.

I shall now try to get this by deriving rational constraints on first values from the principles of rationality applied to the special case of first values, and from reflection on the metaphysics of values *qua* reasons, reflection, in particular, on the nature of preferences. I shall argue that, applied to the choice of first values, the metaphysics of values *qua* reasons give rationality a character akin to the categorical rationality articulated by Kant.

4. Hypothetical Rationality and the Rationality of First Values

It is sometimes said that morality is indifferent as between persons, being universal, impartial and fair, while rationality is particular as between persons, being individual, partial, and self-serving. But in fact, rationality is indifferent as between persons too; it is just that it is not indifferent as between values: what is rational for someone in a given circumstance depends on what she values. Different people with different values, in the same circumstances, will find different actions rational. However, people with the same values, in the same circumstances, will find the same actions rational.

Still, the whole reason for doubting that morality could reduce to rationality is that, while one's moral duties hold categorically, independent of one's values, one's rational duties seem to hold only hypothetically, depending on one's values. And it seems it could only be a lucky accident if the actions recommended by hypothetical rationality to a person given her values were to be the same as those obligated by morality. For suppose some behaviours to be morally required independently of rationality (e.g., by the requirements of universality and impartiality, or of equal concern and respect for all affected parties); and suppose rationality, at least instrumental rationality, to be some preference-sensitive function (like the maximization function) from preferences to choices of behaviour (preference-sensitive in that the actions it recommends vary with preferences): then for any such function, it is possible to specify preferences which the function does not take over into moral behaviour. And so agents who happen to have such preferences will not find it rational to behave morally.

There are only three ways out of this: revise our conception of what is morally required; find further constraints on the rationality of preferences; or show that the functions determining the morality and rationality of actions are co-intensional. I shall develop the latter two options, deriving the third from new arguments for the second.

The project of reducing morality to rationality was never to show that, given arbitrary values, moral conduct is rationally obligatory. Indeed, no one has ever held this. Hobbes and Gauthier are sometimes wrongly believed to have held it. But Hobbes does not describe morality as emerging rationally from arbitrary values. Rather, he assumes that people have selfish values, but not malevolent ones. They care about themselves, but they do not have values logically (definitionally) able to be satisfied only by the non-satisfaction of the values of another agent. They are, if you will, not malevolently tuistic (where to be tuistic is to see intrinsic value in the welfare or illfare of others, to be altruistic, to see it in their welfare, malevolent, in their illfare). Meanwhile, Gauthier thinks morality emerges from rationality for agents assumed to have non-tuistic values, values (again) not *defined* as targeting the satisfaction or non-satisfaction of the values of others. (He was trying not to assume that agents had fellow-feeling, which assumption would have trivialized the reduction of morality to rationality, but only by basing it on a mere contingency). But in stipulating non-tuism, he in effect helped himself to the assumption that agents do not value each other's illfare. Indeed, it is because these authors merely assume these constraints on values, rather than showing the rational obligatoriness of one's so restricting one's values, and of having somewhat altruistic values, that their projects of reducing morality to rationality fail.

The closest one comes to a philosopher who thinks he can motivate arbitrary agents, with arbitrary values, into acting morally, is Kant. And he only got this by stipulating that agents had the power to go against hypothetical rationality given their values, and to choose instead by principles they could will to be laws of nature. (Or rather, he claimed it was a condition on agents being moral agents, ones able to do the right thing simply because it is right; if agents could not so choose, they were just machines controlled by the values given them by nature and culture – they did not properly choose at all, and so their behaviours, not really being actions, would not be subjectable to moral evaluation.[4] But he also argued that actual agents *can* make categorically rational choices, by arguing that the structure of nature might accord with that of noumenal choice.) In effect, then, while he took agents as he presumed they were and claimed they had motivations to morality, this was not provided by their values, which, therefore, serve as idle cogs in the generation of moral behaviours.

The more plausible hope of a reduction will not consist in rationally deriving morality from arbitrary preferences, but from the conditions on the possibility and rationality of preferences. There are two reasons one might not think this possible. First, since Hume, it has been thought that the only rationality constraint on one's values is that they be consistent with one's beliefs: if you value A, and believe doing B will get you A, *ceteris paribus*, you must come to value doing B. That apart, rationally, you may value *anything*. If, however, I was right that it can be rationally obligatory to acquire certain preferences when that would advance one's prior preferences, then there is a further constraint on rational value: if one values X, and if coming to value Y for itself would help to get you X, you must come to value Y – see my (1992). But furthermore, I shall argue, the very concept of a value – and so of the values which can figure in arguments for the hypothetical rationality of values given prior values – can constrain the rationality of *initial* values, of values given *no* values. For it is in the concept of preferences that they must be able to serve as reasons in the circumstances in which they are held, which in turn requires that there be actions there available to agents which these values could give them reason to do. And this constrains rationally willable first values: you cannot will to have those that no action could serve in the circumstance of holding them.

The second reason one may not have thought it possible to derive morality from rationality is that while morality has a universalizability requirement, rationality seems not to. But in fact, both do. What made hypothetical rationality particular to individual persons, rather than indifferent as between them, was that it was sensitive to values, which vary across persons. But in a rational choice of values made prior to having any values, rationality becomes impartial to individual persons, indifferent as between them, because there is not yet anything for it be partial to a person by. The principles determining the rationality of actions and of new values are only *non*-indifferent to persons so far as persons *have* (possibly different) values. The principles are only particular to or dependent on values, and thereby persons, for choice situations where the principles operate given people's prior values. But if I am right on the first point, that rationality also operates where an agent has no prior values, we get the surprising result that if it here recommends anything to one agent, it recommends the same to all agents: it is indifferent to persons because, in this condition, persons do not have different values to make rationality yield different enjoinments for them. So rationality has a universalizability requirement in the situation of choice of first values: it is only rationally permissible to will to have values one could will everyone to have in the same circumstance. This requirement follows from the one normally operant even in standard applications of instrumental rationality to the choice of values given values: values can be ones I rationally should adopt given my prior ones in my circumstances, only if they would be ones rational for anyone to adopt had they similar prior values, and were they in similar circumstances. Now, if the antecedent of this requirement is false, as it is for everyone when they have no prior values, then everyone, prior to their having values, is in the same circumstance, that of having no values. So whatever rationality would there recommend for one valueless person is what it would recommend for any – indeed, every – valueless person; thus, a value can be rational for one person only if everyone could rationally have it. And I shall argue that values which would pass these constraints of rationality in this circumstance would be moral values, ones rationally inducing of moral behaviour.

We shall proceed, then, in two steps. First, we show that rationality constrains first values. Second, we show that values are choosable by oneself as one's first ones, only if so choosable by everybody; in choosing one's first values, one must ask if everyone could have them, i.e., one must, in effect, choose for one's self only values one could still choose if one were choosing that everyone have them. The only values which will pass this test will be moral ones.

5. On Rationally Possible Values

It might be thought that making a rational choice of first values is impossible. For in choosing without prior preferences, one has none to use in deciding which to adopt. But there are other constraints on rationally appropriate first preferences. To see these, we must consider what preferences are, what role they play in rational choice.

Preferences are just reasons for choices; they rationalize the choices believed to make most probable the conditions they target. A preference is characteristically a reason for choice of action, but it may also be a reason to *be* a certain way, e.g., to have some other preference; for sometimes, as we saw for the PD, the very having of a new preference can advance the targets of ones now held. It is then rational, given the old preferences, to supplant them with the new. Here, we might speak of the old as being reasons for choosing the new. And for simplicity, we shall count revising, adopting, having, and keeping a preference as possible choosable actions – see my (1992).

That a preference is a reason for choice, and that the having of a current preference cannot be rational if having some other one would better advance the former's target, are connected: having a preference is rational only if having it advances its target. Normally, a preference has this effect by rationally motivating its holder to act so as to make the obtaining of its target more likely. A preference can do this only if its target is one for which, in the circumstances where the preference is held, there is some action available to its holder whose performance would advance its target. That is, a preference can do this only if it is "actionable." So, normally, a preference for some target can exist only if, when the preference is held, there is some action it then gives its holder reason to do, one that would advance that target. And that a preference is, normally, nothing but what can play this role means that, normally, one can only prefer targets some action available to one in the circumstances can help make obtain. (So it is not enough for a preference now to be rationally permissible that its target is one some action would advance in some other possible circumstance than that in which the preference is actually held; rather, what you can prefer varies with which possible targets are such that you can now do something to advance their obtaining – see my (1994).) Summarizing: a preference is just a reason for a choice; so one can only have a preference for conditions which are such that, when preferring them, there is some action one could do that would better the odds of those conditions' obtaining (would "probabilify" those conditions), so that, in preferring those conditions, one has reason to do that action.

But it is also possible for you to have a preference you cannot act on, but which I will act on for you. For recall that being a certain way, namely, *having* a certain preference, can sometimes be a way of causing states – e.g., in PCSs, you having certain preferences induces others to act so as cause certain states, as when you acquire the preference to co-operate in the PD in order to induce others to co-operate with you, those who will co-operate with those who prefer to co-operate too. So a preference you cannot act on yourself remains one you can have if you having it makes it likely that someone else will advance it for you. Here, you merely having or adopting that preference is an "action" which advances its target, since it gets another agent to advance its target; and so the preference is permissible. We can assimilate a preference's having, in order to be rationally permissible, to motivate its holder or someone else to advance its target, into one compendious condition on the rationality of a preference: a rationally permissible preference must be "self-advancing," or "self-maximizing" when held.

There are two other constraints on the preferences which might be rational as one's first ones. One derives from the fact that some states of affairs are impossible (either inherently, or given the circumstances): since no action can probabilify an impossible state, and since preferences are only individuated by which actions would probabilify their target states, there can be no preferences for impossible states – see my (1993).

Because one cannot prefer impossible conditions, nor (as we saw earlier) possible ones not probabilifiable by any of one's available actions (not even the "actions" of merely having some preferences), excluded as possible targets of preferences are things like the past, the contradictory, the known-to-be-contrary to fact, the already-known-to-be-a-fact, and anything contrary to or ineluctable given natural law. (Unless it is one's preferring it that makes it ineluctable; the point is, you cannot prefer anything whose likelihood your motivated actions cannot increase. But because your actions *can* sometimes make a difference, determinism would not mean nothing can be preferred.)

The final constraint derives from the preceding one: since were there to be preferences for impossibilia, they could not be satisfied (for what satisfies a preference is the obtaining of its target state, and an impossible state is one that cannot obtain, so a preference for it could not be satisfied), not only can one not prefer impossible states, one cannot have impossible-to-satisfy preferences. This also follows from how having a preference relates one to its possible satisfaction. Say you prefer that condition *x* obtain. For *x* to obtain, given that you prefer this, is for the preference for *x* to be satisfied. So in preferring *x*, either you are also preferring that the preference for *x* be satisfied, or once you saw that to get *x*, you had to satisfy the preference for *x*, rationally you would prefer its satisfaction. But then you cannot have an unsatisfiable preference. For in having one, you would be preferring its satisfaction; it cannot be satisfied, so you would be preferring the impossible; but you cannot prefer the impossible, and so cannot have that preference. A necessary condition of having a preference, then, is that it is at least possible that it be satisfied in the circumstance in which it is held.

So we have found four constraints on the possibility and so rationality of a preference, first or otherwise: it must be self-advancing to have it, it must be actionable, it must not target the impossible, and it must be satisfiable.

We may now apply all this to the rational choice of first values. Perhaps someone with no preferences would have no reason to come to prefer anything. Still, *were* he to form preferences, for them to be rationally permissible they must meet our four constraints. And we shall take the question, which first preferences is it rational to have? as the question, given that one has no preferences, if one *were* to form some, which would it be permissible to form after all the irrational or impossible ones are ruled out?

If we combine the constraints on first preferences with its being rationally obligatory to maximize on one's preferences, and with my extension of this into the rationality of preferences given prior preferences, we have a general theory of rationality in values. If one has, as yet, no preferences, it is rationally permissible to adopt any ones, in any combination, provided they jointly meet our four conditions. Thereafter, it is rationally obligatory to revise one's preferences in whatever ways would be maximizing on them (likewise for one's

revised preferences
constraints; and it w
preferences provid
not be anti-maxim
current preferenc
missible first prefe
mizing accretions
rationally permitt

6. Catego
for

Supposing then
ones that they be
held, we turn to
if actionable and

Recall that o
nality is that m
ticular and hyp
particular to th
as before acqu
there may be
cal. So if it is
certain first va
one. But then
to have certa
rationally w
rationally w
on rational

But this
in the sam
not mean
ing them
actionabl
which m
ues. Ag
which v
ers give
by the
becaus
ration
ferent
agent
ferent

revised preferences, and so on), provided the revisants meet the four constraints; and it would be permissible at any time to acquire any new preferences provided they meet the constraints, and provided it would not be anti-maximizing on preferences one already has. Thus, one's current preferences are rationally permissible just if derived from permissible first preferences by maximizing revisions and non-anti-maximizing accretions. Rational actions, in turn, are ones maximizing on rationally permitted current preferences.

6. Categorical Rationality and the Values Rational for Me Only if Rational for Everyone

Supposing then that it is necessary to values being rational as one's first ones that they be (among other things) actionable and satisfiable when held, we turn to the second matter: showing that they are rational only if actionable and satisfiable if all agents had them as their first values.

Recall that one reason it seems morality cannot be reduced to rationality is that morality is universalized and categorical, rationality, particular and hypothetical. But rationality is like this only because it is particular to the individual values of agents. If agents have no values, as before acquiring their first ones, then such rational constraints as there may be on their first values are neither particular nor hypothetical. So if it is rationally required or permitted for anyone that she have certain first values, it is also rationally required or permitted for everyone. But then it can only be rationally required or permitted for anyone to have certain first values if it can be so for everyone; one can only rationally will that one's self have a certain first value if one could rationally will that everyone have it. This constitutes a fifth constraint on rationally permissible individual first values.

But this may be too quick. That agents choosing their first values are in the same situation *value*-wise (that of having no values as yet), does not mean each agent rationally must choose her values as if by choosing them for all agents. For even if agents could only rationally choose actionable and satisfiable preferences, it is false that the only thing which makes rationality particular to persons is their individual values. Agents also differ in their powers; and so surely they differ in which values they can have, because they differ in whether their powers give them an action to advance some goal which might be targeted by their values. Thus, values which may be rational for me to adopt, because my powers would let me advance them, might not be ones rational for you; you might be too weak to advance them. But then different values are rationally permissible for different agents. So if the agents knew their individual powers, they could, rationally, choose different first values; I can rationally will to have certain values, without having to be able to will that everyone have them. But then, if I knew I was strong and well-situated, surely I could adopt, say, malevolent values, confident of them being actionable for me. The constraints of rationality would not filter out immoral values.

So to keep faith with our moral intuitions it seems we must deny agents such knowledge in choosing their first values. Our reduction project must show it rational for agents to choose their values indifferently to their individual powers and circumstances. But how can we justify this using only rationality as the test of proper first values? How do we get from one's choosing values prior to having values, to choosing ones prior to knowing one's circumstances and powers?

Well, persons are individuated by their values; so if a hypothetical first-value-chooser is valueless, it is no given person; and nor, then, does it yet have such and such powers, in such and such circumstances. So it cannot help but choose values in ignorance of such features of determinate persons. One's identity as a person is given by the rationally self-updating conating cognizer one is. If no one yet has any conations (as where all agents are choosing their first values), no one yet has a personal identity; no one is yet you, for example. So no one is such that the principles rationalizing values figure with partiality for you. So, that some person will have special powers does not make it rationally permissible for an *ex ante* chooser to pick values which, were she the powerful person, would mean she had satisfiable values, though a weaker person would not. For no chooser of first values knows she is especially powerful.

But in imagining yourself choosing your first values, if you are no person then, in what sense are you choosing *your* values? Well, it is not that you are no one, but that who you are is indeterminate as between the agents for whom you are choosing values. Thus, your choice problem is to choose values rationally appropriate no matter which of the people who will come to exist by the infusion of values will be you.

This makes the choice of first values a problem whose solution requires agents to imagine the choice from an original position behind something like a Rawlsian veil of ignorance about the identities of those who must live with the values chosen. For Rawls, this was just a thought-experimental apparatus for elucidating the concept of fairness, something of no interest to, and with no power to influence, those currently indifferent to justice – someone could make Rawls' calculations, and yet not be rationally required to change his values or behaviour. But our original position represents the vantage from which is decided the rationality of one's first values, and so of all later ones derived from them. And so it speaks to all rational agents, not just ones who already value, say, justice. An agent whose current values could not have been derived from first values chosen from our original position, would have irrational values.

So the question is now, which preferences is it rationally permissible to prefer people to have, given that you know no particular facts about your powers and circumstances, but only general facts about everyone's? You know about logic, laws of nature, the different powers of different agents, and so on, but not about whether you are an agent with such and such powers.

As we argued above, it is only rationally permissible to prefer to have satisfiable first preferences. To decide which ones are rational, then, we must consider what makes a preference impossible to satisfy, thence to see what it is impossible to prefer. We saw that one cannot have preferences for things one can now do nothing about. But there are also social limits on what one can prefer, on which, more shortly.

Another reason morality has seemed irreducible to rationality is that morality seems to have a determinate content computable from the moral requirement on everyone to treat everyone with equal concern and respect; but no such content seemed present in the requirement that one have had rationally permissible first values. For it seemed that, prior to one's having any values, there are none for instrumental rationality (long thought to exhaust practical rationality) to permute into a determinate recommendation for choice, neither for choice of values, nor of actions. Without given values, the duty of instrumental rationality – to advance one's values – seems empty. But it has proved wrong to think rationality contentless where one has no values; for there are constraints on rationally choosable first values: to value x is to have a reason to act, and to value satisfaction of the value for x; thus, x can be valued only if the preference for x is actionable and satisfiable, and so the first values one may will to have must have these properties. If a choice of first values is rational for a given person only if rational for all persons (which follows from all first-value-choosers being in the same predicament), and if, to be rational for one person, the values must be actionable and satisfiable for him, then they can be rationally havable for him only if they would remain actionable and satisfiable for all persons were all to have them. But now we have determinate constraints on the content of rational first values. Values are rationally permissible just if, were all agents to have them as their first values, all would find them actionable and satisfiable. So a rational agent can only will to have as his first values, ones he could will that everyone have; and in so willing, he is willing that everyone have ones everyone could act on and satisfy should every person have those values.

This gives us a formal structure for rational values, one very like that Kant proposed with his test for categorically rational (and so possibly morally obligatory) actions: for him, an action is categorically rational just if you could will its motive principle to be a law of nature, one all agents followed in like circumstances. For us, a value is rational just if you could will that all agents have it, which they could only if all could then act on and satisfy it. Both tests work independently of the values agents happen now to have; and for both, something can pass them for one agent only if it could do so for all. Both yield principles of rational choice appropriate no matter who you are, nor what your circumstance, principles categorical in holding no matter what, rather than hypothetical in holding only on hypothesis of certain givens which may vary across individuals and circumstances. But while Kant's test seemed imposed on the motivations of agents, ours derives from what it is to *have* a motive.

But now we must show how our criterion entails the rational impermissibility of the values which posed difficulties for Gauthier's project.

7. Values Rationally Permissible and Impermissible

We identified several values as inconsonant with morality in morally criticizing Gauthier. To complete the reduction project, we must show the irrationality of malevolent values (ones logically defined as aiming at the non-satisfaction of the values of others), slavish values (ones aiming only at satisfying the values of others), bullying values (ones inclining one to profit at the expense of others), and stingy values (ones inclining one to withhold aid to another, even where it would involve little cost to oneself). Are these values irrational by our measure?

The question what values to have, asked prior to having any, must be answerable the same way for everyone. So I can only say I ought to (or may) have such and such values, if I can say everyone ought to (or may). Values are reasons for action. A proposed value is only such a reason if some action is such that doing it advances that value. If the value is for an impossible state, there is no such action, and so that state cannot be the target of a value. Suppose (as we just argued) that values can only be those one may have if everyone can have them. Suppose everyone can have them only if, if everyone had them, everyone would have an action to advance them. At least two sorts of ends are ruled out by this test: the end of having (only) others' ends attained (the end of slavish values, those feminists bemoan in "real women"), and the end of having the non-attainment of others' ends (the end of malevolent values). The first fails because if everyone has as their only end, the obtaining of others' ends, no action advances anyone's ends; it is impossible that the state obtain that everyone's ends so defined are satisfied, because unless someone has ends defined independently of the attainment of the ends of others, no state of affairs is described as an end as such, and so no action is made such that one has reason to do it by virtue of its probabilifying some state. The second fails because, if everyone wants that no one (else) get what they want, again, their wants do not combine to define a state such that some choice could help procure it. Unless one of us wants something other than that others not get what they want, no state is one everyone wants not to be brought about, and so again, because of the circular interdependence of these values, they are not actionable, nor satisfiable.

That leaves bully and stingy preferences. Bully preferences are ones to profit at the expense of another. If everyone so preferred, would everyone have actions to advance their preferences? No. For everyone can profit only if no one is successfully bullied; for if you are bullied, you have been deprived of possible profit – impossible if everyone profits. So it is impossible for everyone to advance their preferences if everyone has bully preferences (for there is no such state as the one in which everyone's bully preferences are satisfied). And it is rationally permissible to have bully preferences only if, if everyone had them, everyone could advance them. They could not, so you may not have them.

But what of stingy preferences? These are ones not to give help, even at little cost to oneself. Could everyone prefer to refrain from helping? Would everyone's so preferring be consistent with everyone's having an action available to advance their preferences? Let us consider. To have stingy preferences is to prefer to withhold aid from those who need it to advance their values. They only need it if they cannot advance their values without it. If you have stingy preferences, either others have preferences they can only advance/satisfy if you help, or they do not. If they do not, your preference is idle. If they do, your preference is advanced/satisfied only if theirs is not. But a distribution of preferences among agents is permitted only if the preferences so distributed are co-advanceable and co-satisfiable. That is false of distributions featuring non-idle stingy preferences. So in any situation where a moral problem could arise, i.e., where another's welfare (*qua* preference satisfaction) is at risk, you cannot have such first preferences.

It would seem, then, that it is rationally impermissible to have as one's first values, malevolent, slavish, bully and stingy preferences; and since these exhaust immoral preferences, it has proved rationally impermissible to have immoral preferences. The requirements on a satisfactory reduction of morality to rationality are met for first preferences.

8. On Values Rationally Permissible for Me Given the Possible Values of Others (Objections and Possible Renovations)

The argument (in Section 6, above) from the identity conditions on persons to the notion that all agents are in the same predicament in selecting first values, and so to each agent's representing all agents in such choices, is pretty shaky. Can we dispense with it? Perhaps. Here is an argument I build on in my (1994). Suppose your identity is determinate when you choose your first values, so that you know your powers and circumstances. In acquiring preferences, you are forming a preference-ranking of every possible state of affairs. Some of such states contain other people. People have preferences. The conditions on preferences possible for them are the same as for you – their preferences must be self-advancing, actionable, must not target the impossible, and must be satisfiable. Your preferences determine how you will use your powers, and this may lead you to do things that would make certain preferences of others unactionable, impossible of satisfaction, and so on, contradicting the supposition that they have those preferences. So you cannot prefer conditions in which they have preferences which would not satisfy the conditions on possible preferences given the preferences you have, given your powers, and given the actions these two things will conspire to make you do. Even knowing your powers, then, the preferences you can will to have are limited to those action on which would not make impossible others having the preferences you suppose them to have in the situations as you preference-rank them. Rationally, you must limit the preferences you choose for various situations to those action on which is compatible with other agents' preferences in those situations remaining actionable, satisfiable, etc. But then you cannot rationally will to have stingy, bully, or malevolent preferences. For the obtaining of their targets requires that other people have certain preferences, but, paradoxically, have them when, due to the actions your preferences would rationalize, the conditions needed for their preferences even to be possible, would fail. That is, to will to have such preferences would be to will to prefer the impossible, which is itself impossible.

So no matter how powerful you (know you) are, it is rationally impossible for you to have any of those three sorts of immoral preferences, since they are preferences both that other agents have certain preferences, and that the conditions required in order for them to be able to have such preferences not be satisfied. Since that is an impossible compound state, and since one cannot prefer the impossible, one cannot will to have those immoral preferences.

This approach is also better in another way. For arguably the universalizability test for rational first values was more strict than was appropriate given our analysis of the nature of preferences and of the implications of this for which ones are individually rationally permissible. Our analysis showed that a preference is possible (and so rationally permissible) only if actionable and satisfiable. This does not entail that a preference is rational only if it would remain actionable (etc.) were all to have one like it. It only shows that one cannot will both that one have certain preferences in certain circumstances, and will that others there have preferences unactionable and unsatisfiable given one's proposed preferences and the consequences of one's holding them in those circumstances.

But there may be trade-offs in this approach. True, one need not restrict one's preferences to ones everyone could jointly have; one need only restrict them to ones consistent with those others are imagined to have in the circumstances in which one proposes to have certain preferences of one's own. But then are not *prima facie* morally problematic values rationally permissible? For could one not will to have the values of masters in situations in which others are stipulated to have the values of slaves? If so, either we must repudiate that part of our morality, developed by feminists, which holds it morally problematic for people to prefer nothing but the satisfaction of the preferences of others; or we have not found a way to perfectly moralize values from rationality alone.

Fortunately, such values would not be rationally permissible, at least on a worst-case reading of their content; for they would not be co-satisfiable. Suppose slaves have the slavish values discussed above, masters, the malevolent ones: then we have, again, circularity. I am the slave, you the master. I want that you get what you want; you want that I not get what I want. But unless one of us has an independently defined want, nothing is such that you want it and I want you to get it. And there is in this something like the liar paradox: if I want that you get what you want, and you want that I not get what I want, then you want that you not get what you want. To get what you want, you must both get it and not. Impossible. So even by this test, one cannot rationally will to have either of malevolent or slavish values in situations where other agents are imagined to have the other value. Still, it seems possible to will to have slavish values where others would have otherwise morally innocent first-order values. Fortunately, I think I have a way around this, but I must leave it for another occasion, namely, my (1994).

9. The Reconciliation of Morality and Rationality

I argued that certain values are rationally impermissible. But the foregoing is also informative on which ones are permissible, and on which moral system these would constitute. You may have any preferences provided if everyone had them, everyone could advance and satisfy them; and agents may differ in their preferences, provided given those they have, all may advance and satisfy them. Some values are such that a given agent can only pass these tests in having them if her having them happens under a co-ordination constraint with other agents, such that the resulting pattern of values among agents consists of co-satisfiable (etc.) values. And in some patterns, some of the values in them are satisfiable just if some of the value-holders in the pattern have values inclining them to help some of those with certain other values. For example, one agent can only have the desire, as a handicapped person, to live a relatively normal life, if non-handicapped agents desire to see that the handicapped live such a life, even should they need help in this. So that pattern of distribution of rational first values in a community must be such that the several values of the agents in effect implement social arrangements and goods distributions akin to those which agents would choose in something like Rawlsian ignorance (thence to ensure the actionability of the values of the least powerful agent). And while it is impossible that everyone have logically tuistic benevolent values, everyone *can* have ones consistent with those that citizens of a Rawlsian-chosen society would have; for such citizens would have various individual projects, but would share the project of having arrangements advantageous to the least well-off agent. Agents could still get into PDs, for sometimes, by compromise in pursuing their several life-projects, the agents could enjoy a co-operative surplus; and our previous results on the rationality of revising one's values when about to face a PD will guarantee that agents behave non-exploitively when in the PD. But it would not be possible for agents to fail to be in PDs as a result of the malevolence of agents' values, nor as a result of bullying or stingy inclinations. (And nor would it be possible as a result of agents having slavish values, at least not if the universalizability test works; or failing that, at least not if I can defuse the objection to the co-ordination test that it leaves slavish values rationally permissible.) So we have here an argument for the co-intensionality of the notion of justice as fairness, and the notion of practical reason as maximization on individual preferences, where all preferences must be actionable and satisfiable; for the two notions unify at the point of rational choice of first values.

Rationality proves normatively thick, like a moral system. But its "content" derives from the structure of all valuing and choosing, not from some arbitrary conception of what is truly valuable. It does by deduction from the nature of value and the principles of instrumental rationality what moral systems do by deduction from specific value premises.

Acknowledgments

For helpful comments, my thanks to my colleagues and students at Dalhousie University, and to my co-participants in the conference of which this volume is the proceedings. They have made me aware of many possible objections, ones to which I cannot here reply, given the limitations of space and of my ingenuity. But I hope I say enough to make the project seem worth exploring. I enlarge on and defend it in my (1994). The writing of this essay was supported by a research grant from the Social Sciences and Humanities Research council of Canada.

[1] Actually, it is on
interactive oppo
when facing age
values to recipro

[2] These preferen
making the age
revised their pr
defined relativ
erences incline

[3] For arguments
losophers with

[4] For a good in
pp. 693–709.

Gauthier, David
MacIntosh, Dun
theoretic
problem
——— (1991a)
ty of mor
——— (1991b)
lemma
——— (1991
Studies
——— (199
dilem
——— (199
ty
——— (19
tiona
——— (1
Unp
Solomon

Notes

1 Actually, it is only rational to co-operate in PDs where one has had a pre-interactive opportunity to undergo alterations in one's values, and only when facing agents whom one knows will likely be caused by one's altered values to reciprocate co-operation. See my (1991a, 1991b, and 1991c).

2 These preferences need a rather complicated structure in order to avoid making the agent vulnerable to exploitation from those who have not revised their preferences, and in order for his preferences not to be circularly defined relative to the preferences of those agents with whom the new preferences incline him to co-operate. See my (1991a, 1991b, and 1991c).

3 For arguments to this effect, see my (1991a). And for references to other philosophers with similar reservations, see my (1991d and 1988).

4 For a good interpolative exposition of Kant's views, see Solomon (1993), pp. 693–709.

References

Gauthier, David (1986). *Morals By Agreement.* Oxford: Clarendon Press

MacIntosh, Duncan (1988). Libertarian agency and rational morality: Action-theoretic objections to Gauthier's dispositional solution of the compliance problem. *The Southern Journal of Philosophy,* 26: 499–525.

——— (1991a). Preference's progress: rational self-alteration and the rationality of morality. *Dialogue: Canadian Philosophical Review,* 30: 3–32.

——— (1991b). McClennen's early co-operative solution to the Prisoner's Dilemma. *The Southern Journal of Philosophy,* 29: 341–58.

——— (1991c). Co-operative solutions to the Prisoner's Dilemma. *Philosophical Studies,* 64: 309–21.

——— (1991d). Retaliation rationalized: Gauthier's solution to the deterrence dilemma. *Pacific Philosophical Quarterly,* 72: 9–32.

——— (1992). Preference-revision and the paradoxes of instrumental rationality. *Canadian Journal of Philosophy,* 22: 503–30.

——— (1993). Persons and the satisfaction of preferences: Problems in the rational kinematics of values. *The Journal of Philosophy,* 90: 163–80.

——— (1994). Rational first values and the reduction of morality to rationality. Unpublished manuscript. Halifax, NS: Dalhousie University

Solomon, Robert C. (1993). *Introducing Philosophy: A Text With Integrated Readings.* 5th Edition. Toronto: Harcourt Brace Jovanovich.