The Blackwell Guide to Ethical Theory
edited by Hugh LaFollette
(Blackwell, 1999), pp. 247-267.

Contemporary Contractarian Moral Theory[1]

by
Geoffrey Sayre-McCord

Introduction

Contractarianism, as a general approach to moral and political thought, has had a long and distinguished history -- its roots are easily traced as far back as Plato's Republic, where Glaucon advanced it as a view of justice, and its influential representatives include Pufendorf, Hobbes, Locke, Rousseau, Hume, and Kant. In various ways, to various purposes, and against the background of various assumptions, each of these philosophers offered contractarian arguments for the views they defended. What binds the tradition together, in the face of this variety, is the conviction that moral norms or political institutions find legitimacy, when they do, in their ability to secure (under the appropriate conditions) the agreement of those to whom they apply.

As long as the tradition is, it seemed, until recently, to have quietly faded into the past. Several things were responsible for contractarianism's decline. The first was the unsettling realization that even evidently legitimate governments had never actually secured the consent of those governed, which pushed contractarianism to an appeal to the hypothetical consent of hypothetical people under hypothetical circumstances. The second was the rise of utilitarianism and, later, Marxism, as substantive alternatives that advanced their own positive views of political legitimacy along with their own criticisms of contractarianism. And the third was the influence of positivist criticisms of moral and political theory that seemed to undermine all attempts to develop a rationally defensible normative theory.

Recently, however, contractarianism has enjoyed a dramatic resurgence in popularity. This striking renewal of interest is due not only to the eventual rejection of positivism but, in part, to developments in formal decision and game theory (developments that promise a clarity and rigor of formulation all too rare in moral theory), in part, to an increasing dissatisfaction with traditional arguments for utilitarianism and its competitors, and, in part, to the sense that individuals deserve a pre-eminent place in any plausible account of moral and political obligation.

The variety of views that count themselves contractarian is daunting. Some hark back to Hobbes (and his emphasis on a preoccupation with one's own advantage), others to Kant (and his emphasis on a respect for others as ends-in-

themselves). Some restrict themselves to subjective theories of value and maximizing conceptions of rationality, others embrace objective theories of value and more elaborate accounts of practical reason. Some give a prominent role to game theory and principles of bargaining, others emphasize consensus and conciliation. And some defend familiar forms of utilitarianism (albeit with new foundations), while others defend theories that give a prominent role to rights.

In what follows I hope both to place the contemporary work into historical perspective and to set out some distinctions and contrasts that might help organize, and explain the shape of, contemporary work. The historical perspective I offer, however, is not scrupulously historical. I smooth over a good deal of the twists and turns that due care to the historical record would sanction, and I leave out of account almost completely the various social and political forces that induced not just the twists and turns but the main line of development I identify. For the sake of making clear the positions that emerged as contractarianism developed, and the philosophical considerations that recommended them, the historical picture offered here is more than a little contrived. While the views described emerged (for the most part) in the order suggested and (I believe) for the reasons offered, the succession of views was not nearly so clean as my description implies. Indeed, because each view won devotees long after others saw reasons for change, a number of the positions characterized here as long superceded kept a grip on life even as their off-spring thrived and attempted patricide. So, while I describe one view as giving way to another in the face of its perceived weaknesses, I force on the history of contractarian thought an air of inevitable development that is only partially born out by what actually happened. Still, I hope, this bit of sanitized analytical history will provide a way of thinking about contractarianism that sheds some light on why it developed as it did.

The Background

Contractarianism came into its own in the seventeenth and eighteenth centuries, primarily as a political theory. It developed directly as a response to concerns about the legitimacy of government and the grounds of political obligation. As faith in the divine right of kings evaporated, and the assurance that some were by nature born to rule waned, people came to see political authority as a reflection of human convention. The question arose: what could possibly justify the state and explain our obligation to it? "Man was born free; and everywhere he is in chains" Rousseau (1978: 46) famously noted, "What can make [this change] legitimate?" Social contract theories offered an appealing answer that traced the grounds of political legitimacy and obligation not to God or to nature, but to the wills of the people who were affected. In the process, it promised to articulate the origins of, obligations to, and limits on, legitimate government.

Viewing government either as a conventionally established arrangement among people, or between a people and their sovereign, the social contract was called on in three capacities. First, at least initially, it was offered as an explanation of how governments actually arose. As the explanation would have it, governments emerged in a pre-political context -- the "state of nature" -- as conventional solutions to the problems that inevitably arise among people in the absence of a state. Second, the social contract was offered as an account of why people have an obligation of allegiance to their government. The idea was that people have such an obligation because either they had consented, or they had good reason to consent, to the government's authority (as a way to avoid the hardships threatened by the 'state of nature'). And third, the social contract was advanced as a justification for limiting the powers of government. The suggestion was that a government's powers are properly limited to those necessary for solving the problems that give rise to government in the first place, since more extensive powers would go beyond what people would have reason to consent to were they in a 'state of nature'.

Contractarianism provided, as a result, a normative framework that might be used to defend or to attack the legitimacy of particular governments. As it happened, contractarian arguments were in fact relied upon both to foment and to resist revolutionary pressures. Thus Rousseau's The Social Contract came to the fore as a natural source for the condemnation of virtually all existing governments while Hobbes' Leviathan was, in its implications and intention, conservative in the extreme. When the point was to attack the legitimacy of some particular government, contractarians would argue that the relevant people had no reason to recognize its authority because the life they could realistically expect to face without it would be better. When the point was to defend the legitimacy of some government, contractarians would argue that the relevant people had reason to recognize the authority of some government because the life they could realistically expect to face without the government would be palpably worse.

Originally, and especially while natural law theory still held sway, the contractarian arguments played out against two substantive assumptions: that real consent had (sometimes) actually been given and that people had a moral obligation to keep the agreements they had made. The first assumption legitimated the sovereign's power, since the relevant people were supposed to have given their permission for its exercise. The second established the political obligations of those party to the agreement, since they were seen as having undertaken, by way of the agreement, certain specific moral obligations.

Just what constituted giving consent, however, became a ticklish and difficult issue for those who thought it required. Instances of explicit consent seemed rarely on offer. So tacit consent, given (say) by the acceptance of benefits or participation in certain conventional practices, emerged as the only sort of consent that might actually be given on a regular basis. Yet tacit consent,

if characterized in a way that allowed it to be often enough secured, was so easily given (especially under conditions where no viable options existed) as to be of little use in justifying political authority and obligation. To the extent those too poor to move accepted the benefits of the state, for instance, their tacit consent seemed not so much a reflection of their wills as the inevitable up-shot of their situation. Virtually unavoidable, tacit consent seemed insufficient grounds for distinguishing legitimate from illegitimate government.

Questions arose as well about the need for real consent, explicit or otherwise. For it appeared, on the one hand, that particular legitimate governments had never received consent of any kind, and, on the other hand, that whatever might justify a person's moral obligation to keep agreements would serve as well to justify a political obligation (even in the absence of actual agreement). (Hume, 1985) Whatever did legitimize government eventually seemed not to depend on the government's consensual pedigree.

Thus actual consent looked to be neither sufficient nor necessary for legitimate government and political obligation. Nonetheless, the idea that people would, with good reason, willingly give their consent, seemed to offer a compelling endorsement of whatever they would agree to -- even if they hadn't actually given their consent. Conversely, the idea that people would, with good reason, not willingly offer their consent, seemed a powerful condemnation -- even if their consent had in fact been given either for bad reasons or unwillingly.

As a result, appeals to actual consent, along with the fanciful histories that accompanied early attempts to identify when it had been given, were replaced by appeals to hypothetical consent -- appeals to what people would agree to, if only they were rational, and not to what they had agreed to. But this raised difficulties of its own, for while real agreement might establish real obligations, hypothetical consent, which was no consent at all, apparently established no obligations whatsoever. Still, that people would have given consent, if they were given the chance and were rational, does seem to establish something: that they had reason to support what they would have consented to. So, leaving behind the suggestion that consent (of any sort) was the source of the obligation, contractarians shifted to the claim that what mattered was that people had reason to give their consent if it were needed either to establish or to maintain the government in question. The reasons for giving consent, not the consent itself, were taken as establishing obligation. The authority of some government, in turn, was seen to depend on people having reason to recognize it as authoritative, not on their actually having given that recognition (via consent or contract, or indeed in any other way).

The switch to hypothetical consent (which plays a role in Hobbes and is clear in both Rousseau and Kant) allowed contractarians to avoid explicit consent's implausible histories and tacit consent's excessively lenient account of commitment, as well as their shared reliance on the assumption that people have

a moral obligation to keep their agreements. Appeals to hypothetical consent emphasized instead the reasons agents had for reaching agreement in the first place. And they allowed the theories to rely not on what people might actually have done (perhaps for bad reasons or unwillingly) but on what they had good reason to do.

The reasons people were seen as having were not, importantly, provided by some independently specifiable theory of what constituted legitimate government. The argument was not that people had reason to give their consent because the government was legitimate, but that it was legitimate because they had reason to give their consent. The reasons people (supposedly) had needed to be found, therefore, in considerations that did not presuppose an account of legitimacy. The considerations offered were, in a straightforward sense, standardly practical, in that they emphasized the advantages to each that would come from the state.

Of course the details of the theory shifted substantially as different accounts of the pre-political state of nature were offered, since those accounts influenced significantly what people might reasonably have agreed to and so what powers they might have recognized as necessary or desirable. Hobbes, for instance, saw the state of nature as so threatening that it called for an absolute sovereign limited only by an inability to demand that citizens willingly submit to death. Whereas Locke, convinced that a pre-political state of nature would be a tolerably harmonious community suffering primarily from the undesirable effects of people trying to enforce individually what they each regard as their rights, saw the government's legitimate role as really quite limited.

In any case, and whatever the description of the pre-political state of nature, contractarians offered the description as a realistic characterization of how things would actually be without government. The contractarian approach asked people to consider seriously what life would really be like without government. In an age of revolutions, this was not a call to imaginative flights of fancy, it was the pressing of a seemingly realistic possibility. To the extent the prospect of life without government was palpably worse than life with government, people could see themselves as more than willing to consent to some government or other, rather than face the alternative.

Crucial to these contractarian arguments, of course, was their effectiveness in advancing a credible view of how life would be in the absence of government. Those who saw themselves as facing the prospect described saw as well that they had reason to give their consent, if it were required, to some political authority. Against the background of a compelling description of state of nature, the contractarian argument had a great deal of appeal.

As long as the state of nature represents a genuine threat, and as long as staying out of that situation requires mutual cooperation in support of government conventionally established, the real people facing the threat will

each find they have reason to recognize the authority of the state. It won't be actual consent that carries the burden so much as there being compelling grounds for giving it if asked -- grounds provided by the thought that without the government life would be much worse. So while the proof of acceptability was originally thought to be found in actual acceptance, via real consent, that proof came to seem securable, and sufficient, in the absence of consent.

To admit that having some particular government is better than having none, however, is not yet to hold that any government whatsoever is acceptable. Some governments might still be so terrible that people would reasonably prefer no government. Moreover, those governments that do improve on the state of nature will not do so in the same way, and if people faced a choice they might well prefer one of these to the others. No doubt, for any given government, some people falling under it would prefer a different government even if not the state of nature. Yet the question people faced wasn't just which government (or form of government) would they prefer to a 'state of nature' nor was it which would they most prefer given their particular convictions, tastes, skills, weakness.... It was, instead, which government (or form of government) would they prefer from among those that might also secure the support and agreement of the others who are to establish and maintain it. The question became: what form of government would properly (and plausibly) secure the agreement of each in light of the fact that no acceptable government would be establishable without the consent of all (or at least most). Indeed, at the core of contractarianism is the insistence that the arrangements must prove acceptable to all who would fall under them.

To this point, the argument for being willing to give consent was not merely addressed to real people, it was put to them in terms they were supposed to see as accurately characterizing a choice they might in fact have. It was hypothetical consent theory in the face of a possibly real choice. And the reasons people were seen as having were reasons they supposedly actually had in their circumstances.

Unfortunately, while people might willingly consent to (maybe almost any) government rather than face the state of nature, that willingness could well reflect aspects of their actual situation that are morally suspect. That a person has reason to give consent to enslavement rather than face the (perhaps quite real) prospect of painful death at the hands of the would-be master does nothing to establish the legitimacy of the enslavement. Real reasons, under coercive circumstances, may legitimize giving consent, but they won't legitimize others acting on that consent. Thus, in order to justify the authority of some government, hypothetical consent needed to emerge in situations reasonably viewed as morally untainted.

The pressure to purify the circumstances of agreement naturally led the hypothetical consent contractarians to advance as well idealized circumstances

for that consent. Those who resisted an appeal to idealized circumstances did so for either of two reasons. On the one hand, some thought the actual circumstances, and thus the 'state of nature', are appropriately untainted even if life is less good than it might be. On the other hand, some thought that whatever reasons a person might have for making an agreement under idealized conditions, those reasons would be irrelevant to real agents unless those reasons were likewise reasons they would have under suitably realistic conditions -- in which case there was no point in appealing to the idealization. Still, most contractarians were moved by considerations of fairness to hold that the relevant choice situation was one made under idealized conditions.

Similar considerations worked to recommend too that the people whose consent mattered were not people as they actually are -- sometimes irrational and often ignorant -- but those people as they would be were they (for instance) perfectly rational and appropriately informed. That a person, when irrational or ignorant, would or even does give consent under certain (perhaps non-coercive) circumstances to some arrangement, does nothing to establish that he has reason to give that consent. Consent under non-coercive circumstances may legitimize others acting on that consent, but it won't legitimize thinking the person has reason to consent.

All the while recognizing that the question is whether real people have reason to endorse the government whose legitimacy is at issue, contractarians began to distinguish these real people and their actual circumstances from the (suitably idealized) people who are supposed to reach agreement and the (appropriately idealized) circumstances under which their agreement was supposed to be secured. That suitably idealized people would, under appropriately fair circumstances, willingly agree to some (form of) government, came to seem the plausible standard of legitimacy. What grounds could there possibly be, one is inclined to ask, for objecting to such a government?

A natural worry, however, is that the rhetorical force of this question was bought at the expense of vacuousness. For it began to look as if all the interesting justificatory work would be done in specifying who might count as suitable parties to the agreement, and what would count as appropriately fair circumstances. The contractarian test came to seem empty without the addition of some non-contractarian theory that would identify not only who counts as suitably rational, and what circumstances are appropriately non-coercive, but also what reasons there are for consenting under those circumstances. Such a theory would presumably support its own substantive account of political legitimacy. Whereas, initially, contractarianism appealed to real people in their actual circumstances facing a real choice, it now was so removed from the real world, and so normatively laden in its assumptions, that the contractarian framework seemed at best a useful heuristic for discovering some independently specifiable criterion that must be defended on some other, non-contractarian, grounds.

A number of theories emerged as candidates for the role. The most influential, early on, was utilitarianism, which held, in effect, that what rational people would agree to (under actual as well as idealized circumstances) is precisely whichever government would maximize over-all welfare. The legitimacy of a government turned, utilitarians argued, on how well the government advanced the interests of everyone concerned. And the reason any particular person had for recognizing the legitimacy of the government -- and so for consenting to its authority -- was traceable not (for instance) to how that person would fare, but to how people in general would. Of course this theory didn't require the contractarian framework for its articulation or deployment. Yet if one asked whether, on the utilitarian view, all rational people could, under appropriately fair conditions, willingly give their consent to the (form of) government it endorsed, the answer was an easy 'yes'.

Other theories, such as natural rights theories and (on some interpretations) Marxism, also stepped into the breach and offered accounts of what might constitute fair conditions of, and good reasons for, agreement. Each of them, though, was in a position to side-step an appeal to the contractarian framework even as it had the resources to say that all (really) rational people would, in the appropriate circumstances, willingly agree to the (form of) government the theory legitimized. To the extent these theories didn't co-opt the rhetorical force of the contractarian framework, they were used instead to undermine it on the grounds (for instance) that the framework illegitimately valorized the individual or ignored the value of community or substituted market relations for moral ones (see Pateman 1988, Sandel 1982).

Once contractarianism stepped away from reliance on the real consent of real people, and once it moved on to embrace as relevant only the hypothetical consent of idealized people in idealized circumstances, it not only invited non-contractarian additions, it seemed to need them. And once the additions were at hand, they didn't just add to contractarianism, they displaced it.

Recent Contractarianism

That is pretty much where things stood until the middle of the twentieth century. Although, in the first half of this century things got even worse, thanks to the influence of logical positivism. For according to the positivists grand attempts at moral theory and political justification are, despite pretensions to the contrary, actually only elaborate devices for bringing others onto one's own side.

The relatively recent revival of contractarianism has depended upon an emerging conviction that, contra positivism, there must be room for reasoned argument about normative matters. But the revival has required two other things as well: first, a growing dissatisfaction with the theories that had displaced contractarianism, and, second, the prospect that recognizably contractarian considerations might after all contribute non-trivially to moral theory.

Moreover, just as political contractarianism emerged as a response to the recognition that political legitimacy and obligation could not be traced to God or nature, moral contractarianism's appeal has grown substantially with the sense that moral constraints must in some way be a reflection of human reason or social convention, not of God or (non-human) nature. Contractarianism holds out the seductive prospect of a theory that demystifies morality's status and shows it to be a compelling expression of humanity's nature. For if morality finds its source and authority in our capacity to embrace its demands, then understanding morality will ultimately require appealing to what we would need in any case to explain our own capacities and practices. Nothing occult or mysterious or supernatural need be implicated (Mackie 1977, Milo 1995).

The contractarian framework, with its appeal to what people would agree to under appropriate circumstances, has found a natural home in two very different approaches that take their inspiration (though frequently little else) from Kant and Hobbes. The Kantian approach begins with our natural concern with morality and uses the contractarian framework to specify and draw out the implications of that concern. Contractarianism, in this case, is advanced as a way to articulate the content of morality. The Hobbesian approach, in contrast, acknowledges our concern with morality but sees that concern itself as properly called into question and uses the contractarian framework to show why and to what extent we have (non-moral) reason to embrace morality. Contractarianism, in this case, is advanced as a way to justify a concern for morality's content and demands. On either approach, contractarianism's distinctive commitment to seeing legitimacy as grounded in what people might willingly agree to under the appropriate circumstances finds a central role.

Kantian Contractarianism

The Kantian approach has famously been pursued by John Rawls, who introduces the contractarian framework to articulate morality's impartiality. In the process, he hopes to take "seriously the distinction between persons" in a way that other attempts to capture impartiality do not (1971:187).

That moral demands are impartial is, of course, acknowledged by virtually all moral theories. Yet the nature of that impartiality and its implications for morality's authority and content are quite controversial.

One familiar way to capture morality's impartiality is to suppose that moral demands flow from a source equally concerned for all who fall under them. Thus many religious views portray morality's demands as reflecting God's equal love for all, while Ideal Observer theories treat the demands as an expression of what an equi-sympathetic observer would approve of, and utilitarian views see the demands as giving equal weight to the welfare of all. This way of capturing impartiality leads naturally (although not inevitably) to moral principles that are decidedly utilitarian in their implications.

Another way of capturing morality's impartiality, however, is to see its principles as those we would each, individually, choose to govern everyone's behavior if our choice was made in ignorance of how we, as we actually are, might benefit or suffer as a result (Harsanyi 1953, Rawls 1971). With this in mind, Rawls describes the appropriate circumstances of agreement -- the relevant 'state of nature' --as including a 'veil of ignorance' that shields from view all information concerning the particular talents, tastes, history and situation of those seeking agreement. Impartiality is achieved by eliminating all the information that would engage partial concern. Collective and partial choice under circumstances of radical ignorance is substituted for individual choice under circumstances of extraordinary impartiality and knowledge. And with the substitution comes the appeal to contractarianism. For now the legitimacy of certain principles, and their standing as distinctively moral, appropriately impartial, principles, turns on whether people would -- under the relevant circumstances (in this case, circumstances of ignorance that neutralize partiality) -- choose the principles. Hypothetical choice, under hypothetical circumstances, sets the standard for moral legitimacy, on this view, because such choice embodies impartiality.

One apparently significant advantage of the contractarian approach to impartiality is that it need appeal neither to interpersonal utility comparisons nor to any general method of balancing the interests or welfare of people. In contrast, when impartiality is embodied in equal love, or sympathy, or concern for other's welfare, an appeal to interpersonal utility comparisons and an over-all balance of advantages seems inevitable. Many see this difference as grounds for thinking that the contractarian embodiment of impartiality captures especially well the moral significance of the individual and the idea that one person's loss cannot always be morally compensated by another's gain (Rawls 1971, but see Harsanyi 1953).

Impartiality is only one aspect of morality that invites contractarian elaboration. Rather than starting from the conviction that moral demands are impartial or fair, some versions of contemporary contractarianism focus instead on the idea that moral reasons are public and shared -- they provide reasons for all. These approaches shift attention away from conflicting interests that call for impartial arbitration towards a collective concern to accept principles all can embrace as reasonable. Contractarianism's appeal to mutual agreement (under appropriate circumstances) strikes many as doing a uniquely satisfying job of articulating the sense in which morality's demands can lay claim to the allegiance of all. Indeed, by treating moral norms as just those everyone has reason to accept, contractarianism not only articulates the connection between morality and mutual acceptability but takes that connection as definitive. A concern to act morally, on this view, is a concern to act in light of principles that everyone might reasonably embrace. To determine what these principles are we need to ask the distinctively contractarian question: to what could people, under

the appropriate circumstances, reasonably agree? This time, though, the appropriate circumstances are conceived not as involving radical ignorance but instead as being occupied by participants who offer considerations for and against various principles in a context where all are supposed to be both reasonable and concerned to settle on principles all the participants can accept. (See Scanlon, 1998; Habermas, 1990).

While impartiality and mutual acceptability have both played crucial roles in making contemporary contractarianism attractive, they themselves find support in a third aspect of morality -- the evident importance of equal concern and respect. Many hold that the moral significance of individuals is best captured by a view that treats morality's demands as themselves a reflection of what each person, uncoerced and conceived of as a full participant in the process, could rationally embrace. To treat a people with equal concern and respect, on this view, is to see them, no less than oneself, as having a legitimate say in the principles that should govern your interactions. By governing oneself by principles others could endorse, one thereby gives expression to the equal concern and respect that is distinctive of morality.

As I have suggested, contractarians disagree among themselves as to which, if any, of these various considerations ought to be given primacy. Even among those who agree on that, there is significant disagreement as to how impartiality, or mutual acceptability, or equal concern and respect, might find their best expression. Despite the disagreement, however, there is consensus among those taking this approach, that the relevance of the contractarian framework is found in its capacity to articulate crucial and distinctive features of morality. Thought of in these terms, contractarianism addresses those who are already concerned to do as morality demands but are trying to figure out what, precisely, that might be. Contractarianism is offered as a way of specifying those demands and is defended as appropriate by appeal to its capacity to articulate and embody crucial features of morality.

The Kantian approach to contractarianism faces two related problems. One concerns whether, in the end, any real work is being done by the appeal to the agreement of people, properly situated. The more successful an account is in eliminating the influence of individual differences on choice, in the name of impartiality, the less room there seems to be for the idea that the choices of distinct individuals matter to the outcome. When it comes to choices behind a veil of ignorance, for instance, asking what all might agree to under that circumstance appears functionally equivalent to asking what any one person might agree to, since the veil hides from the scene all the features of a person that might distinguish one person from others. Do the notions of collective choice or mutual agreement really have any substantive place in such theories? It is not at all clear. In any case, a second familiar problem emerges when one asks on what basis the people are to reach a collective choice or mutual agreement, supposing they do. Any grounds people who are "properly situated"

might have for settling on one choice or agreement rather than others threaten to stand independently of any choice or agreement at all. Even hypothetical agreement among people seems to drop out of the picture. The concern underlying both of these problems is that the contractarian appeal to what people -- in the plural -- might agree to, under whatever circumstances, seems not to be playing anything other than a heuristic role.

This is a serious concern. If it is not met, the contractarian framework will stand as mere window dressing, a decorative over-lay that might have evocative advantages but that contributes not at all to the substance of a theory or the justification of the principles it endorses. The central challenge is to find a role for the distinctively contractarian idea that morality's demands reflect, in some non-trivial way, what people might reasonably agree to under the appropriate circumstances. Of course, a range of obviously non-contractarian theories can end up saying that the principles they advance might be chosen by reasonable people properly situated. When they do say this, however, the appeal to what the people might agree to (or choose) swings off the side as a fifth wheel rolling along with the theory but driving no part of it. The content of the principles advanced is wholly unaffected by consideration of what people might reasonably agree to -- all the influence goes the other direction.

In order to meet the central challenge a contractarian theory has to show that the appeal to what people might agree to is sensitive in some way either (i) to the variety of people participating in the choice, or (ii) to the variety of people to be governed by what is chosen, or (iii) to the variety of people being addressed by the argument. Only then will the idea that morality turns on what distinct individuals might agree to have a significant role. If, alternatively, the key choice or agreement in play might as well be made by and for a single individual, talk of what people (as opposed to a person) might agree to will have no substantive impact on the nature of the principles that are supported by the argument and the contractarian framework will be making no significant contribution to the theory. As it happens, all three of these options have been explored, exploited, and defended.

Thus some contractarians argue that the appropriate circumstances of choice leave intact, in the way a veil of ignorance might not, the key fact that those who are seeking to reach agreement differ from one another in ways that influence which principles might be genuine candidates for mutual agreement. This sort of argument usually characterizes the relevant agreement as being a result of bargaining or some kind of balanced accommodation, the particular content of which reflects differences among the participants. By leaving intact individual differences and allowing those differences to have some impact on the nature of the principles that are supported by the argument, such views make genuine room for the contractarian thought that the legitimacy of certain principles depends non-trivially on what people collectively might agree to.

Other contractarians argue that the outcome of the choice in question is shaped substantially by the fact that what is being chosen (a set of principles to govern interactions among individuals, or a set of basic institutions to structure society, or whatever) is chosen for a potentially diverse group of people who differ in talents, values, personality, etc. This sort of argument turns our attention to the ways in which the choice problem is shaped by the prospect of the results applying to different people. Even if, in the appropriate circumstances, one chooser is as good as another and more than one is no real addition (as might be the case behind a veil of ignorance), it may be that what such a chooser would select is influenced by the fact that those for whom she is choosing are different in ways that need to be accommodated, from the start, by the choice she makes.

Still other contractarians argue that the whole choice situation -- the circumstances under which it is to be made and the nature of those who are to make it -- is answerable, in a non-trivial way, to the fact that the over-all contractarian argument is being offered not to a single person but to people insofar as they see themselves as together trying to settle on acceptable principles for interacting with each other. This sort of argument focuses on the situation of the actual people to whom the arguments are addressed and maintains that their differences have an impact on just how the choice situation is to be described. A crucial feature of our actual situation, it seems, is that we can expect reasonable people to disagree fundamentally about central philosophical and moral issues in ways that mean there is no real prospect of reaching an across the board consensus. Nonetheless, there may be room for all reasonable people to agree (for their different reasons) that there are reasons to regulate our interactions by norms that are mutually acceptable by all who are reasonable and we may see the original contract situation as articulating the common ground shared by all those who are reasonable (Rawls, 1993).

All three lines of argument have more than a little plausibility. At the very least, they suggest there might be room to defend the view that the contractarian framework, in some guise or other, can play more than a heuristic role in a theory. Still, those who offer some version of contractarianism as the best articulation of moral concerns we are assumed to share face two additional worries.

The first is that the variety of contractarian theories itself testifies to the fact that people's prior understandings of morality differ significantly, even when it comes to thinking through what impartiality, say, consists in. And this raises a problem since, in the face of a range of different contractarian theories, the question naturally arises: which, if any, of these articulations of our prior concern captures accurately the object of that concern? Asking what people, appropriately situated, might agree to seems to provide no purchase whatsoever on that question, since all the different versions of contractarianism will travel with their own preferred description of the circumstances of choice and of the

grounds on which the people so situated will reach agreement. Whatever counts as good grounds for settling on one (contractarian) characterization of our moral concern rather than another, it seems it won't be grounds that turn in any interesting sense on what people would agree to, if they were properly situated. The fundamental argument for one view rather than another looks as if it will have to be decidedly non-contractarian.

The second worry arises even if a particular characterization of the contractarian framework settles out as successfully capturing our moral concerns. It centers on the question: what reason is there to embrace that moral concern? Even those who are concerned to act as morality requires might, on reflection, wonder whether they have any good reason to retain or act on that concern, especially in situations where morality quite clearly requires sacrifice. Why not think of the concern as merely a reflection of socialization that one would do better to be without? Insofar as contractarianism is offered solely as a way to articulate a concern for morality that we are assumed to share, it will in effect ignore the issue. But it is an issue, many think, that should not be put to one side casually, not least of all because so often peoples' actual concerns reflect ignorance, superstition, and prejudice. Morality of course presents itself as legitimately commanding allegiance and sacrifice. But do we really have reason to offer the allegiance and make the sacrifices, when called for?

The Hobbesian approach to contractarianism takes this challenge seriously and sees the contractarian framework as offering a uniquely compelling answer to it. Before turning to this approach, though, I should mention one tempting answer that is available to the Kantian contractarian. As this answer would have it, those who acknowledge that some course of action is morally right or required, but wonder whether they have reason to act accordingly, are failing to appreciate something that follows directly from what they have acknowledged: that they have (moral) reason to act as required. Acknowledging a moral demand, it seems reasonable to think, carries in its wake recognition of a moral reason to act accordingly. But this observation just pushes the problem back a step, since now the question is whether one has any good reason to give weight to moral reasons in one's deliberations. One might insist at this point that moral reasons are necessarily weighty so that once we admit there are moral reasons there's no good sense to be given to wondering about the weight of those reasons. Yet it is not hard to sympathize with those who would feel cheated by such an answer -- cheated of a non-question-begging defense of the importance of morality.

Hobbesian Contractarianism

The Hobbesian approach to contractarianism offers such a defense. Those who take this approach argue that we have non-moral reasons to embrace morality. The distinctively contractarian element in this approach comes with explaining the way in which the reasons we each have for embracing morality

are reasons that reflect the interdependence of our interests and the opportunities we have for mutual benefit. Speaking in broad terms, the Hobbesian approach views morality as constituted by a set of principles the adoption of which is advantageous for everyone in a way that means each person would have (non-moral) reason to adopt the principles as long as others did as well. And it sees the legitimacy of morality's demands as turning on our having (non-moral) reason to support them.

One of the earliest versions of contractarianism, advanced in Plato's Republic, contained the core elements of the Hobbesian strand of contemporary moral contractarianism. Put in Glaucon's mouth, this version of contractarianism is pleasingly direct. According to the view he sets out, the rules of justice are conventional and represent a compromise. On the one hand, people would prefer to have their wills unchecked by others. On the other hand, they would prefer not to suffer the unchecked wills of others. Recognizing that they can't enjoy the first without suffering the second, and rightly fearing the second, they band together to establish and enforce mutually agreeable limits on each other's wills. These limits, Glaucon suggests, simply are the rules of justice. Thus, as Glaucon tells the story, the constraints justice imposes are a reflection of convention, yet the reflection of a convention we each have (non-moral) reason to encourage and embrace (given the human condition). The convention that constitutes morality, while a compromise of sorts, is nonetheless a reasonable one that redounds to our mutual advantage. (Gauthier 1986, Buchanan 1975, and Harman 1978, all offer contemporary defenses of this sort of view).

Game theory provides resources for representing perspicuously the underlying structure of social interactions that give point, in the way Glaucon suggests, to moral principles. In the process it has made possible a sophisticated investigation of the various different ways in which the reasons any particular person might have to act in one way or another depend upon what others have reason to do. Perhaps most influential on this front has been the Prisoners Dilemma which models a situation where the options and available benefits are such that, if each person directly maximizes her expected utility, they will together predictably end up worse off than they would have been had they cooperatively forgone immediately available benefits.[2] But various other notions from game theory and economics have played crucial roles in recent discussions of contractarianism. Especially important on this front have been developments concerning the understanding of Free Riders (who enjoy a benefit thanks to the efforts of others without themselves participating in producing that benefit), externalities (which are costs imposed by decisions that are shifted to those who have had no say in their production), and assurance problems (where a potentially available benefit for all will be beyond reach unless all have assurance that others will do their parts). In each of these cases, it looks as if a successfully established and internalized set of principles requiring certain sorts

of acts, demanding the consideration of others, and underwriting confidence that others will act in concert, would alleviate problems we would all otherwise face. Reciprocal constraints, intelligently selected, lead to mutual advantage.

The hope held out by Hobbesian contractarianism is that, at least to some extent, moral principles might ultimately be justified by showing the extent to which we all benefit from living in a community of people who constrain their pursuit of interest by those principles. At the same time, though, the hope must be balanced by the recognition that in many ways the Hobbesian approach will likely support principles that match commonsense at best only imperfectly.

On the Hobbesian view, for instance, the advantages we each enjoy from morality come primarily from others embracing moral principles and secondarily from our avoiding the burdens we would suffer were others to punish us for violating those principles. In particular situations, the balance of advantages may fall in favor of violating particular principles, especially if one can do so undetected (and so unpunished) by others. As a result, even where there might be mutual advantage in establishing recognizably moral principles to govern our interactions, there may in certain cases be no advantage from -- and so no reason, on this view, for -- compliance. And even when there is an advantage to be gained, the motive for so complying appears to be distinctly non-moral. Thus at most the Hobbesian approach seems to underwrite acting morally for non-moral, indeed apparently selfish, reasons. To the extent a full justification of being moral involved justifying doing as morality requires *because* morality requires it, the Hobbesian might seem incapable of providing the justification (but see Gauthier 1986 and Sayre-McCord 1989).

Moreover, on this view, the principles that would be mutually advantageous overlap only contingently, and then pretty clearly only partially, with those we currently recognize as moral principles. After all, to the extent the principles we have reason to embrace turn on what others too have (non-moral) reason to embrace the principles will almost surely reflect the differential power, wealth, and general situation of those party to the arrangement. Similarly, when it comes to those whose protection brings no advantage to others (e.g., the weak and infirm) the principles the adoption of which would bring mutual benefit would presumably not offer them protection, since such protection would bring no advantage to others. In both cases, the resulting -- mutually advantageous -- principles will presumably differ from those recommended by commonsense morality.

The tension between commonsense morality and the principles that would be recommended by Hobbesian contractarian is due in no small part to holding that principles are legitimate only if they can be should to be advantageous to real people in their actual circumstances. For, almost inevitably, morally suspect differences among people will then influence the content of the principles that will qualify as legitimate (because genuinely advantageous). Yet

the more one corrects for these morally suspect differences by focussing not on actual advantages people can expect but on the advantages that would be secured under hypothetical circumstances, the less one can claim real people have (non-moral) reason to care. If I have been born to comfort and wealth or have secured such a life through force or fraud or cunning, any subsequent agreements I might make would no doubt be distorted by my initial advantages. Of course someone might, on grounds of fairness, say, insist on disallowing the influence of these advantages, but then the prospect of mutual advantage plummets as the real benefits to me disappear. The moral appeal of the resulting principles seems to be inversely proportional to their claim on actually being advantageous to all. Be that as it may, if one takes seriously the idea that people should act as they have reason to, and if one thinks what one has reason to do is whatever is personally advantageous, then a mismatch between advantage and commonsense morality would be all the worse for commonsense.

Fortunately, Hobbesian contractarians can and usually do admit that people's interests and preferences may be other-regarding, sympathetically directed, and broadly sensitive in ways that mean a true appraisal of how their interests are intertwined with other's will reveal, after all, an argument from mutual advantage to principles that are recognizably moral. Their appeal to interest and advantage in defending moral principles does not have to be an appeal solely to self-interest and private advantage. By taking honest account of human nature, and the extent to which we can be engaged by the welfare of others (to a greater or lesser degree, in response to both nature and nurture), it is at least plausible to think real advantage for all may be secured by the adoption of moral principles already securely established in commonsense.

No doubt any defense of morality that needs to appeal, in this way, to our fellow feeling leaves moral principles contingent in two ways that may be disturbing. First of all, the content of the principles is, on such a view, contingent upon the existence and shape of our concern for others. Second of all, the force of the argument offered for giving allegiance to the principles will be contingent as well on the actual concerns of those addressed. Although, unlike Kantian contractarianism, the Hobbesian variety need not suppose that the people addressed by the argument already possess a distinctively moral concern.

Significantly, the Kantian and Hobbesian approaches may compliment rather than compete with each other. For it may well be that the concern we have non-moral reason to embrace (as the Hobbesian would argue) is a distinctly moral concern the content of which calls for contractarian elaboration (as the Kantian would maintain).

Conclusion

There is a third version of contractarianism, inspired by Hume, that takes for granted neither a concern for morality nor any particular account of what

people have reason to do or accept (Hume 1978). It sets out to explain why evaluative concepts and commitments would naturally emerge among beings with our capacities, concerns, strengths and weaknesses. The distinctly contractarian elements in the evolutionary story revolve around the evaluative concepts and commitments themselves being conventional solutions to problems people would otherwise face. In setting out to explain our evaluative concepts without presupposing others the Humean contractarianism is more ambitious than either the Kantian or the Hobbesian approaches. Yet the ambition is mitigated by the fact that the Humean approach is concerned neither to establish any particular substantive moral view nor to argue that people have reason, of any particular sort, to be moral. Instead, it hopes to account for the evaluative concepts we actually possess (contractarian in content or not, rationally embraced or not) by appealing initially only to non-evaluative features of our situation and the de facto advantages that come with the capacity to think in evaluative terms.

Once evaluative concepts are up and running, and have a life of their own, the Humean -- no less than others -- will rely on them in justifying or criticizing not only particular actions and institutions but also, in some cases, the conventions that give shape to the concepts themselves. As a result, the Humean contractarian might well end up defending a particular evaluative stance concerning morality and practical reasons more broadly. So she may embrace the Kantian view that we possess a moral concern that is best articulated by appeal to the contractarian framework, or she may share the Hobbesian view that a proper understanding of what people have reason to do shows that their reasons are essentially bound up with their own advantage. Or she may reject both views. Her commitment is to seeing the evaluative concepts she relies on as being grounded in, and shaped by, a distinctive set of conventions. Just as moves in a game of chess make sense only in the context defined by the rules of the game, so too, the Humean maintains, evaluative judgments make sense only in the context defined by the conventional rules governing the concepts that are deployed in those judgments. Our capacity to think in moral terms and to talk of reasons depends, on this view, upon resources that are available only once certain conventions and practices have been established.

Significantly, Humeanism offers an account of the conventions that give place and point to distinctively evaluative concepts, not (or at least not merely) an account of fellow feeling or altruism or cooperative dispositions. Thus it hopes to explain our capacity to make evaluative judgments and not (merely) our capacity to get along or respond emotionally. Presumably the conventions that define our evaluative concepts require the presence of various affective reactions and dispositions. But the Humean's focus is on those conventions themselves and the way in which they serve to constitute our evaluative concepts by setting standards for their correct application.

The Humean approach resembles the Hobbesian, in that the introduction of evaluative concepts (and the principles or standards that specify their content) is seen as an advantageous solution to a problem people collectively would otherwise face. Yet there are some crucial differences. In particular, on the Humean view, while the concepts do have this benefit they are not seen as deliberately introduced on the basis of reasons people recognize (since, by hypothesis, there is no substantive concept of reason yet in play and so no sense to be made of people actually recognizing reasons). Thus, in the first instance, the concepts are seen as arising in an explicable way, given the situations in which people would find themselves, but not as rational solutions to a problem of collective choice. Of course, once the relevant concepts are in place, it is possible to reflect back on the introduction and evolution of the evaluative concepts. And on reflection, the Humean approach assumes, one will discover there are good grounds for being glad something like the original concepts were introduced and for endorsing what they have become as they have evolved. However, at this point, the grounds for approving of the evaluative concepts we share will go beyond the austere resources of an appeal to self-interest and will implicate substantive considerations of fairness, justice, and value in ways that a Hobbesian excludes from consideration. (See Sayre-McCord 1994.)

Of course, reflecting on the origin and nature of our evaluative principles may well reveal deep problems with our current understanding of our evaluative concepts. But then the grounds for criticism and the justifications offered for altering our understanding of what justice, say, requires, will of necessity invoke evaluative concepts we have and can understand. The original, mutually advantageous, conventions will be providing, in these cases, both the resources and the reasons for reflectively correcting the conventions as they stand. The process of reflective adaptation that is then in play is a crucial element in making sense of an otherwise puzzling and anyway distinctive feature of evaluative concepts -- their essential contestability.

The process of reflective adaptation plays an important role in addressing two worries. The first worry is that Humean conventionalism is committed to an objectionable form of relativism since the concepts that may emerge in one community might well differ substantially from those in another community. The second worry is that by giving a central role to mutual advantage the approach will inevitably underwrite moral principles that are arbitrarily parochial in their focus and implications. After all, the concepts that do emerge in a particular society, it seems, will be shaped by the interests of those in the community without regard to others. Both considerations raise serious worries, of course, but only insofar as the relativism involved is objectionable and the parochialism arbitrary. There is no doubt that some versions of relativism are objectionable and that parochial concerns are often arbitrary. Still, that different communities may develop different evaluative concepts to answer to their particular situations seems not only something that obviously happens but also

unobjectionable (as long as the concepts in questions are unobjectionable). Similarly, that the concepts that develop within a community answer to that community's needs and interests seems not at all arbitrary. Nor does it seem disturbing on other grounds once we notice that the content of the concepts we have an interest in having may well, and in fact do, bring within their scope the interests of others. To the extent there are reasons to expand the scope of our principled concern or adjusting our understanding of our commitment's implications, those reasons are articulated using our current concepts. And these are concepts the currency of which finds an explanation in the Humean story of their social role. Our capacity reasonably to criticize principles and practices cannot outstrip the conceptual resources we have for identifying and articulating the supposed difficulties and what the Humean view offers is an explanation -- a metaphysically and epistemically modest explanation -- of those resources. Barring the discovery that our evaluative concepts carry the seeds of their own destruction, Humean contractarianism is well placed to accommodate and even embrace whatever substantive considerations might be mobilized for thinking there is reason re-evaluate our evaluative commitments.

This very capacity to accommodate and adapt to new considerations calls into question the value of Humean contractarianism, to the extent one hopes to use contractarianism to identify and defend some particular (and fixed) set of evaluative principles. It is important to recognize that this approach cannot offer, and does not pretend to offer, such a defense. Instead, the aim of Humean contractarianism is to explain the origin and nature of our evaluative concepts in a way that shows them to find their source in human nature. At the same time, though, the hope is to show that in discovering the origin and nature of evaluative principles we simultaneously show them, thereby, to have a claim on our allegiance. In playing the role they do in social interaction, in serving as the medium (so to speak) through which people can coordinate actions and recommendations and resolve conflicts, our evaluative concepts at least in part earn their own endorsement.

Whether and how the Humean approach might mesh with the Kantian and Hobbesian approaches to contractarianism is unsettled. Those tempted by Humean contractarianism, myself included, suspect that it can offer a philosophically satisfying account of (i) when the Kantian appeal to what people might find mutually agreeable under fair conditions is relevant to determining moral demands (and when it is not) and (ii) why the Hobbesian appeal to our non-moral interest in resolving conflict and coordinating behavior is relevant to morality's demands (but why it does not ultimately limit their scope).

Whether and how contractarianism of any sort might ultimately be defended is also unsettled. However, those tempted by contractarianism suspect that a proper understanding of morality must see morality as a reflection of what those subject to its demands might reasonably accept.

Bibliography

Buchanan, James (1975). *The Limits of Liberty*. Chicago: University of Chicago Press.

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

Gauthier, David (1991). "Why Contractarianism?" In P. Vallentyne (ed.) *Contractarianism and Rational Choice*. New York: Cambridge University Press, 15-30.

Gough, J. W. (1957). *The Social Contract.* Oxford: Clarendon Press. 2nd ed.

Habermas, Jürgen (1990). 'Discourse Ethics: Notes on a Program of Philosophical Justification.' In Christian Lenhardt and Shierry Weber Nicholasen (trans.), *Moral Consciousness and Communicative Action*. Cambridge: MIT Press,

Harman, Gilbert (1978). 'Relativistic Ethics: Morality as Politics'. *Midwest Studies in Philosophy*, III, pp. 109-121.

Harsanyi, John (1953). 'Cardinal Utility in Welfare Economics and the Theory of Risk-Taking'. *Journal of Political Economy*, 61, 309-321.

Harsanyi, John (1976). *Essays on Ethics, Social Behavior, and Scientific Explanation.* Dordrecht: D. Reidel.

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

Hume, David (1985). 'Of the Original Contract'. In Eugene Miller (ed.), *Essays: Moral, Political and Literary*. Indianapolis: LibertyClassics, pp. 465-487.

Kant, Immanuel (1964). *Groundwork of the Metaphysic of Morals*. H. J. Paton (trans.). New York: Harper & Row.

Kant, Immanuel (1970). "On the Common Saying: 'This May be True in Theory, But It Doesn't Apply In Practice," in *Kant's Political Writings*, edited by Hans Reiss. Cambridge: Cambridge University Press.

Lessnoff, Michael (1986) *Social Contract.* New York: Macmillan.

Luce, R. D. and Howard Raiffa (1957). *Games and Decisions*. New York: John Wiley and Sons.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. Marmondsworth: Penguin Books.

Milo, Ronald (1995). 'Contractarian Constructivism'. *Journal of Philosophy*, 92, 181-204.

Pateman, Carole (1988). *The Sexual Contract*. Stanford: Stanford University Press.

Plato (1992). *The Republic*. G.M.A. Grube (trans.) with revisions by C. D. C. Reeve. Indianapolis: Hackett Publishing Company.

Rawls, John (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rawls, John (1993). *Political Liberalism*. New York: Columbia University Press.

Rousseau, Jean-Jacque (1978). *On the Social Contract*. Roger and Judith Masters (trans.). New York: St. Martin's Press.

Sandel, Michael (1982). *Liberalism and the Limits of Justice*. Cambridge: Cambridge University Press.

Sayre-McCord, Geoffrey (1989) "Deception and Reasons to Be Moral," *American Philosophical Quarterly* 26, 113-122.

Sayre-McCord, Geoffrey (1994). "On Why Hume's 'General Point of View' Isn't Ideal -- and Shouldn't Be," *Social Philosophy & Policy*, volume 11, 202-228.

Scanlon, Thomas (1982). 'Contractualism and Utilitarianism'. In Amartya Sen and Bernard Williams (eds.), *Utilitarianism and Beyond.* Cambridge: Cambridge University Press, pp. 103-128.

Scanlon, Thomas (1998). *What We Owe to Each Other.* Cambridge, MA: Harvard University Press.

Skyrms, Brian (1996). *Evolution of the Social Contract.* New York: Cambridge University Press.

Vallentyne, Peter (ed.) (1991). *Contractarianism and Rational Choice*. New York: Cambridge University Press.

ENDNOTES

---

[1] Thanks are due to Robert Goodin, Philip Pettit, Michael Ridge, and Michael Smith for comments on an earlier draft of this essay.

[2] See Luce and Raiffa (1957) for the classic description of the dilemma. Here is the dilemma. Imagine two prisoners find themselves facing the following offer. If neither confesses to the crime they are being charged with, they will both be convicted of some lesser crime (that carries a penalty, let's say, of a year in prison). If they both confess, then both will be convicted of the more serious crime but will receive some leniency for having confessed (so they will each, say, serve five years). But if one confesses and the other doesn't, the person who confesses will get off free with no penalty and the other will serve the maximum sentence (of, say, ten years). Assuming the various years in prison represent costs to the individuals in proportion as they add up, the prisoners face a kind of dilemma: each sees that whether the other person confesses or not she does better to confess, since if the other person confesses she can, by confessing herself, spend only five years, rather than 10, in prison and if the other person doesn't confess she can, by confessing herself, spend no

time at all in prison rather than a year. But if each acts according to this reasoning, they will together confess themselves into five years of jail each rather than the one that they would have been sentenced had they both kept quiet. Yet as soon as one has reason to think the other will not confess she finds herself again with compelling reason to confess... The structure of the dilemma remains even if the costs and benefits at stake are radically different and regardless of whether they represent the selfish concern of the criminal simply to stay out of jail or the selfless preoccupation with worries about the welfare of her children. Moreover, assuming the pay-offs have the Prisoners' Dilemma structure, prior promises to remain silent leave the dilemma in place. What is needed to solve it is either something that will change the pay-offs so as to eliminate the dilemma or grounds for reasoning in some way that breaks free from simply maximizing expected utility.