2. Why contractarianism?*

David Gauthier

I

As the will to truth thus gains self-consciousness – there can be no doubt of that – morality will gradually *perish* now: this is the great spectacle in a hundred acts reserved for the next two centuries in Europe – the most terrible, most questionable, and perhaps also the most hopeful of all spectacles.

– Nietzsche[1]

Morality faces a foundational crisis. Contractarianism offers the only plausible resolution of this crisis. These two propositions state my theme. What follows is elaboration.

Nietzsche may have been the first, but he has not been alone, in recognizing the crisis to which I refer. Consider these recent statements. "The hypothesis which I wish to advance is that in the actual world which we inhabit the language of morality is in . . . [a] state of grave disorder . . . we have – very largely, if not entirely – lost our comprehension, both theoretical and practical, of morality" (Alasdair MacIntyre).[2] "The resources of most modern moral philosophy are not well adjusted to the modern world" (Bernard Williams).[3] "There are no ob-

*Two paragraphs of Section II and most of Section IV are taken from "Morality, Rational Choice, and Semantic Representation – A Reply to My Critics," in E. F. Paul, F. D. Miller, Jr., and J. Paul (eds.), *The New Social Contract: Essays on Gauthier* (Oxford: Blackwell, 1988), pp. 173–4, 179–180, 184–5, 188–9 (this volume appears also as *Social Philosophy and Policy* 5 [1988], same pagination). I am grateful to Annette Baier, Paul Hurley, and Geoffrey Sayre-McCord for comments on an earlier draft. I am also grateful to discussants at Western Washington University, the University of Arkansas, the University of California at Santa Cruz, and the University of East Anglia for comments on a related talk.

[1] *On the Genealogy of Morals*, trans. by Walter Kaufmann and R. J. Hollingdale (New York: Random House, 1967), third essay, sec. 27, p. 161.

[2] *After Virtue* (Notre Dame, IN: University of Notre Dame Press, 1981), p. 2.

[3] *Ethics and the Limits of Philosophy* (Cambridge, MA: Harvard University Press, 1985), p. 197.

jective values. . . . [But] the main tradition of European moral philosophy includes the contrary claim" (J. L. Mackie).[4] "Moral hypotheses do not help explain why people observe what they observe. So ethics is problematic and nihilism must be taken seriously. . . . An extreme version of nihilism holds that morality is simply an illusion. . . . In this version, we should abandon morality, just as an atheist abandons religion after he has decided that religious facts cannot help explain observations" (Gilbert Harman).[5]

I choose these statements to point to features of the crisis that morality faces. They suggest that moral language fits a world view that we have abandoned – a view of the world as purposively ordered. Without this view, we no longer truly understand the moral claims we continue to make. They suggest that there is a lack of fit between what morality presupposes – objective values that help explain our behavior, and the psychological states – desires and beliefs – that, given our present world view, actually provide the best explanation. This lack of fit threatens to undermine the very idea of a morality as more than an anthropological curiosity. But how could this be? How could morality *perish*?

II

To proceed, I must offer a minimal characterization of the morality that faces a foundational crisis. And this is the morality of justified constraint. From the standpoint of the agent, moral considerations present themselves as constraining his choices and actions, in ways independent of his desires, aims, and interests. Later, I shall add to this characterization, but for the moment it will suffice. For it reveals clearly what is in question – the ground of constraint. This ground seems absent from our present world view. And so we ask, what reason can a person have for recognizing and accepting a constraint that is independent of his desires and interests? He may agree that such a constraint would be *morally* justified; he would have a reason for accepting it *if* he had a reason for accepting morality. But what justifies paying attention to morality, rather than dismissing it as an appendage of outworn beliefs? We ask, and seem to find no answer. But before proceeding, we should consider three objections.

The first is to query the idea of constraint. Why should morality be seen as constraining our choices and actions? Why should we not rather say that the moral person chooses most freely, because she chooses in the light of a true conception of herself, rather than in the light of the false conceptions that so often predominate? Why should we not link

[4] *Ethics: Inventing Right and Wrong* (Harmondsworth: Penguin, 1977), pp. 15, 30.
[5] *The Nature of Morality* (New York: Oxford University Press, 1977), p. 11.

morality with self-understanding? Plato and Hume might be enlisted to support this view, but Hume would be at best a partial ally, for his representation of "virtue in all her genuine and most engaging charms, . . . talk[ing] not of useless austerities and rigors, suffering and self-denial," but rather making "her votaries . . . , during every instant of their existence, if possible, cheerful and happy," is rather overcast by his admission that "in the case of justice, . . . a man, taking things in a certain light, may often seem to be a loser by his integrity."[6] Plato, to be sure, goes further, insisting that only the just man has a healthy soul, but heroic as Socrates' defense of justice may be, we are all too apt to judge that Glaucon and Adeimantus have been charmed rather than reasoned into agreement, and that the unjust man has not been shown necessarily to be the loser.[7] I do not, in any event, intend to pursue this direction of thought. Morality, as we, heirs to the Christian and Kantian traditions, conceive it, constrains the pursuits to which even our reflective desires would lead us. And this is not simply or entirely a constraint on self-interest; the affections that morality curbs include the social ones of favoritism and partiality, to say nothing of cruelty.

The second objection to the view that moral constraint is insufficiently grounded is to query the claim that it operates independently of, rather than through, our desires, interests, and affections. Morality, some may say, concerns the well-being of all persons, or perhaps of all sentient creatures.[8] And one may then argue, either with Hume, that morality arises in and from our sympathetic identification with our fellows, or that it lies directly in well-being, and that our affections tend to be disposed favorably toward it. But, of course, not all of our affections. And so our sympathetic feelings come into characteristic opposition to other feelings, in relation to which they function as a constraint.

This is a very crude characterization, but it will suffice for the present argument. This view grants that morality, as we understand it, is without purely *rational* foundations, but reminds us that we are not therefore unconcerned about the well-being of our fellows. Morality is founded on the widespread, sympathetic, other-directed concerns that most of us have, and these concerns do curb self-interest, and also the favoritism and partiality with which we often treat others. Nevertheless, if morality depends for its practical relevance and motivational efficacity entirely on our sympathetic feelings, it has no title to the prescriptive grip with which it has been invested in the Christian and Kantian views to which I have referred, and which indeed Glaucon and Adeimantus demanded

[6] David Hume, *An Equiry Concerning the Principles of Morals*, 1751, sec. IX, pt. II.
[7] See Plato, *Republic*, esp. books II and IV.
[8] Some would extend morality to the nonsentient, but sympathetic as I am to the rights of trolley cars and steam locomotives, I propose to leave this view quite out of consideration.

that Socrates defend to them in the case of justice. For to be reminded that some of the time we do care about our fellows and are willing to curb other desires in order to exhibit that care tells us nothing that can guide us in those cases in which, on the face of it, we do not care, or do not care enough – nothing that will defend the demands that morality makes on us in the hard cases. That not all situations in which concern for others combats self-concern are hard cases is true, but morality, as we ordinarily understand it, speaks to the hard cases, whereas its Humean or naturalistic replacement does not.

These remarks apply to the most sustained recent positive attempt to create a moral theory – that of John Rawls. For the attempt to describe our moral capacity, or more particularly, for Rawls, our sense of justice, in terms of principles, plausible in the light of our more general psychological theory, and coherent with "our considered judgments in reflective equilibrium,"[9] will not yield any answer to why, in those cases in which we have no, or insufficient, interest in being just, we should nevertheless follow the principles. John Harsanyi, whose moral theory is in some respects a utilitarian variant of Rawls' contractarian construction, recognizes this explicitly: "All we can prove by rational arguments is that anybody who wants to serve our common human interests in a rational manner must obey these commands."[10] But although morality may offer itself in the service of our common human interests, it does not offer itself only to those who want to serve them.

Morality is a constraint that, as Kant recognized, must not be supposed to depend solely on our feelings. And so we may not appeal to feelings to answer the question of its foundation. But the third objection is to dismiss this question directly, rejecting the very idea of a foundational crisis. Nothing justifies morality, for morality needs no justification. We find ourselves, in morality as elsewhere, in mediis rebus. We make, accept and reject, justify and criticize moral judgments. The concern of moral theory is to systematize that practice, and so to give us a deeper understanding of what moral justification is. But there are no extramoral foundations for moral justification, any more than there are extraepistemic foundations for epistemic judgments. In morals as in science, foundationalism is a bankrupt project.

Fortunately, I do not have to defend *normative* foundationalism. One problem with accepting moral justification as part of our ongoing practice is that, as I have suggested, we no longer accept the world view on which it depends. But perhaps a more immediately pressing problem is that we have, ready to hand, an alternative mode for justifying our

[9] John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971), p. 51.
[10] John C. Harsanyi, "Morality and the Theory of Rational Behaviour," in *Utilitarianism and Beyond*, edited by Amartya Sen and Bernard Williams (Cambridge: Cambridge University Press, 1982), p. 62.

choices and actions. In its more austere and, in my view, more defensible form, this is to show that choices and actions maximize the agent's expected utility, where utility is a measure of considered preference. In its less austere version, this is to show that choices and actions satisfy, not a subjectively defined requirement such as utility, but meet the agent's objective interests. Since I do not believe that we have objective interests, I shall ignore this latter. But it will not matter. For the idea is clear; we have a mode of justification that does not require the introduction of moral considerations.[11]

Let me call this alternative nonmoral mode of justification, neutrally, deliberative justification. Now moral and deliberative justification are directed at the same objects – our choices and actions. What if they conflict? And what do we say to the person who offers a deliberative justification of his choices and actions and refuses to offer any other? We can say, of course, that his behavior lacks *moral* justification, but this seems to lack any hold, unless he chooses to enter the moral framework. And such entry, he may insist, lacks any deliberative justification, at least for him.

If morality perishes, the justificatory enterprise, in relation to choice and action, does not perish with it. Rather, one mode of justification perishes, a mode that, it may seem, now hangs unsupported. But not only unsupported, for it is difficult to deny that deliberative justification is more clearly basic, that it cannot be avoided insofar as we are rational agents, so that if moral justification conflicts with it, morality seems not only unsupported but opposed by what is rationally more fundamental.

Deliberative justification relates to our deep sense of self. What distinguishes human beings from other animals, and provides the basis for rationality, is the capacity for semantic representation. You can, as your dog on the whole cannot, represent a state of affairs to yourself, and consider in particular whether or not it is the case, and whether or not you would want it to be the case. You can represent to yourself the contents of your beliefs, and your desires or preferences. But in representing them, you bring them into relation with one another. You represent to yourself that the Blue Jays will win the World Series, and that a National League team will win the World Series, and that the Blue Jays are not a National League team. And in recognizing a conflict among those beliefs, you find rationality thrust upon you. Note that the first two beliefs could be replaced by preferences, with the same effect.

Since in representing our preferences we become aware of conflict among them, the step from representation to choice becomes compli-

[11] To be sure, if we think of morality as expressed in certain of our affections and/or interests, it will incorporate moral considerations to the extent that they actually are present in our preferences. But this would be to embrace the naturalism that I have put to one side as inadequate.

cated. We must, somehow, bring our conflicting desires and preferences into some sort of coherence. And there is only one plausible candidate for a principle of coherence – a maximizing principle. We order our preferences, in relation to decision and action, so that we may choose in a way that maximizes our expectation of preference fulfillment. And in so doing, we show ourselves to be rational agents, engaged in deliberation and deliberative justification. There is simply nothing else for practical rationality to be.

The foundational crisis of morality thus cannot be avoided by pointing to the existence of a practice of justification within the moral framework, and denying that any extramoral foundation is relevant. For an extramoral mode of justification is already present, existing not side by side with moral justification, but in a manner tied to the way in which we unify our beliefs and preferences and so acquire our deep sense of self. We need not suppose that this deliberative justification is itself to be understood foundationally. All that we need suppose is that moral justification does not plausibly survive conflict with it.

III

In explaining why we may not dismiss the idea of a foundational crisis in morality as resulting from a misplaced appeal to a philosophically discredited or suspect idea of foundationalism, I have begun to expose the character and dimensions of the crisis. I have claimed that morality faces an alternative, conflicting, deeper mode of justification, related to our deep sense of self, that applies to the entire realm of choice and action, and that evaluates each *action* in terms of the reflectively held concerns of its *agent*. The relevance of the agent's concerns to practical justification does not seem to me in doubt. The relevance of anything else, except insofar as it bears on the agent's concerns, does seem to me very much in doubt. If the agent's reflectively endorsed concerns, his preferences, desires, and aims, are, with his considered beliefs, constitutive of his self-conception, then I can see no remotely plausible way of arguing from their relevance to that of anything else that is not similarly related to his sense of self. And, indeed, I can see no way of introducing anything as relevant to practical justification except through the agent's self-conception. My assertion of this practical individualism is not a conclusive argument, but the burden of proof is surely on those who would maintain a contrary position. Let them provide the arguments – if they can.

Deliberative justification does not refute morality. Indeed, it does not offer morality the courtesy of a refutation. It ignores morality, and seemingly replaces it. It preempts the arena of justification, apparently leaving morality no room to gain purchase. Let me offer a controversial comparison. Religion faces – indeed, has faced – a comparable foundational crisis. Religion demands the worship of a divine being who purposively orders the universe. But it has confronted an alternative mode of explanation. Although the emergence of a cosmological theory based on efficient, rather than teleological, causation provided warning of what was to come, the supplanting of teleology in biology by the success of evolutionary theory in providing a mode of explanation that accounted in efficient-causal terms for the *appearance* of a purposive order among living beings, may seem to toll the death knell for religion as an intellectually respectable enterprise. But evolutionary biology and, more generally, modern science do not refute religion. Rather they ignore it, replacing its explanations by ontologically simpler ones. Religion, understood as affirming the justifiable worship of a divine being, may be unable to survive its foundational crisis. Can morality, understood as affirming justifiable constraints on choice independent of the agent's concerns, survive?

There would seem to be three ways for morality to escape religion's apparent fate. One would be to find, for moral facts or moral properties, an explanatory role that would entrench them prior to any consideration of justification.[12] One could then argue that any mode of justification that ignored moral considerations would be ontologically defective. I mention this possibility only to put it to one side. No doubt there are persons who accept moral constraints on their choices and actions, and it would not be possible to explain those choices and actions were we to ignore this. But our explanation of their behavior need not commit us to their view. Here the comparison with religion should be straightforward and uncontroversial. We could not explain many of the practices of the religious without reference to their beliefs. But to characterize what a religious person is doing as, say, an act of worship, does not commit us to supposing that an object of worship actually exists, though it does commit us to supposing that she believes such an object to exist. Similarly, to characterize what a moral agent is doing as, say, fulfilling a duty does not commit us to supposing that there are any duties, though it does commit us to supposing that he believes that there are duties. The skeptic who accepts neither can treat the apparent role of morality in explanation as similar to that of religion. Of course, I do not consider that the parallel can be ultimately sustained, since I agree with the religious skeptic but not with the moral skeptic. But to establish an explanatory role for morality, one must first demonstrate its justificatory credentials. One may not assume that it has a prior explanatory role.

The second way would be to reinterpret the idea of justification, show-

[12] This would meet the challenge to morality found in my previous quotation from Gilbert Harman.

ing that, more fully understood, deliberative justification is incomplete, and must be supplemented in a way that makes room for morality. There is a long tradition in moral philosophy, deriving primarily from Kant, that is committed to this enterprise. This is not the occasion to embark on a critique of what, in the hope again of achieving a neutral characterization, I shall call universalistic justification. But critique may be out of place. The success of deliberative justification may suffice. For theoretical claims about its incompleteness seem to fail before the simple practical recognition that it works. Of course, on the face of it, deliberative justification does not work to provide a place for morality. But to suppose that it must, if it is to be fully adequate or complete as a mode of justification, would be to assume what is in question, whether moral justification is defensible.

If, independent of one's actual desires, and aims, there were objective values, and if, independent of one's actual purposes, one were part of an objectively purposive order, then we might have reason to insist on the inadequacy of the deliberative framework. An objectively purposive order would introduce considerations relevant to practical justification that did not depend on the agent's self-conception. But the supplanting of teleology in our physical and biological explanations closes this possibility, as it closes the possibility of religious explanation.

I turn then to the third way of resolving morality's foundational crisis. The first step is to embrace deliberative justification, and recognize that morality's place must be found within, and not outside, its framework. Now this will immediately raise two problems. First of all, it will seem that the attempt to establish any constraint on choice and action, within the framework of a deliberation that aims at the maximal fulfillment of the agent's considered preferences, must prove impossible. But even if this be doubted, it will seem that the attempt to establish a constraint *independent of the agent's preferences*, within such a framework, verges on lunacy. Nevertheless, this is precisely the task accepted by my third way. And, unlike its predecessors, I believe that it can be successful; indeed, I believe that my recent book, *Morals by Agreement*, shows how it can succeed.[13]

I shall not rehearse at length an argument that is now familiar to at least some readers, and, in any event, can be found in that book. But let me sketch briefly those features of deliberative rationality that enable it to constrain maximizing choice. The key idea is that in many situations, if each person chooses what, given the choices of the others, would maximize her expected utility, then the outcome will be mutually disadvantageous in comparison with some alternative – everyone could do

[13] See David Gauthier, *Morals by Agreement* (Oxford: Oxford University Press, 1986), especially chaps. V and VI.

better.[14] Equilibrium, which obtains when each person's action is a best response to the others' actions, is incompatible with (Pareto-)optimality, which obtains when no one could do better without someone else doing worse. Given the ubiquity of such situations, each person can see the benefit, to herself, of participating with her fellows in practices requiring each to refrain from the direct endeavor to maximize her own utility, when such mutual restraint is mutually advantageous. No one, of course, can have reason to accept any unilateral constraint on her maximizing behavior; each benefits from, and only from, the constraint accepted by her fellows. But if one benefits more from a constraint on others than one loses by being constrained oneself, one may have reason to accept a practice requiring everyone, including oneself, to exhibit such a constraint. We may represent such a practice as capable of gaining unanimous agreement among rational persons who were choosing the terms on which they would interact with each other. And this agreement is the basis of morality.

Consider a simple example of a moral practice that would command rational agreement. Suppose each of us were to assist her fellows only when either she could expect to benefit herself from giving assistance, or she took a direct interest in their well-being. Then, in many situations, persons would not give assistance to others, even though the benefit to the recipient would greatly exceed the cost to the giver, because there would be no provision for the giver to share in the benefit. Everyone would then expect to do better were each to give assistance to her fellows, regardless of her own benefit or interest, whenever the cost of assisting was low and the benefit of receiving assistance considerable. Each would thereby accept a constraint on the direct pursuit of her own concerns, not unilaterally, but given a like acceptance by others. Reflection leads us to recognize that those who belong to groups whose members adhere to such a practice of mutual assistance enjoy benefits in interaction that are denied to others. We may then represent such a practice as rationally acceptable to everyone.

This rationale for agreed constraint makes no reference to the content of anyone's preferences. The argument depends simply on the *structure* of interaction, on the way in which each person's endeavor to fulfill her own preferences affects the fulfillment of everyone else. Thus, each person's reason to accept a mutually constraining practice is independent of her particular desires, aims and interests, although not, of course, of the fact that she has such concerns. The idea of a purely rational agent,

[14] The now-classic example of this type of situation is the Prisoner's Dilemma; see *Morals by Agreement*, pp. 79–80. More generally, such situations may be said, in economists' parlance, to exhibit market failure. See, for example, "Market Contractarianism" in Jules Coleman, *Markets, Morals, and the Law* (Cambridge: Cambridge University Press, 1988), chap. 10.

moved to act by reason alone, is not, I think, an intelligible one. Morality is not to be understood as a constraint arising from reason alone on the fulfillment of nonrational preferences. Rather, a rational agent is one who acts to achieve the maximal fulfillment of her preferences, and morality is a constraint on the manner in which she acts, arising from the effects of interaction with other agents.

Hobbes's Foole now makes his familiar entry onto the scene, to insist that however rational it may be for a person to agree with her fellows to practices that hold out the promise of mutual advantage, yet it is rational to follow such practices only when so doing directly conduces to her maximal preference fulfillment.[15] But then such practices impose no real constraint. The effect of agreeing to or accepting them can only be to change the expected payoffs of her possible choices, making it rational for her to choose what in the absence of the practice would not be utility maximizing. The practices would offer only true prudence, not true morality.

The Foole is guilty of a twofold error. First, he fails to understand that real acceptance of such moral practices as assisting one's fellows, or keeping one's promises, or telling the truth is possible only among those who are disposed to comply with them. If my disposition to comply extends only so far as my interests or concerns at the time of performance, then you will be the real fool if you interact with me in ways that demand a more rigorous compliance. If, for example, it is rational to keep promises only when so doing is directly utility maximizing, then among persons whose rationality is common knowledge, only promises that require such limited compliance will be made. And opportunities for mutual advantage will be thereby forgone.

Consider this example of the way in which promises facilitate mutual benefit. Jones and Smith have adjacent farms. Although neighbors, and not hostile, they are also not friends, so that neither gets satisfaction from assisting the other. Nevertheless, they recognize that, if they harvest their crops together, each does better than if each harvests alone. Next week, Jones's crop will be ready for harvesting; a fortnight hence, Smith's crop will be ready. The harvest in, Jones is retiring, selling his farm, and moving to Florida, where he is unlikely to encounter Smith or other members of their community. Jones would like to promise Smith that, if Smith helps him harvest next week, he will help Smith harvest in a fortnight. But Jones and Smith both know that in a fortnight, helping Smith would be a pure cost to Jones. Even if Smith helps him, he has nothing to gain by returning the assistance, since neither care for Smith nor, in the circumstances, concern for his own reputation, moves him. Hence, if Jones and Smith know that Jones acts straightforwardly to

[15] See Hobbes, *Leviathan*, London, 1651, chap. 15.

maximize the fulfillment of his preferences, they know that he will not help Smith. Smith, therefore, will not help Jones even if Jones pretends to promise assistance in return. Nevertheless, Jones would do better could he make and keep such a promise – and so would Smith.

The Foole's second error, following on his first, should be clear; he fails to recognize that in plausible circumstances, persons who are genuinely disposed to a more rigorous compliance with moral practices than would follow from their interests at the time of performance can expect to do better than those who are not so disposed. For the former, constrained maximizers as I call them, will be welcome partners in mutually advantageous cooperation, in which each relies on the voluntary adherence of the others, from which the latter, straightforward maximizers, will be excluded. Constrained maximizers may thus expect more favorable opportunities than their fellows. Although in assisting their fellows, keeping their promises, and complying with other moral practices, they forgo preference fulfillment that they might obtain, yet they do better overall than those who always maximize expected utility, because of their superior opportunities.

In identifying morality with those constraints that would obtain agreement among rational persons who were choosing their terms of interaction, I am engaged in rational reconstruction. I do not suppose that we have actually agreed to existent moral practices and principles. Nor do I suppose that all existent moral practices would secure our agreement, were the question to be raised. Not all existent moral practices need be justifiable – need be ones with which we ought willingly to comply. Indeed, I do not even suppose that the practices with which we ought willingly to comply need be those that would secure our present agreement. I suppose that justifiable moral practices are those that would secure our agreement ex ante, in an appropriate premoral situation. They are those to which we should have agreed as constituting the terms of our future interaction, had we been, per impossible, in a position to decide those terms. Hypothetical agreement thus provides a test of the justifiability of our existent moral practices.

IV

Many questions could be raised about this account, but here I want to consider only one. I have claimed that moral practices are rational, even though they constrain each person's attempt to maximize her own utility, insofar as they would be the objects of unanimous ex ante agreement. But to refute the Foole, I must defend not only the rationality of agreement, but also that of compliance, and the defense of compliance threatens to preempt the case for agreement, so that my title should be "Why Constraint?" and not "Why Contractarianism?" It is rational to dispose

oneself to accept certain constraints on direct maximization in choosing and acting, if and only if so disposing oneself maximizes one's expected utility. What then is the relevance of agreement, and especially of hypothetical agreement? Why should it be rational to dispose oneself to accept only those constraints that would be the object of mutual agreement in an appropriate premoral situation, rather than those constraints that are found in our existent moral practices? Surely it is acceptance of the latter that makes a person welcome in interaction with his fellows. For compliance with existing morality will be what they expect, and take into account in choosing partners with whom to cooperate.

I began with a challenge to morality – how can it be rational for us to accept its constraints? It may now seem that what I have shown is that it is indeed rational for us to accept constraints, but to accept them whether or not they might be plausibly considered moral. Morality, it may seem, has nothing to do with my argument; what I have shown is that it is rational to be disposed to comply with whatever constraints are generally accepted and expected, regardless of their nature. But this is not my view.

To show the relevance of agreement to the justification of constraints, let us assume an ongoing society in which individuals more or less acknowledge and comply with a given set of practices that constrain their choices in relation to what they would be did they take only their desires, aims, and interests directly into account. Suppose that a disposition to conform to these existing practices is prima facie advantageous, since persons who are not so disposed may expect to be excluded from desirable opportunities by their fellows. However, the practices themselves have, or at least need have, no basis in agreement. And they need satisfy no intuitive standard of fairness or impartiality, characteristics that we may suppose relevant to the identification of the practices with those of a genuine morality. Although we may speak of the practices as constituting the morality of the society in question, we need not consider them morally justified or acceptable. They are simply practices constraining individual behavior in a way that each finds rational to accept.

Suppose now that our persons, as rational maximizers of individual utility, come to reflect on the practices constituting their morality. They will, of course, assess the practices in relation to their own utility, but with the awareness that their fellows will be doing the same. And one question that must arise is: Why these practices? For they will recognize that the set of actual moral practices is not the only possible set of constraining practices that would yield mutually advantageous, optimal outcomes. They will recognize the possibility of alternative moral orders. At this point it will not be enough to say that, as a matter of fact, each person can expect to benefit from a disposition to comply with existing

practices. For persons will also ask themselves: Can I benefit more, not from simply abandoning any morality, and recognizing no constraint, but from a partial rejection of existing constraints in favor of an alternative set? Once this question is asked, the situation is transformed; the existing moral order must be assessed, not only against simple noncompliance, but also against what we may call alternative compliance.

To make this assessment, each will compare her prospects under the existing practices with those she would anticipate from a set that, in the existing circumstances, she would expect to result from bargaining with her fellows. If her prospects would be improved by such negotiation, then she will have a real, although not necessarily sufficient, incentive to demand a change in the established moral order. More generally, if there are persons whose prospects would be improved by renegotiation, then the existing order will be recognizably unstable. No doubt those whose prospects would be worsened by renegotiation will have a clear incentive to resist, to appeal to the status quo. But their appeal will be a weak one, especially among persons who are not taken in by spurious ideological considerations, but focus on individual utility maximization. Thus, although in the real world, we begin with an existing set of moral practices as constraints on our maximizing behavior, yet we are led by reflection to the idea of an amended set that would obtain the agreement of everyone, and this amended set has, and will be recognized to have, a stability lacking in existing morality.

The reflective capacity of rational agents leads them from the given to the agreed, from existing practices and principles requiring constraint to those that would receive each person's assent. The same reflective capacity, I claim, leads from those practices that would be agreed to, in existing social circumstances, to those that would receive ex ante agreement, premoral and presocial. As the status quo proves unstable when it comes into conflict with what would be agreed to, so what would be agreed to proves unstable when it comes into conflict with what would have been agreed to in an appropriate presocial context. For as existing practices must seem arbitrary insofar as they do not correspond to what a rational person would agree to, so what such a person would agree to in existing circumstances must seem arbitrary in relation to what she would accept in a presocial condition.

What a rational person would agree to in existing circumstances depends in large part on her negotiating position vis-à-vis her fellows. But her negotiating position is significantly affected by the existing social institutions, and so by the currently accepted moral practices embodied in those institutions. Thus, although agreement may well yield practices differing from those embodied in existing social institutions, yet it will be influenced by those practices, which are not themselves the product of rational agreement. And this must call the rationality of the agreed

practices into question. The arbitrariness of existing practices must infect any agreement whose terms are significantly affected by them. Although rational agreement is in itself a source of stability, yet this stability is undermined by the arbitrariness of the circumstances in which it takes place. To escape this arbitrariness, rational persons will revert from actual to hypothetical agreement, considering what practices they would have agreed to from an initial position not structured by existing institutions and the practices they embody.

The content of a hypothetical agreement is determined by an appeal to the equal rationality of persons. Rational persons will voluntarily accept an agreement only insofar as they perceive it to be equally advantageous to each. To be sure, each would be happy to accept an agreement more advantageous to herself than to her fellows, but since no one will accept an agreement perceived to be less advantageous, agents whose rationality is a matter of common knowledge will recognize the futility of aiming at or holding out for more, and minimize their bargaining costs by coordinating at the point of equal advantage. Now the extent of advantage is determined in a twofold way. First, there is advantage internal to an agreement. In this respect, the expectation of equal advantage is assured by procedural fairness. The step from existing moral practices to those resulting from actual agreement takes rational persons to a procedurally fair situation, in which each perceives the agreed practices to be ones that it is equally rational for all to accept, given the circumstances in which agreement is reached. But those circumstances themselves may be called into question insofar as they are perceived to be arbitrary – the result, in part, of compliance with constraining practices that do not themselves ensure the expectation of equal advantage, and so do not reflect the equal rationality of the complying parties. To neutralize this arbitrary element, moral practices to be fully acceptable must be conceived as constituting a possible outcome of a hypothetical agreement under circumstances that are unaffected by social institutions that themselves lack full acceptability. Equal rationality demands consideration of external circumstances as well as internal procedures.

But what is the practical import of this argument? It would be absurd to claim that mere acquaintance with it, or even acceptance of it, will lead to the replacement of existing moral practices by those that would secure presocial agreement. It would be irrational for anyone to give up the benefits of the existing moral order simply because he comes to realize that it affords him more than he could expect from pure rational agreement with his fellows. And it would be irrational for anyone to accept a long-term utility loss by refusing to comply with the existing moral order, simply because she comes to realize that such compliance affords her less than she could expect from pure rational agreement.

Nevertheless, these realizations do transform, or perhaps bring to the surface, the character of the relationships between persons that are maintained by the existing constraints, so that some of these relationships come to be recognized as coercive. These realizations constitute the elimination of false consciousness, and they result from a process of rational reflection that brings persons into what, in my theory, is the parallel of Jürgen Habermas's ideal speech situation.[16] Without an argument to defend themselves in open dialogue with their fellows, those who are more than equally advantaged can hope to maintain their privileged position only if they can coerce their fellows into accepting it. And this, of course, may be possible. But coercion is not agreement, and it lacks any inherent stability.

Stability plays a key role in linking compliance to agreement. Aware of the benefits to be gained from constraining practices, rational persons will seek those that invite stable compliance. Now compliance is stable if it arises from agreement among persons each of whom considers both that the terms of agreement are sufficiently favorable to herself that it is rational for her to accept them, and that they are not so favorable to others that it would be rational for them to accept terms less favorable to them and more favorable to herself. An agreement affording equally favorable terms to all thus invites, as no other can, stable compliance.

V

In defending the claim that moral practices, to obtain the stable voluntary compliance of rational individuals, must be the objects of an appropriate hypothetical agreement, I have added to the initial minimal characterization of morality. Not only does morality constrain our choices and actions, but it does so in an impartial way, reflecting the equal rationality of the persons subject to constraint. Although it is no part of my argument to show that the requirements of contractarian morality will satisfy the Rawlsian test of cohering with our considered judgments in reflective equilibrium, yet it would be misleading to treat rationally agreed constraints on direct utility maximization as constituting a morality at all, rather than as replacing morality, were there no fit between their content and our pretheoretical moral views. The fit lies, I suggest, in the impartiality required for hypothetical agreement.

The foundational crisis of morality is thus resolved by exhibiting the rationality of our compliance with mutual, rationally agreed constraints on the pursuit of our desires, aims, and interests. Although bereft of a basis in objective values or an objectively purposive order, and con-

[16] See Raymond Geuss, *The Idea of a Critical Theory: Habermas and the Frankfurt School* (Cambridge: Cambridge University Press, 1981), p. 65ff.

fronted by a more fundamental mode of justification, morality survives by incorporating itself into that mode. Moral considerations have the same status, and the same role in explaining behavior, as the other reasons acknowledged by a rational deliberator. We are left with a unified account of justification, in which an agent's choices and actions are evaluated in relation to his preferences – to the concerns that are constitutive of his sense of self. But since morality binds the agent independently of the particular content of his preferences, it has the prescriptive grip with which the Christian and Kantian views have invested it.

In incorporating morality into deliberative justification, we recognize a new dimension to the agent's self-conception. For morality requires that a person have the capacity to commit himself, to enter into agreement with his fellows secure in the awareness that he can and will carry out his part of the agreement without regard to many of those considerations that normally and justifiably would enter into his future deliberations. And this is more than the capacity to bring one's desires and interests together with one's beliefs into a single coherent whole. Although this latter unifying capacity must extend its attention to past and future, the unification it achieves may itself be restricted to that extended present within which a person judges and decides. But in committing oneself to future action in accordance with one's agreement, one must fix at least a subset of one's desires and beliefs to hold in that future. The self that agrees and the self that complies must be one. "Man himself must first of all have become *calculable, regular, necessary,* even in his own image of himself, if he is to be able to stand security for *his own future,* which is what one who promises does!"[17]

In developing *"the right to make promises,"*[18] we human beings have found a contractarian bulwark against the perishing of morality.

[17] Nietzsche, *On the Genealogy of Morals,* trans. by Walter Kaufmann and R. J. Hollingdale (New York: Random House, 1967), second essay, sec. 1, p. 58.
[18] Ibid., p. 57.