I

This paper explores a certain class of "reductive" moral philosophies. Generally speaking, reductive accounts attempt to analyze the nature and content of morality without relying on evaluative concepts, or at least none other than those employed in the natural and social sciences. The aim is to describe what people are doing when they engage in moral evaluation and what would give them reason to accept a set of norms as justified. At the same time, the hope is to avoid endorsing any substantive position within ethics. Moral philosophy thus attempts to be less controversial and more objective. Jürgen Habermas, for instance, criticizes John Rawls's aim of elaborating an idea of a just society that citizens could use to evaluate existing political arrangements and policies. Habermas proposes instead "that philosophy limit itself to the clarification of the moral point of view and the procedure of democratic legitimation, to the analysis of the conditions of rational discourses and negotiations. . . . It leaves substantial questions that must be answered here and now to the more or less enlightened engagement of participants, which does not mean that philosophers may not also participate in the public debate, though in the role of intellectuals, not of experts."³

Habermas's concern is that philosophy not encroach into the province of democracy, but his conclusion is echoed by developments in analytic
meta-ethics that have been motivated by other considerations. Analytic moral philosophers have emphasized a division between the project of clarifying the logic and meaning of our moral concepts, on the one hand, and the task of a moral theory that draws on and attempts to systematize our substantive moral intuitions, on the other. The division makes possible disinterested inquiry within the former domain, or so it has been thought. The enterprise of meta-ethics came to be identified with this ambition and is characterized by several prominent positions. In response to G. E. Moore's open question argument and in view of the development of causal theories of meaning and reference, Richard Boyd and others formulated a variety of moral naturalism that they refer to as "moral realism." Moore argued that naturalistic definitions of ethical concepts will fail to be adequate since questions about the normative significance of any set of natural properties remain open. Boyd argues that our normative notions may refer to natural properties that help to explain our moral practices and conclusions whether or not we identify our notions with those properties. Thus empirically identifiable properties may constitute moral facts even though questions about whether those properties are normatively significant seem open.

Emotivists and other reductive noncognitivists took a different approach in response to Moore. They disputed the apparently cognitive content of our evaluative judgments by maintaining that moral judgment consists not in a cognitive state but in the expression of a feeling or of a person's disposition to have certain feelings. This finding bodes well for the prospects of an objective meta-ethical analysis, since the relevant feelings and dispositions can be described from an evasively neutral stance, or so these philosophers have maintained. Moral philosophers could then inform and expand their analysis by utilizing research in psychology and evolutionary biology.

Emotivists have notoriously had trouble analyzing moral disagreement and understanding how disagreement may be addressed through an exchange of reasons. Allan Gibbard's expressivism is more sophisticated. He proposes that moral judgments express our acceptance of certain norms, norms that we consult to determine whether certain reactive feelings are warranted. Norms can be discussed and critically evaluated. Nevertheless, it is not clear, on Gibbard's account, how fully the grounds for our acceptance of norms could be evaluated for their reasonableness or how disagreement about which norms we ought to accept could be rationally resolved. The ultimate ground for a claim of rationality, on Gibbard's analysis, is our disposition to accept a norm that supports it, and

this is largely a matter of feelings we in fact experience. What seems to be missing is an analysis of the ethical importance of responding to moral disagreement by offering considerations we think others have reason to accept. We want to know what counts as a plausible basis for arguing that mutual acceptance of proposed norms would be reasonable. The relevant question may be posed in various ways. We want to know what reasons we have to accept norms we would be disposed to accept, why we ought to accept them, or why our acceptance of them is justified. Christine Korsgaard calls this the normative question.

Moral realists might seem better equipped to engage the normative question and, in particular, to understand how it is that moral deliberation could issue in justifiable results. They can affirm that parties in disagreement address a common question and may reason together about it. Furthermore, they maintain that we can describe the cognitive activity of moral deliberation without employing evaluative concepts and entering the debate. We can instead position ourselves to draw upon the resources of the natural and social sciences in order to identify the causes and effects of this activity. Using empirical methods we may arrive at the content of justifiable norms. In this way, the relevant natural and social facts would serve to clarify and to demystify the requirements of morality without having to explain away its apparently cognitive character. Appreciation of the relevant empirical facts could provide us with, as Jeremy Bentham put it, a "clue to the labyrinth."

In order to engage the normative question, moral philosophers must offer a plausible description of the normative activity of moral reasoning and defend its criteria of justification. Throughout this paper I will understand moral reasoning as an activity that aims to arrive at mutually acceptable norms to govern social life. Persons who care about morality are concerned to establish the interpersonal justification of shared norms: each person should take the interests of others into account in order to arrive at principles that could be accepted by all. The reductive accounts that I will be concerned with all aim to provide a naturalistic understanding of what people would think and how they would proceed when engaged in this sort of mutual task. My disagreement with them is thus framed by some common assumptions about the aims of morality.

I have chosen to focus on a number of views that all characterize morality as involved with the aim of establishing mutually acceptable social norms both because this is a compelling account of the practical nature and social role of morality and because it gives naturalism a good run for its money. The idea that morality is concerned with interpersonal
justification acknowledges and analyzes our reasoned response to moral disagreement. It allows us to abstract from the possible distortions and biases of actual negotiations and instead to propose conditions under which agreement would ideally be reached. Yet it still appears to describe an empirically identifiable social activity and role for moral norms. This makes it attractive to naturalists. The activity of moral evaluation takes place in a social context within which a claim about justifiable norms could apparently be appraised for its psychological and sociological plausibility. Justifiable norms could it seems be appealed to in order to explain how we would feel and act in certain kinds of situations. If the activity of finding mutually acceptable norms is in this way open to empirical inquiry, the prospects for using the methods of natural and social science to discover the content of morality appears promising. If, on the other hand, this practical and contextual conception of morality cannot be understood naturalistically, the promise of naturalism in ethics more generally may appear to be dim.

This paper argues that naturalistic approaches come up short. They fail to appreciate the normative content of ethics. An adequate account of the nature of our morally legitimate interests and of the norms compatible with due respect for those interests must rely upon substantive moral judgments. These are judgments that in effect take a normative position in ethics. They employ rather than simply describe or reference our evaluative concepts. An empirical reduction of these judgments cannot fully be carried out and thus no empirical reduction of the justification of moral norms can ultimately be given.

II

The first attempt at reduction we will consider identifies the content of moral norms by reference to certain idealized responses of rational agents. One of the most prominent accounts of this kind is found in Peter Railton's interpretation of moral realism. Specifically, Railton defends the claim that moral norms are norms that reflect "what would be rationally approved of were the interests of potentially affected individuals counted equally under circumstances of full and vivid information." Railton maintains that his approach is reductive because the relevant facts about people's responses pick out properties that can be described without using evaluative concepts. It is naturalistic because these properties are natural properties that figure into causal accounts of our behavior given by the natural and social sciences.

Railton arrives at his idealized response criterion of moral rightness by aiming to identify a criterion that would be supported by reference to sociological, physiological, and psychological accounts of our moral dispositions and conduct. For example, when social arrangements fail to satisfy the relevant criterion, this might help to explain why people get angry, organize and protest, or become depressed. A criterion of moral rightness that has this kind of support is one that facilitates "criterial explanation" of our behavior. Because the criterion is reductive and naturalistic, the behavioral explanations that Railton thinks are supported by reference to moral rightness are explanations that will refer only to natural (as opposed to nonnatural) facts. He refers to these natural facts as "moral facts." They might include, for example, psychological facts about what distribution of scarce resources we would be disposed to approve of when we aim for consensus on questions about the distribution of social goods. Like Boyd, Railton believes that reference to moral facts could help to explain our behavior, whether or not we believe those facts ought to guide our behavior, that is, independently of our beliefs about what is morally right. Because the moral facts do not depend on what we believe they are, he thinks we may understand the view he sets forth as a form of moral realism.

Railton launches his analysis of moral norms with a brief account of when the assessment and regulation of behavior should be thought, generally speaking, to take place on moral grounds. Here he appeals to some convergence among moral philosophers on the idea that moral norms concern the assessment of conduct or character where the interests of more than one individual are at stake. In adjudicating the interests of many persons, morality would appear to aim for impartiality and to be equally concerned with all those potentially affected. Railton thinks these are the aims of a form of social rationality and that the substance of moral realism emerges in the analysis it gives of the nature and content of social rationality, for instance, in its treatment of "equal concern." As we have seen, central to the account Railton defends is the idea that the content of moral norms is a matter of what would be approved of by persons who are symmetrically situated and fully informed about the nature of each other's interests.

The strength of this account, according to Railton, and the source of its validity lies in its explanatory potential. He argues that moral facts about what is socially rational could help to explain the dissatisfaction and unrest that may characterize social arrangements that significantly discount the interests of a particular group. It may also be the case that social arrangements that are nonoptimal with respect to social rationality display "a tendency toward certain religious or ideological doctrines, or
toward certain sorts of repressive apparatus; they may be less productive in some ways (for example, by failing to develop certain human resources) and more productive in others (for example, by extracting greater labor from some group at less cost).”¹⁴ Causal connections between these tendencies and failures of social rationality would be empirically significant. Furthermore, such an account may help us to understand how persons acquire the belief that their society is unjust and to explain why their society may come to develop norms that better approximate social rationality.¹⁵ Although Railton hesitates to endorse a general thesis of historical moral progress, he does identify certain historical trends that he thinks are amenable to criterial explanation by reference to moral rightness.¹⁶ These include a growing tendency in moral theory to draw a connection between normative principles and human interests.

The terms of this conception of moral rightness are appealing in several respects. A concern with our idealized responses would seem to acknowledge the importance within morality of giving the interests of all equal consideration and, ultimately, of reaching agreement. It is relevant to abstract from the epistemic limitations of our actual circumstances and to consider instead the ideal circumstances in which rational agreement on the content of moral norms could be worked out. These would seem to be core aspects of moral reasoning. Yet the finer details of the account, designed to enable us to identify people’s idealized responses, are less convincing. This significantly compromises the potential of Railton’s account to illuminate our moral practices.

On Railton’s approach, identifying moral norms will depend on our understanding of what count as people’s interests. To address this need he formulates what is, despite some counterfactual idealization, a reductive, psychological description of people’s interests. He thinks my interests can be understood in terms of what I would want my nonidealized self to want, with a view only to self-interest and not to the interests of other persons. Specifically, my interests are a matter of what I would want my nonidealized self to want were I to consider the matter from a position in which I have unqualified cognitive and imaginative powers and full factual and nomological information about my physical and psychological constitution, capacities, circumstances, history, and so on.¹⁷ More simply put, “a person’s interests are a matter of what he would want himself to seek if he knew what he were doing.”¹⁸

This stance for identifying an agent’s interests is meant to be predictive. It orients us to natural facts relevant to predicting what a person would approve of, given full information, et cetera. Railton’s reference to “unqualified cognitive and imaginative powers” is supposed to be normatively thin; he rejects anything but an instrumental conception of rationality. We may think of his understanding of an agent’s objective interests as fitting within Bernard Williams’s understanding of internal reasons. Williams claims that what a person has reason to do must have the support of his or her present motives or of reasoning that supports those motives.¹⁹

Railton thinks that once we have identified all the interests that members of a given social group have, we can formulate moral norms for that group that set the standard of right action for it. We do this by predicting which norms would be judged to be of instrumental value by persons who count each other’s interests equally under certain counterfactual circumstances: circumstances in which persons have full and vivid information about the consequences for their interests of compliance with the proposed norms. Railton thinks this points us toward norms that would maximize the overall satisfaction of interests, but this is, on his approach, an empirical matter.²⁰

I will focus on three problems with this approach. The first two concern the status of Railton’s definitions. It is not clear that his analysis provides an empirical reduction after all. The reference to “full and vivid information,” for instance, would appear to involve evaluative concepts and appeal to “rational approval”; while normatively thin, this is still nonreductively normative. I will not pursue this criticism in any detail, however, and will simply accept Railton’s view as minimally normative.

Second, even if what Railton offers as the relevant notion of interests is nonreductively normative, it appears to be insufficiently normative. I turn briefly to an account of normative reasons developed by Michael Smith in order to make out this criticism. Smith endorses an analysis that overlaps with Railton’s in important respects. Like Railton, he argues that “what it is desirable for us to do in our actual circumstances is what our more rational selves, looking down on ourselves as we actually are from their more privileged position, would want us to do in our actual circumstances.”²¹ He claims that a fully rational agent is one who suffers no psychological disorders, has no false beliefs, has all relevant true beliefs, and engages in sound means-ends reasoning. Nevertheless, Smith thinks the satisfaction of these conditions does not entail that a person’s preference is fully rational, for they do not capture an important respect in which rational reflection can lead us to revise our current aims. This concerns whether our current aims are “systematically justifiable.”²² Our desires and aims are systematically justified when they fit together in a coherent and unified way. The process of achieving coherence and unity may lead us to give up certain desires and to acquire others. Making
adjustments in our overall set of desires and values for the sake of coherences is the process Rawls has described as attaining “reflective equilibrium.” This is a plausible requirement of moral reasoning, especially in view of the role within morality that principles play. We could supplement Raiton’s account of the relevant understanding of people’s interests by adding to his criteria of rationality the requirement of reflective equilibrium described by Smith.23

Even when it is supplemented by this aspect of Smith’s general account of normative reasons, however, Raiton’s analysis is inadequate. The problem is that Raiton treats the content of persons’ morally relevant interests as fixed prior to moral reasoning. In fact, he identifies them as aspects of a person’s “non-moral good.”24 He does not attempt to characterize moral criteria for determining what to count as legitimate interests, and thus he gives us no sense of why and to what extent we should expect others, if they are reasonable, to accept claims based upon our individual interests. As we have seen, morality must posit a set of fundamental interests as a justifiable basis of claims, but if morality is to have the interpersonal role I have imputed to it, this set of interests cannot include all the interests that persons may affirm.25 There is no reason to think that all the interests that a person would affirm under conditions of full information and in reflective equilibrium would help to comprise an acceptable basis for identifying social norms of right action.26 Raiton concedes that some persons may turn out to have objective interests in things we find harmful or repulsive and that we may well judge these interests to be morally wrong.27 Why, then, should such interests be counted equally along with morally unobjectionable interests when it comes to determining which norms are socially rational? It is not enough for Raiton to argue that objectionable interests will get “overruled”: they simply should not count.28 Other interests, while perhaps not morally repugnant, presuppose ideas about what is good that are the subject of genuine and not unreasonable disagreement. Interests rooted in the endorsement of a particular religious doctrine or ethnic identity, for instance, are not generally shared. If such interests are to ground legitimate moral claims, say, for a distributive claim to a share of social goods or to special rights, they must be reinterpreted in terms of values that persons more generally have reason to care about, such as the value of freedom of conscience and freedom of association or the moral importance of rectifying an injustice. This is because such distributive claims impose costs on other persons and it is objectionable to impose costs on persons in order to advance values they could not reasonably be expected to recognize. Interests that cannot be generalized in the way I have described should drop out of consideration as a reasonable basis for making demands on other people.

Although Raiton’s account of social rationality is constructed in accordance with a requirement of interpersonal agreement, it seems inadequate as an account of the reasoning that makes agreement possible and underwrites its validity. As we have seen, we need an account of the reasoning that could plausibly narrow the class of relevant interests. Without this kind of filtering, we could not expect to reach agreement, at least not for the right reasons. Our morally legitimate interests must have the support of values that all relevant participants to moral deliberation could reasonably be expected to recognize, in full view of the burdens that realization of those values may impose.29 This characterization of morally relevant interests does not attempt to fix their content prior to moral reasoning, but would use moral reasoning to understand them. It will rule out certain interests, for example, state-sponsored religion in a religiously pluralistic society or school prayer in a society that contains nonbelievers.

In terms of the notion of reflective equilibrium to which Smith appeals we might say: morality requires that reflective equilibrium be not only wide but also general.30 Interests that provide a reasonable basis for making moral claims on other people must fit into a moral conception that could be the object of wide and general reflective equilibrium. Wide reflective equilibrium is attained when a moral agent achieves coherence in her moral judgments, not just through making minimal adjustments to fit the closest approximating theory but after more wide-ranging reflection about possible alternatives. Reflective equilibrium is general when different persons affirm the same moral theory, or as Rawls puts it, “the same conception is affirmed in everyone’s considered judgments.”31 He continues, “In such a society not only is there a public point of view from which all citizens can adjudicate their claims, but also this point of view is mutually recognized as affirmed by them all in full [that is, wide and general] reflective equilibrium.”32 What this means is that our understanding of other people’s basic interests must fit into a conception of morality that we may reasonably think could be affirmed from their perspective as well as our own. Raiton has not provided a reductive analysis of these judgments. Indeed, he has not attempted to, since he appears not to acknowledge the importance of establishing that a person’s morally relevant interests should meet the requirements of wide and general reflective equilibrium.

I have considered two problems with Raiton’s account. The first concerned whether Raiton’s account provides an empirical reduction after
all. The second concerned whether his account of people’s morally relevant interests is sufficiently normative. Finally, I turn to a third problem with Railton’s account. This is a problem with Railton’s claim that the interests of all be counted equally. As I have indicated, Railton suggests that equal consideration of the interests of all would lead us to aggregate and to maximize the overall satisfaction of interests. But why is this so? Why should it not instead lead us to affirm a conception of equal rights? Or the value of equal outcomes? Or the importance of equal opportunity for all? The basis for predicting that we would adopt an aggregative and maximizing interpretation of “equal consideration of interests” is not obvious. In any case, we want to know why we ought to accept an aggregative and maximizing approach. Missing from Railton’s description of the standards of moral deliberation is a normative criterion and analysis of reasonable moral judgment.

III

Let us now turn to a prominent attempt by Jürgen Habermas to analyze the requirement of interpersonal justification. His approach interests me because it appears to include an analysis of moral reasoning that aims, among other things, to narrow the class of relevant interests. At the same time, the analysis of moral reasoning he offers is reductive. In particular, Habermas claims that his principles of discourse ethics can be represented procedurally without reference to the normative content of the particular outcomes compatible with these principles.

Habermas’s principles of discourse ethics hold that justifiable moral norms are those that would be affirmed by free and symmetrically situated participants in ideal discourse. Participants in ideal discourse attempt to determine whether all persons affected by a proposed norm could accept the intended consequences and side effects that the general observation of the norm could be expected to have for the interests of each person. This criterion of legitimacy for ideal discourse is expressed by a “Universalization Principle” that Habermas refers to as Principle U. Principle U states, “A norm is valid when the foreseeable consequences and side effects of its general observance for the interests and value-orientations of each individual could be jointly accepted by all concerned without coercion.”

Principle U bears some similarity to Railton’s criterion of social rationality. Recall that Railton proposed that moral norms are norms that reflect what would be rationally approved of were the interests of all counted equally under circumstances of full and vivid information. Both approaches aim for ideal agreement, abstracting from morally distorting features of actual negotiations, such as the influence of ignorance, an imbalance of power, or disproportionate attention to self-interest. In two ways, however, U appears to capture the requirement of interpersonal justification better than Railton’s criterion of social rationality does. First of all, Railton believes that counting the interests of each person equally supports the idea that we ought to aggregate and to maximize the overall satisfaction of interests. By contrast, Habermas affirms the equal status of other persons by having us take up their perspective and determine whether the proposed norms are acceptable in view of their interests. This reciprocal perspective-taking would seem to block Railton’s move toward aggregation. Moral norms must represent what all consider fair to and in the best interests of each person, and this means that the basic interests of each person must be protected; the violation of some persons’ basic interests cannot be outweighed by greater gains to others. This gives good sense to the idea of justification to all.

Second, as Habermas understands Principle U, the content of people’s fundamental interests is not fixed prior to moral discourse but rather is subject to negotiation and agreement. Principle U governs our reasoning about which interests are morally relevant as well as which norms ought to guide us in acting. Relevant interests must be “generalizable”: they must be interests whose legitimacy can be affirmed from all points of view. Interests are relevant only when they meet a requirement of interpersonal justification.

The reductive aspect of Habermas’s account relates to his claim that the moral content of Principle U can be represented procedurally via formal constraints on the ideal discourse situation. These formal constraints are captured by a set of “general symmetry conditions.” Following Robert Alexy, Habermas states these conditions as follows:

1. Every subject with the competence to speak and act is allowed to take part in a discourse.
2. a. Everyone is allowed to question any assertion whatever.
   b. Everyone is allowed to introduce any assertion whatever into the discourse.
   c. Everyone is allowed to express his attitudes, desires, and needs.
3. No speaker may be prevented, by internal or external coercion, from exercising his rights as laid down in 1 and 2.

The reason Habermas thinks Principle U can be represented procedurally by way of these symmetry conditions on discourse is that he thinks these conditions represent the formal or structural features of discursive
practices per se and that \( U \) in fact derives from the formal features of practical discourse.\(^{38} \) Practical discourse is the discursive context within which we evaluate reasons for accepting possible action-guiding norms.\(^{39} \) Drawing upon a general theory of argumentation, Habermas claims that when we advance arguments that aim for agreement on action-guiding norms, we commit ourselves to the idea that an argument can rationally be accepted as valid only when the context of discussion guarantees in principle freedom of access, equal rights to participate, truthfulness on the part of participants, and the absence of coercion in adopting positions.\(^{40} \) These are the presuppositions captured by the general symmetry conditions listed above. Arguments that depend on violating any of these conditions fall short of being rationally convincing, and thus must fail to demonstrate the validity of principles or other claims put forth. Since the above presuppositions are unavoidable and rationally necessary constraints on argumentation, or so Habermas claims, any attempts to jettison them will be inconsistent. Consider the following statement: “By excluding questions that slowed down the discussion, we demonstrated that a proposed ethical norm is rationally defensible.” Such a statement, argues Habermas, involves a “performative contradiction.”\(^{41} \) Argumentation that claims to establish the validity of a norm implies that a mutual agreement on that norm by way of the full discursive participation of rational and relevantly situated persons could be worked out. In this example, however, it is claimed that validity has been established under conditions that preclude the free and full participation of all. This is inconsistent.

Habermas’s argument, then, is that a set of general symmetry conditions, together with a weakly normative understanding of what it means to justify norms, entail Principle \( U. \)\(^{42} \) In other words, if one engages in argumentation with the aim of establishing that certain norms of right action ought to be adopted, the presuppositions of discourse (captured by the general symmetry conditions) entail that one also presupposes Principle \( U. \) Habermas thus takes the position that justifiable norms must have the support of an agreement that would emerge from free and full discussion, the conditions of which appear to be empirically verifiable. This proposal is appealing for its simple, democratic, and naturalistic character.\(^{43} \) Let us now evaluate its plausibility.

In claiming that the principles of discourse ethics can be represented as purely formal constraints on deliberation, Habermas is claiming that the discourse ethical principles do not restrict the content of norms that can be proposed in ideal discourse, only the conditions under which discourse proceeds. This means that whatever rational persons agree to under those conditions would be justified. Habermas does not attempt to eliminate references to rationality, but he believes that its noninstrumental requirements can be represented formally.\(^{44} \) Hence the justification of particular norms or principles, such as principles of justice, can be understood in terms of instrumentally rational agreement under the formal constraints on joint deliberation represented by the general symmetry conditions. Provided that we can establish that the symmetry conditions and the (weak) requirements of instrumental reason are satisfied, determining which norms would be the object of interpersonal agreement becomes an empirical matter.

I do not find this account plausible. The reason is that Principle \( U \) appears to convey substantive moral content not contained in the general symmetry conditions Habermas elaborates. It would appear to be derived on the basis of premises not represented in those conditions and to act as a more substantive guide to deliberation. Principle \( U \) formulates an idea of the aims it is reasonable for participants to ideal discourse to attempt to advance: participants should propose as valid only principles they believe express equal respect for and consideration of the interests and value-orientations of all. Thus, if participants do not conform to this requirement, they violate the conditions of ideal discourse. This narrows the range of considerations participants to ideal discourse could put forth, increasing the likelihood of convergence. In this way, \( U \) articulates substantive normative constraints on proposals that are worth taking seriously. What counts as legitimate reasons within discourse must be compatible with a certain ideal of equality and mutual respect.\(^{45} \) But then it would appear no longer to be true that in ideal discourse, “anyone can advance any claims whatever.” Certain sorts of proposals will be discounted, even though they may generally be compatible with the communicative aim of persuading others on the basis of reasons. Such proposals could be those that fail to draw upon shared values and interests, appealing instead, for instance, to private experience and anecdotal evidence, the value of particular aesthetic or religious conceptions, or the importance to the majority of an established way of life.\(^{46} \)

Habermas might reply that objectionable proposals would be filtered out in practice as they fail to generate agreement. But this may not be true. It seems possible that group consensus about what constitutes a reasonable proposal could be distorted in significant ways even when participants to discourse are not obviously unreasonable. The possibility of collective yet unreasonable agreement could be traced to the ways in which people's life experience and social position may distort their judgment. Some persons have low self-esteem and underestimate the value
of their claims. Others are overconfident and self-important and press their claims too far. These distortions are not psychological disorders but common and sometimes subtle biases. A group as a whole could overestimate the value of austere measures of self-sacrifice, for example, or be overly invested in punishing offenders due to collective guilt for the misdeeds of previous generations. These factors may lead a group to agree on principles that are not entirely fair and responsive to the needs and interests of all persons. The idea that satisfying the general symmetry conditions is incompatible with these biases is not plausible unless those conditions contain normative content that Habermas does not analyze.

The question that moral theory should try to answer is what would guide us to reasonable agreement. Once we understand how the substantive requirements of moral reasoning extend beyond the general symmetry conditions Habermas elaborates, we are in a position to see that those conditions do not provide us with any real guidance in determining which claims we have moral reasons to take seriously. Principle $U$ tells us that the claims people advance in practical discourse need to be assessed for their moral content in order to determine their relevance: we need to assess whether proposals express due consideration for the interests and value-orientations of all and thus could be affirmed in a procedure of universal perspective-taking. The mere fact that somebody has freely advanced a claim in discourse carries no weight. There would even appear to be some tension between promoting Habermas's symmetry conditions and the discriminatory judgment required by $U$.

An advantage of what we might think of as the reductive interpretation of $U$—which more or less equates the content of $U$ with the alleged formal features of discourse elaborated by the general symmetry conditions—is that it makes it easier to envisage what would be involved in the realization of practical discourse procedurally governed by $U$. This could put us in a position to study the activity of moral justification empirically and to identify its results. But read as a substantive moral principle, $U$ cannot be implemented as straightforwardly as the general symmetry conditions could be. The implementation of $U$ requires the subtle and discriminatory judgment required in perspective-taking and the conduct of fair play. The moral sensitivities of those persons capable of making use of $U$ as a regulative principle must already be considerably developed. Participants in discourse must understand the difference between what is respectful and what is inconsiderate or exploitative, for instance, if they are to succeed in implementing $U$. They must maintain a sense of fairness and goodwill toward one another, and be capable of discerning which proposals fall short of demonstrating full commitment to the aims of practical discourse. Also, they must be able to make out clear and cogent arguments that are responsive to the relevant information and be careful in addressing matters that are commonly the subject of disagreement. Discourse ethics must rely irreducibly on reasonable judgments about the moral relevance of various sorts of considerations, judgments that Habermas's procedural reduction does not analyze.

IV

The notion of what is reasonable plays a larger role in moral deliberation than either Habermas or Raiton would seem to admit. It characterizes not only the symmetrical situation of participants to moral discourse and the information on the basis of which their deliberation proceeds, but also their orientation and judgment within deliberation. The fact that an agreement emerges from discourse characterized by Habermas's general symmetry conditions or meeting Raiton's criterion of social rationality is not sufficient to ensure its reasonableness; it must also be supported by the right reasons, say, by reasonable judgments of the legitimacy and relative urgency of the needs and interests of affected persons.

The notion of the reasonable is not itself fundamental. The normative force and content of the reasonable judgments derives from arguments about the relevance to moral deliberation of invoking particular values. Moral principles to govern social life will be justified by reference to the values of fairness, respect, equality, reciprocity, and so on. Various values together help to explicate the requirement of interpersonal justification because they each provide compelling and rationally defensible grounds for preferring some proposed principles to others. For example, as T. M. Scanlon puts it, we have reason to reject principles when they arbitrarily favor the claims of some persons over the identical claims of others: that is to say, because they are unfair. The same may be true about principles that are incompatible with the reasons we have to want to live our lives in ways that realize important values such as friendship.

The point is that in analyzing the justifiability of possible action-guiding norms we must look beyond the psychological and sociological fact that persons take certain reasons to provide compelling and rationally defensible grounds for preferring some moral norms and rejecting others. We must look to why, if at all, these reasons should play this role. The clue to the labyrinth of our moral notions lies in evaluating the normative force of the moral arguments into which they figure. Only doing this will enable us to understand what will have made it possible to reach reasonable agreement on moral norms and to correct distortions in judg-
ments, when agreement has provisionally been attained. It is not just the relevant notion of interests that resists reduction, but a whole family of moral reasons and the arguments upon which they rest for their relevance.

We should reject reductive accounts of how the work of moral reasoning is done. When norms that satisfy naturalistic criteria of moral rightness fail to square with our settled moral convictions, there is reason to doubt that these criteria illuminate the nature of morally defensible norms and practices or that they are adequate to guide us in deliberating about which social norms are justifiable. The requirements of reason in morality cannot be codified by a list of formal constraints on a procedure of reasoning or by a causal account of the content of our moral judgments. In order to illuminate the nature of our moral aims and judgments, a moral philosopher must engage in substantive moral reasoning, however indeterminate and fallible it might be—that is, she must engage as a participant in moral discourse. There is, philosophically, no shortcut around practical reasoning itself. We identify morally justifiable norms on the basis of compelling reasons to think that a proposed position reasonably takes everyone's interests into account. No set of social, psychological, or historical facts can replace the role that good arguments play in the enterprise of identifying justifiable moral norms.54

This does not mean that our moral notions could not figure into empirical theories. Our normative notions may well have explanatory value. Let's hope they do. But unless empirical explanations of social behavior incorporate our reasonable judgments, and not just the nonnormative fact that we have and regard as reasonable such judgments, those explanations will not count as moral explanations, that is, explanations that rest on "moral facts," as relative and relational as those facts may be.