Ratio Juris. Vol. 17 No. 1 March 2004 (27–51)

Moral Objectivity and Reasonable Agreement: Can Realism Be Reconciled with Kantian Constructivism?

CRISTINA LAFONT

Abstract. In this paper I analyze the tension between realism and antirealism at the basis of Kantian constructivism. This tension generates a conflictive account of the source of the validity of social norms. On the one hand, the claim to moral objectivity characteristic of Kantian moral theories makes the validity of norms depend on realist assumptions concerning the existence of shared fundamental interests among all rational human beings. I illustrate this claim through a comparison of the approaches of Rawls, Habermas and Scanlon. On the other hand, however, objections to moral realism motivate many Kantian constructivists to endorse the antirealist claim that reasonable agreement is the source of the validity of social norms. After analyzing the difficulties in the latter strategy, I try to show how a balance between the realist and antirealist elements of Kantian constructivism can be reached by drawing a sharper distinction between the justice and the legitimacy of social norms.

One striking feature of contemporary debates in metaethics is the proliferation of all kinds of moral realisms, antirealisms, quasi-realisms, and an endless variety of combinations of them. Most of these metaethical debates can be traced back to a remarkable feature of our practices of normative assessment, namely, the purported objectivity and unconditional validity of our normative judgments. Moral realists try to explain this feature of our normative judgments by assimilating them to factual judgments. Accordingly, normative judgments are supposed to describe an order of moral facts that subsists entirely independent of human beliefs and attitudes. Moral antirealists try to avoid the implausibility and the problematic consequences of such metaphysical assumptions by embracing different versions of what these days is called *expressivism*, namely, the view that in making moral judgments we do not even purport to make claims about what is objectively right

or wrong, but rather simply to give expression to our non-cognitive attitudes. The desired metaphysical parsimony is thus bought at the price of renouncing the claim to objectivity entailed by our normative judgments and thus embracing a revisionary approach vis-à-vis our current moral practices. In this metaethical context moral Kantianism seems particularly hard to situate. On the one hand, Kantians explicitly oppose moral realism. Thus, they agree with the antirealists that our normative judgments do not purport to describe a pregiven moral order, heteronomously imposed on us independently of our practical reason. Contemporary Kantians emphasize this opposition to realism by characterizing their approaches as *constructivist*. But on the other hand they do not want to renounce the claim to objectivity of our normative judgments as moral antirealists do. They are moral cognitivists and not moral expressivists. However, given that expressivism is the paradigmatic feature of moral antirealism, it seems that Kantian constructivism can be at most an *anomalous antirealism*.

This anomaly makes Kantian constructivism an essentially unstable position.[1] It seems only possible to fully develop it into either a consistently realist or a consistently antirealist approach. One can follow a consistent antirealist strategy and claim that moral rightness is exclusively a function of our beliefs and attitudes. As I will try to show in what follows, this strategy is incompatible with moral cognitivism and leads inevitably to a decisionist approach. Given that cognitivism is an essential feature of Kantian moral theory, this relativist strategy would lead to a theory that would no longer be recognizably Kantian. Or one can stick to the claim of objectivity and recognize that the moral rightness of norms is not a function exclusively of our beliefs and attitudes. But in so doing one has already conceded everything that is required by a realist strategy. In this case what remains to be shown is that the realist presuppositions implicit in this strategy do not amount to the assumption of a moral order of facts that subsist independently of our moral practices.

In order to show this, I would like to first identify the core of realist assumptions built into Kantian moral theories through a short comparison of the approaches of some of its most important defenders (I). This will make possible to see what justifies the claim that our normative judgments can be objectively valid within the framework of moral Kantianism. In a second step, I will then address some of the standard objections to moral realism that motivate many Kantian constructivists to provide a decidedly antirealist account of the validity of social norms (II). In this context, I will try to show not only that the antirealist strategy fails to avoid the objections, but also that what is required to avoid them successfully is to find the right

[1] Many critics of Rawls's *Kantian constructivism* (Rawls 1999a) have pointed out the instability of this position, which aims to be neither realist nor relativist. See O'Neill 1989, 206–18; also Brink 1989, 303–22. The same point is made with regard to Kantian theories in general by Darwall, Gibbard, Railton 1997, 12.

balance between the realist and the antirealist elements inherent in Kantian constructivism (III).

I. Realist Assumptions at the Basis of Kantian Constructivism

Kantian moral theories belong to the tradition of social contractualism broadly conceived.[2] The distinctive feature of this tradition is the attempt to explain the validity of social norms in terms of the notion of a possible agreement among those to whom such norms apply. Following this idea, all Kantian moral theories offer some moral principle or procedure to discover valid norms. What distinguishes Kantian contractualism from other contractualist approaches is the claim that such a procedure makes it possible to single out norms valid *for everyone*. Accordingly, the objectivity of our moral judgments is a function of the universal validity of the norms that such a principle or procedure (like Kant's categorical imperative, Rawls's original position, Habermas's principle of universalization, Scanlon's moral principle, etc.) purportedly allows us to select. Given that the results of applying the procedure are assumed to be objectively valid, such an approach must be able to explain in virtue of what this assumption of validity can be granted. It seems clear that if the procedure can single out norms that are equally valid for all of us, regardless of who happens to employ it, something about us must be shared and fixed as well. Only under the assumption of some kind[3] of homogeneity among the interests and needs of all possibly affected by a norm does it make sense to claim that a procedure sensitive to such homogeneity would be able to yield single (i.e., universally valid) outcomes. The claim of objectivity in Kantian approaches stands or falls with this assumption. This can perhaps be seen best, if we compare them with those approaches in the contractualist tradition that are built on noncognitivist premises (such as those developed from Hobbes to Gauthier).

All social contract theories share the assumption (1) that questions of justice arise when there is a conflict of interest between different people, and the claim (2) that a rational answer to questions of justice is one that all possibly affected could reach a rational agreement[4] on. This claim is the

[2] From a purely historical point of view, social contractualism is usually traced back to the view introduced in Plato's *Republic* by Glaucon and its main historical representatives are considered to be authors such as Hobbes, Locke, Rousseau, Hume, and Kant. However, from a systematic point of view, the social contract theories developed by these authors are surely too heterogeneous to be considered as part of a single tradition. For a detailed account of the mutually incompatible conceptions of justice at the basis of social contractualism (justice as impartiality vs. justice as mutual advantage) see Barry 1989.

[3] Of course, as we will see, not just any homogeneity will do. It must be of a morally relevant sort.

[4] Although Kant's procedure (i.e., the categorical imperative) does not make a direct reference to the notion of agreement, one important sense of this notion is operative in his approach, namely, the idea of rational consent implicit in his conception of autonomy, according to which our moral autonomy depends on following a law that our reason has given to itself, i.e. a law

normative core of the otherwise metaphorical idea of a social contract. Of course, the cogency of the contractualist idea of rational agreement turns on two further assumptions, namely, (3) that all parties to the agreement share an interest in solving their conflict by rational means, and (4) that making the resolution dependent on their rational agreement guarantees that the interests of all will be taken into consideration. It is by virtue of the last assumption that any specific version of contractualism can plausibly claim to provide an answer to the question of what *justice* requires, that is, to draw a normative line between just and unjust resolutions to social conflicts.

However, these minimal assumptions are obviously insufficient to distinguish between cognitivist and noncognitivist versions of contractualism. For although both versions consider rational agreement to be a condition for justice, there is nothing in the assumptions mentioned so far that would support the presumption that all such agreements would have to have identical outcomes in order to be just.[5] In fact, the opposite conclusion seems more plausible. For if one assumes that beyond the shared interest in a rational resolution of their conflict all other interests of the affected parties differ or, even worse, are essentially in opposition—as assumption (1) may suggest—, the outcome of each agreement would essentially depend on what happen to be the interests of those affected in each case as well as on their relative willingness to compromise some of them for the sake of reaching agreement. No matter how strict the conditions for the fairness of the procedure were to be set up, the essential differences in the makeup of the participants would necessarily be reflected in different outcomes of their agreements.

Thus, the claim of objectivity entailed in the Kantian versions of contractualism seems to depend on assuming that the interest in the rational resolution of conflict is not the only interest that all affected parties have in common. It is further assumed (5) that they share those basic interests and needs that are necessary to sustain their lives as rational beings.[6] And it is

that we could rationally agree to follow. It is the other sense of the notion, namely, the intersubjectivist sense of an agreement *with others* that is not emphasized in the categorical imperative (although it trivially follows from it: Given the assumption of universal validity, it is taken for granted that in following the categorical imperative my rational agreement would coincide with the agreement of all other rational beings).

[5] In fact, the point of emphasizing rational agreement among the participants to the contract as a condition for justice would seem to be lost, if that presumption is correct. For if questions of justice have fixed right answers and thus the justice of an agreement is a function of the correctness of its outcome, then the sense in which reaching an agreement can nonetheless be a condition for justice is far from clear. I will address this important and difficult issue later.

[6] Assumption (5) may seem incompatible with assumption (1). That this is not the case, though, was forcefully argued by Rousseau in *The Social Contract* with the following remarks: "If the establishment of societies was made necessary because individual interests were in opposition, it was made possible because those interests concur. *The social bond is formed by what these interests have in common; if there were no point at which every interest met, no society could exist*. And it is solely on the basis of this common interest that society must be governed" (Rousseau 1994, 63, my italics).

in virtue of the homogeneity of their basic interests as human beings that the outcomes of their possible agreements can be expected to be homogeneous as well: Norms that protect those interests for all human beings are just, whereas those incompatible with such protection are unjust. Only under this further assumption does the basic claim of contractualism acquire the egalitarian sense characteristic of its Kantian versions. In a Kantian framework the claim that rational agreement among all affected parties guarantees that the interests of all will be taken into consideration does not mean merely that all conflicting interests will be balanced against each other in order to reach a feasible compromise, like in a Hobbesian framework. It means specifically that of all the interests that the different parties may have, those that they cannot fail to share because they are necessary to sustain their lives as rational beings will be *equally protected* by the norms agreed upon.[7] It is the assumption of universally shared interests and needs that in turn gives *prima facie* plausibility to the claim that questions of justice can be answered by a procedure that will yield single answers. The claim of objectivity of Kantian contractualism turns on this assumption, which constitutes the *differentia specifica* with the noncognitivist versions.

This can be seen more clearly if one translates the opposition between the cognitivist and the noncognitivist versions of contractualism into the contrast between realism and antirealism that we mentioned at the beginning. For in this context the question of what it is that the moral Kantian assumes exists and what the moral antirealist assumes does not seems pretty clear. According to the characterization offered above, the justice of a norm depends on whether the norm protects those interests that are generalizable among all rational human beings. If it does, the norm is just; if it does not, it is unjust. Accordingly, a moral antirealist (or relativist) genuinely disagrees with this assumption by claiming that there is no such thing as generalizable interests shared by all rational human beings. For what it is rational for human beings to will essentially depends on the actual desires they have to begin with and those, far from being shared, are actually in opposition. Thus, whereas the Kantian cognitivist is committed to the existence of an overlap among those interests that are unrenounceable for any rational human being, the noncognitivist or antirealist is committed to the non-existence of such overlap (i.e., the claim that the intersection yields an empty set, so to speak).[8] This in turn explains why the noncognitivist can

[7] Here it is important to notice that in order to get this result it is not sufficient to replace the assumption that the parties are moved by self-interest with the assumption that they are moved by the moral interest in reaching an agreement equally good for all. For no matter how genuine this interest were supposed to be, if we did not assume that their basic interests and needs actually overlap, there would be literally nothing equally good for all.

[8] Seen in this light, it should be clear that the Kantian moral realist is not postulating any queer ontology in Mackie's sense, for the entities at issue, namely, the various interests that human beings have, are trivially recognized as existing by both sides. It is the possibility of establishing morally significant distinctions among those interests that divides cognitivists from noncognitivists. On Mackie's argument from queerness against moral realism see Mackie 1977, 38ff.

only take an empirical stand vis-à-vis whatever interests and preferences human beings happen to have (for they are intrinsically arbitrary, according to this view), whereas the Kantian cognitivist can take a normative stand vis-à-vis them and distinguish those that are generalizable and thus legitimate from those that are not.

Of course, different authors in the Kantian tradition offer different accounts of the assumption of common interests and needs shared by all rational human beings. In his *Lectures on the History of Moral Philosophy*, Rawls discusses explicitly the assumption of homogeneity in Kant's moral philosophy and shows its crucial role for defending the claim of moral objectivity. Rawls argues that in order to explain how the categorical imperative can have objective content, that is, how it can specify precepts that are roughly the same for all rational agents, it seems necessary to appeal to what Kant in the *Metaphysics of Morals* calls "true [human] needs": "I understand Kant to say that we have certain true human needs, certain requisite conditions, the fulfillment of which is necessary if human beings are to enjoy their lives" (Rawls 2000, 174). Only under such a presupposition does it make sense to think that what human beings can rationally will is (roughly) the same for everyone. As Rawls expresses it, the contradiction in the will test of the categorical imperative presupposes "that we have such needs and that they are more or less the same for everyone" (ibid. 174; see also Rawls 1999b, 501ff.).[9]

An equivalent assumption is built into the structural features of Rawls's own procedural interpretation of Kant's categorical imperative, namely, the original position. As is well-known, the veil of ignorance is tailored in such a way that it allows the parties to have enough knowledge about the general interests that anyone would have as a rational human being, whereas it rules out knowledge of all particular interests that specific human beings have as a result of their specific biographical circumstances, conceptions of the good, etc. Due to the specific features of Rawls's general approach, he not only assumes the existence of an overlap of unrenounceable interests among all human beings. Moreover, he provides an indexing of some[10] of them in the form of a uniform list of "primary goods", that is, of those things that every

[9] In both contexts, Rawls emphasizes that developing this line of thought requires making sure that the essential elements of Kant's doctrine are not compromised. Although he does not say so explicitly, his warning seems to concern the possible conflict between realism and the central role of the notion of autonomy in Kant's moral theory. I address this issue in the second part of this paper and try to show that there is in fact no such conflict.

[10] Rawls's list of "primary goods" arises out of the specific needs of his theory and does not aim to be exhaustive ('natural' primary goods such as health are explicitly set aside and only "social" primary goods such as liberty, wealth, and the bases of self-respect are included). Here I leave aside all issues concerning the appropriateness and consequences of interpreting human interests in terms of "goods", for they are not directly relevant for my present argument.

rational human being wants whatever else she wants (Rawls 1971, 62, 92, 260).[11]

In the case of Habermas's discourse ethics, the existence of an overlap of generalizable interests among all rational human beings is implicitly assumed[12] in his own version of the categorical imperative, namely, his principle of universalization. According to this principle, "only those action norms are valid to which all possibly affected persons can accept the consequences and the side effects its general observance can be anticipated to have for the satisfaction of everyone's interests" (see Habermas 1990, 65). What is distinctive about this moral principle is that it directly precludes any attempt to single out those interests that can be satisfied equally for everyone outside of the context of participation in real discourses about the justice of norms with all those possible affected. Thus, in this approach the specification of "generalizable interests" is thought of as the *result* of moral discourses and not as something ascertainable prior or independently of participation in moral discourses (Habermas 1990, 2003). However, it seems obvious that they could hardly be the result of moral discourses, if they did not exist at all.[13]

Scanlon's version of the categorical imperative, according to which "the rightness of an action is determined by whether it would be allowed by principles that no one would reasonably reject" (Scanlon 1998, 5),[14] does not make explicit reference to any assumption about the basic interests of those looking for principles of justice beyond their interest in a reasonable agreement.[15] But, as Scanlon makes clear in *What We Owe to Each Other*, the principle's application "cannot be based on the particular aims, preferences, and

[11] See also Rawls 1995, 178, where he remarks that both Habermas' and his own approach "limit relevant human interests to fundamental interests of certain kinds or to primary goods."

[12] Although the term "generalizable interests" is not explicitly used in the formulation of the principle of universalization, it is a notion that Habermas employs throughout his writings on ethics. The most systematic use of it, though, goes back to his *Legitimation Crisis*, where the normative task of a critical theory of society is interpreted as oriented towards the identification of suppressed generalizable interests. See Habermas 1975, chap. 3.

[13] In his latest writings, though, Habermas explicitly opposes the realist strategy that I am proposing here and offers a decidedly antirealist interpretation of discourse ethics. For a more detailed account of the realist interpretation of discourse ethics see Lafont 1998, 1999, 2002, and 2003a. Habermas's objections to this interpretation can be found in Habermas 1998a, 381; and 1999, 271–318.

[14] This is the short version of the principle, that he often uses for brevity. The long version reads as follows: "[T]he rightness of an action is determined by whether it would be allowed by principles that could not reasonably be rejected, by people who were moved to find principles for the general regulation of behavior that others, similarly motivated, could not reasonably reject" (Scanlon 1998, 4).

[15] As in all cognitivist versions of contractualism, the parties's interest in agreement is not interpreted by Scanlon in terms of instrumental rationality as a means to promote their own self-interest, but as a genuine interest in what he calls "justifiability to others." In order to mark this difference with the non-cognitivist versions of contractualism Scanlon speaks of "reasonable" rather than "rational" agreement.

other characteristics of specific individuals. We must rely instead on commonly available information about *what people have reason to want*" (ibid. 204). This information about the "important interests" (ibid.)[16] that any human being would have in a given situation translates into information about "generic reasons" that everyone would have to reject a principle.

As this very short summary already shows, the specific accounts that these authors provide of the assumption of an overlap of basic interests and needs among all rational human beings reflect substantive differences in their overall approaches. How important these differences are in our context, though, depends very much on the exact interpretation of the assumption.

On the weakest interpretation of the assumption's significance, differences in the respective accounts can be seen as merely terminological. For on this reading the assumption entails only the claim that there is an overlap of such interests among all rational human beings, but no further claims about what they may actually be. The minimal claim is thus that questions of justice make sense only under the assumption that there is such an overlap. In this sense, this claim can be seen as part of a conceptual argument. If we came to the conclusion that there are no generalizable interests among all human beings, it would no longer be meaningful to ask whether a norm is not merely good for some people and bad for others, but just or unjust for anyone. As a consequence, the unconditional meaning attached to our current use of the notion of justice would be necessarily lost. To claim that a norm is unjust would be tantamount to claiming that it is not good for some of us. And this, of course, would no longer be the kind of overriding claim that per se invalidates the rightness of a norm, as our current use of the term "unjust" implies. But as long as we can reasonably presuppose that there is an overlap of basic interests among all human beings our judgments about the justice of norms *can already be objectively valid*: If a norm protects those interests for everyone it is just, if it does not, it is unjust.

An altogether different question is whether (and if so, how) we can know which one of the two cases obtains for any specific norm. Designing a procedure to answer this further question may in some cases require a stronger reading of the assumption. In fact, some of the approaches in the Kantian tradition offer substantive characterizations of what those basic interests are. An in-depth analysis of them would most likely show that these character-

[16] In his *Preference and Urgency* Scanlon appeals to the notion of "important interests" in order to make plausible the distinction between subjective and objective criteria of well-being. Whereas on the former criteria a person's subjective preferences and interests "constitute the ultimate standard for judgments about his well-being" (Scanlon 1975, 657), the latter criteria aim at "an objective evaluation of the importance of these interests, and not merely the strength of the subjective preferences they represent" (ibid., 658). On the basis of this distinction, he claims that "the criteria of well-being that we actually employ in making moral judgments are objective." (ibid.) In his *Contractualism and Utilitarianism* he appeals to the notion of "morally legitimate interests" in order to draw a similar distinction (see Scanlon 1982, 119).

izations translate in different metaphysical conceptions of the person, of rational agency, etc. In other cases, most notably in the case of Habermas's approach, a strong reading of the assumption is explicitly undercut by some of the theory's substantive claims (in particular by the claim that only the affected themselves and not the moral philosopher can legitimately determine the substantive content of the assumption). These significant differences with regard to the stronger assumption may even explain the variety of procedures that have been proposed in the tradition of Kantian constructivism. But, no matter whether any of these procedures actually succeeds at its task, it seems clear that none of them would be even intelligible without the minimal assumption of homogeneity.

If we were to seriously doubt the existence of an overlap among the basic interests of all human beings, to follow any of the proposed principles to select our actions would not be just problematic (as it may well be) but entirely arbitrary. With regard to the categorical imperative, to take ourselves as indicators of what the interests and needs of other human beings may be would be *per hypothesis* entirely unwarranted. With regard to Habermas's principle of universalization, to assume that all the affected would agree on the same norms, despite their essentially different interests, would be just absurd. Equally so would it be to assume with Scanlon that their reasons for rejection of principles would coincide. Under these conditions, the specific features of Rawls's design of the original position would be literally incomprehensible for the same reason. Without the assumption of homogeneity among the interests of those possibly affected by a norm there would be literally nothing that would license the claim that some norms are not merely good for some people and bad for others, but just or unjust for anyone. Thus, under such conditions all these putative principles of justice, which as such are designed to detect the difference between the former and the later cases, would be equally condemned to fail.

Now the interesting question is what the implications of the realist assumption of an overlap of generalizable interests among all rational human beings are for the standard claim that Kantian constructivism opposes moral realism.

II. Antirealist Motivations in Kantian Constructivism

II.1. The Worry about Heteronomy

The standard reason that Kantian constructivists adduce against any kind of moral realism is always the concern that any concession to realism unavoidably involves introducing heteronomous considerations about what human beings happen to want or desire which are incompatible with the crucial role that the notion of autonomy plays in Kantian moral theories. However, this concern seems entirely out of place in our context. For given

the specific features of the assumption at issue here, it seems clear that the idea of autonomy is by no means jeopardized by it.

The core of this idea is precisely that our autonomy, to put it in Kant's terms, is a function of the ability to follow our reason rather than our inclinations. That is, it requires the capacity to follow the categorical imperative in order to select from our actual interests and preferences those that are universalizable and only act according to them. Obviously, this could hardly be done (or at least could not be done correctly), if there were no such thing as universalizable interests. But are then those interests just pregiven moral facts, part of the furniture of the universe and as such something heteronomously imposed on our moral practices from the outside? Here the answer very much depends on the exact sense of the question. To the extent that we believe that these interests are those that rational human beings cannot fail to have, we surely must believe that they in fact exist. And this just means that they do so independently of our moral practices. But the issue here can hardly be that these interests should not exist independently of our moral practices (why shouldn't they?). It is just that from a perspective external to our moral practices all other interests and preferences exist as much as these do.[17] Outside of our practices of moral assessment, all human interests and preferences are born equal, so to speak. They either exist or they do not. Only from the normative perspective of asking the moral question about which human interests should be protected or overridden in our social world is it possible to establish a distinction, say, between the interest in killing and the interest in not being killed, whereas from the merely factual perspective of asking the question of which human interests actually exist in our social world no such distinction is possible, for both surely do (if they did not, there would be no conflict and thus no need for a moral regulation of it). Outside the normative horizon of our moral practices, nothing would distinguish them in their moral significance. For moral significance is surely a function of our moral practices. This is the clear sense in which moral facts are not independent of our practices of moral assessment. Outside of these practices there are no moral facts, not because the morally significant facts mysteriously disappear or no longer obtain,[18]

[17] As we saw before, this is precisely the perspective that non-cognitivist versions of contractualism take in considering all human preferences to be intrinsically arbitrary. It is *this* assumption that is incompatible with the Kantian notion of autonomy.

[18] This by no means denies that facts about human interests and preferences can change or even cease to obtain. But if they did, this would be so as much outside as inside our moral practices. As mentioned at the beginning, our moral practices originate in situations of social conflict and thus are essentially dependent on what is usually called "the circumstances of justice." If some of these circumstances were no longer to obtain, many facts about human interests and preferences would surely cease to obtain. But this would remain so even if we were to ask the moral question *of those circumstances*. (The basic human interests and needs that any rational human being would have in our world are surely different from those that they would have in a Robinson Crusoe kind of world, for example. But if we ask the moral question with regard to the latter world, the facts about those interests in that world would surely remain the same before and after we asked the question.)

but because they are indistinguishable *as moral facts* from all other facts.[19]

This is why the realist commitment at the basis of moral Kantianism does not amount to a standard moral realism, either of the non-naturalistic kind defended by rational intuitionists (such as Sidgwick or G. E. Moore) or of the contemporary variety defended by naturalists (such as D. Brink or R. Boyd). By assimilating normative judgments to factual judgments about a subsisting moral order, these varieties of moral realism are committed to the counterintuitive claim that moral facts could in principle be apprehended *as moral facts* from the perspective of an observer entirely detached from the normative presuppositions built in our moral practices.[20] But the strangeness of that kind of moral realism should not lead moral Kantians to embrace antirealism. If the line of argument developed so far is correct, the realist core of moral Kantianism is indeed incompatible with moral antirealism (or noncognitivism), but it is entirely compatible with recognizing our moral practices as a product of our normative constructions. Thus, in light of its ability to account for both the realist and the constructivist elements of morality, moral Kantianism should not be interpreted as an anomalous branch of antirealism but rather as the only plausible kind of moral realism.

But there is another aspect of the Kantian notion of autonomy that is usually thought to be incompatible with any kind of realism. In order to be autonomous, it is not enough that I obey reason in general. I must obey *my* reason in particular. It is the internal connection between autonomy and *free consent* that seems to be lost if, once our moral practices are in place, moral

[19] The best known and most quoted characterization of Kantian constructivism is surely Rawls' statement in *Kantian Constructivism in Moral Theory* that "apart from the procedure of constructing the principles of justice, there are no moral facts" (Rawls 1999a, 307). But his explanation of what this statement means points actually in the same direction that I am defending here. He explains: "Whether certain facts are to be recognized as reasons of right and justice, or how much they are to count, can be ascertained only from within the constructive procedure" (ibid.). He made this position even clearer later in his *Themes in Kant's Moral Philosophy* (Rawls 1999b), where he explains: "To prevent misunderstanding, I should add that Kant's constructivism does not say that moral facts, much less all facts, are constructed. Rather a constructivist procedure provides principles and precepts that specify *which* facts about persons, institutions, and actions, and the world generally, are relevant in moral deliberation. Those norms specify which facts are to *count* as reasons. We should not say that the moral facts are constructed, since the idea of constructing facts seems odd and may be incoherent; by contrast, the idea of a constructivist procedure generating principles and precepts singling out the facts to count as reasons seems quite clear" (Rawls 1999b, 516). In *Political Liberalism* he restates this view and offers a much clearer version of his original statement about constructivism, namely, that "apart from a reasonable moral or political conception, facts are simply facts" (Rawls 1993, 122).

[20] Of course, the naturalistic and the non-naturalistic varieties of moral realism differ widely with regard to the nature of moral facts and the kind of "observation" they require. For naturalists moral facts are naturalistic features of the world and thus are susceptible of regular scientific observation, whereas non-naturalist must postulate some entirely mysterious capacity to detect non-natural properties. What matters in our context, though, is the shared assumption that we could discover moral facts as moral facts just in our capacity as *knowers* (even scientific knowers).

facts are nonetheless imposed on us from the outside regardless of our (possible) acceptance of them. The fundamental moral significance of the notion of voluntary agreement seems threatened by any concessions to realism.

This concern seems to be what leads many Kantian constructivists to defend the decidedly antirealist view that rational agreement is what *constitutes* moral rightness (see Habermas 2003, 297–8; Scanlon 1982, 110, 119; 1998, 1–5; Barry 1989, 268–292; Milo 1995, 184–5, 190).[21] Contrary to the realist view of agreement as an indicator (perhaps even the most reliable indicator) of an independently constituted moral rightness, agreement is seen by these authors as the central moral phenomenon behind our notion of moral rightness. Scanlon explains this idea in *What We Owe to Each Other* with the following remark: "When I ask myself what reason the fact that an action would be wrong provides me with not to do it, my answer is that such an action would be one that I could not justify to others on grounds I could expect them to accept" (Scanlon 1998, 4). Thus, it is the idea of "justifiability to others" that "accounts for the distinctive normative force of moral wrongness."[22] In his latest writings on discourse ethics, Habermas advocates a similar reading of his principle of universalization. In *Truth and Justification*, he remarks that "an agreement about norms or actions that has been attained discursively under ideal conditions carries more than merely authorizing force; it *warrants* the rightness of moral judgments. Ideally warranted acceptability *is* what we mean by moral validity" (Habermas 1999, 297–8).

These accounts seem to be motivated by two correlative aspects of the Kantian notion of autonomy, namely, that to force anyone to act against his own reason is morally wrong and thus that the moral rightness of norms cannot lie beyond the possible reasonable agreement of those to whom these norms apply. There is no moral rightness beyond human rational acceptability. Accordingly, what makes an antirealist strategy *prima facie* more attractive than any realist alternative would be its ability to account for the central moral significance of the notion of mutual agreement and voluntary consent. Unfortunately, as I will try to show in what follows, the antirealist interpretation of Kantian constructivism is actually unable to provide such

[21] In Rawls's case, it is hard to assess whether he would subscribe to this antirealist claim or not. At least since *Justice as Fairness: Political not Metaphysical* (Rawls 1999c) most of Rawls's statements about his Kantian constructivism indicate rather the explicit aim to drop out of the metaethical game entirely. To the extent that endorsing a specific metaethical position would unavoidably require to endorse some comprehensive philosophical doctrine or another, it seems that Rawls' constructivism would have to differ from other versions of Kantian constructivism precisely in declining to endorse any specific metaethical view as the single right interpretation of justice as fairness.

[22] Most of Scanlon's remarks suggest that he intends to defend this claim in its strongest possible sense (e.g., see Scanlon 1998, 4–5; and 1982, 110, 119), but his answers to direct objections against this claim in *What We Owe to Each Other* are so patently evasive that it is actually hard to tell how strong his most considered position should be taken to be (Scanlon 1998, 391, note 20, 393 note 1).

an account. By following the antirealist strategy what is morally significant in the notions of "justifiability to others" and "voluntary consent" gets unavoidably lost, or so I shall argue.

II.2. The Moral Significance of Consent

The difficulty in following an antirealist strategy to account for the moral significance of consent within a Kantian framework is not exactly that its results would be *per se* incoherent or totally indefensible. The problems are actually due to the internal constraints that the acceptance of moral cognitivism and the claim of objectivity impose on the possible ways to follow the antirealist strategy. This can be seen best if we pay attention to how noncognitivist approaches within the tradition of contractualism account for the moral significance of voluntary agreement. Within a noncognitivist framework, an account of the significance of agreement is pretty straightforward. Under the assumption that the interests of the participants to the agreement are essentially in opposition, there is no reason to believe that the moral resolution of their conflict has a right and a wrong answer. There is no right answer to be known, but at most a fair decision to be made. Thus, the rightness of the decision can only depend on whether all participants to the agreement had a fair chance to make their own interests prevail. The moral rightness of their decision is a function of the fairness of the procedure that brought the agreement of the participants about. This provides a clear sense to the claim that agreement constitutes moral rightness: Those norms the participants agree upon under fair conditions, whatever they might be, deserve to be called morally right due precisely to the fact that they were agreed upon in this way. So understood, moral rightness is a purely procedural notion in Rawls's sense.[23]

Following this antirealist strategy, though, makes it impossible to defend the objective and unconditional validity of our moral claims. In this context, it does not make sense to claim that the question of whether a norm is just has an objectively right answer, for it cannot have *any* answer prior to or independently of carrying out the procedure in which a *factual* agreement among the participants is reached.[24] For this reason, a purely procedural

[23] In his *Theory of Justice*, Rawls (1971) characterizes the notion of pure procedural justice in the following terms: "Pure procedural justice obtains when there is no independent criterion for the right result: instead there is a correct or fair procedure such that the outcome is likewise correct or fair, whatever it is, provided that the procedure has been properly followed. This situation is illustrated by gambling. If a number of persons engage in a series of fair bets, the distribution of cash after the last bet is fair, or at least not unfair, whatever this distribution is [. . .] A distinctive feature of pure procedural justice is that the procedure for determining the just result must actually be carried out; for in these cases there is no independent criterion by reference to which a definite outcome can be known to be just [. . .] A fair procedure translates its fairness to the outcome only when it is actually carried out" (Rawls 1971, 86).

[24] See prior footnote.

view of justice leads unavoidably to relativism: norms are not unconditionally valid, but valid only relative to the factual agreements of a specific community at a specific time.

Therefore, in order to defend the claim of objective validity characteristic of Kantian constructivism, moral rightness cannot just be constituted by any and all factual agreements that different communities could reach under more or less fair conditions. At the very least, it must be constituted by an agreement that could be accepted by *everyone*. This constraint leads Kantian constructivists to introduce the distinction between factual and hypothetical agreement in order to avoid a merely decisionistic reading of the claim that agreement constitutes moral rightness. Whereas the straightforwardly antirealist reading of this claim involves embracing pure proceduralism, so that moral rightness is explained in terms of a "factual agreement under fair conditions," the reading that Kantian constructivists favor should be understood as a kind of hypothetical proceduralism (see Darwall, Gibbard, and Railton 1997, 13) that explains moral rightness in terms of a "reasonable agreement under ideal conditions." This strategy of adding further constraints to the conditions of a possible factual agreement in order to avoid the relativist consequence of multiple outcomes can be followed in different ways.

In the case of Rawls' procedure (i.e., the original position), the additional constraints that make it plausible to expect all participants to agree on the same outcome are, on the one hand, the specific features of the situation of agreement (the veil of ignorance), which make the participants to the agreement virtually indistinguishable and, on the other, the single (morally neutral) standard of rationality they are all supposed to apply. Regardless of whether it is indeed plausible to expect a single outcome under these conditions, what seems clear is that the notion of agreement carries no independent weight in determining the outcome of the procedure.[25] The more reasons there are to expect that the parties in the original position reach a specific, single outcome, the less plausible the assumption that the agreement of distinct individuals matters to it seems. Thus, the real theoretical work of explaining what constitutes moral rightness seems to be done by the reasons themselves and not by the agreement (for a clear statement of this problem see Sayre-McCord 2000, 257). To the extent that the notion of agreement or consent plays at most a heuristic role in the theory, its moral significance is clearly not accounted for at all.

This difficulty, though, may seem to depend on the artificiality of the original position. If so, those versions of Kantian constructivism that do not appeal to the hypothetical agreement of hypothetical individuals, but to the

[25] For this reason many interpreters claim that Rawls' problem of reaching mutual agreement in the original position collapses into a decision-theoretic problem of individual choice under uncertainty (e.g., see Barry 1989, 74).

possible agreement of actual individuals, may be better equipped to account for the moral significance of agreement and consent. In the approaches of Habermas, Scanlon, Barry, etc., the conditions for agreement are ideal or hypothetical only in the sense that the participants are supposed to meet some standard of reasonableness. They offer different accounts of what such a standard must be like, but whatever its specific features, the standard is not supposed to be in principle beyond the reach of actual individuals. In the most general terms, the standard of reasonableness involves two general components: a genuinely cognitive motivation (something like the capacity of "following the unforced force of the better argument,"[26] to put it in Habermas's terms) and a genuinely moral motivation (the capacity of adopting an impartial point of view in giving equal consideration to the interests of all). These conditions may be hard to achieve and even harder to assess, if they were to obtain at all. In this sense they are properly called "ideal," but they are certainly not supposed to be "ideal" in the sense of being in principle impossible to meet by actual human beings.[27]

However, there seems to be nothing in the notion of reasonableness alone that can motivate the assumption of single outcomes that these authors build into their respective moral principles. According to these principles, moral rightness requires an agreement that could be accepted by everyone (or not rejected by anyone) under ideal conditions of reasonableness. Thus, in principle no factual agreement short of universal consensus meets the conditions for moral rightness. But if the only resource available in these approaches to motivate the assumption of universal validity is the notion of reasonableness, lack of universal consensus can only mean lack of reasonableness. It is surely uncontroversial to claim that if participants in moral agreements are unreasonable their factual agreement will fall short of universal consensus. But it is very controversial to claim that participants in moral agreements that fall short of universal consensus are thereby necessarily being unreasonable. In view of the multiplicity of hard cases in moral discussion (with regard to norms concerning abortion, euthanasia, animal rights, pornography, etc.), it seems totally implausible to claim that the lack of universal consensus in all such cases is necessarily due to the participants' unreasonableness. As Rawls has forcefully argued, moral disagreement among reasonable people is rather likely to be a permanent condition in pluralistic societies.

Be that as it may, what matters in our context are the implications of this kind of approach for the moral significance of consent. As we mentioned before, one of the motivations behind these approaches seems to be the view

[26] This in turn requires an argumentation process that excludes coercion, deception, bargaining power, etc. See Habermas 1990; Scanlon 1982; Barry 1989.

[27] This claim is explicit in the case of Habermas (1993, 139). It is less clear in the case of Scanlon because of his oscillation between the operational and the achievement sense of the notion of "reasonableness" (see footnote 34). I discuss this issue later.

that social norms can only be valid to the extent that those to whom these norms apply have voluntarily agreed to submit to them. This is what generates the concern for mutual justifiability that is made explicit in their respective moral principles. However, given that these principles link moral validity directly to universal acceptance (or lack of rejection), there is actually no way to account for the significance of consent or dissent under conditions short of total unanimity. At best, these principles are silent with regard to the validity of factual agreements under such conditions. At worst, they seem unable to avoid the suggestion that under those conditions any factual consent or dissent could be overridden in the name of other (allegedly more reasonable) hypothetical participants, whose consent or dissent would actually be correct (on this point see O'Neill 1989, 109). One way or another, these principles seem unable to motivate the concern for mutual justifiability precisely under those conditions in which it matters the most, namely, when there is no universal consensus. But, more importantly, to the extent that the concern for mutual justifiability is not tied to the actual consent of actual people, it is not clear at all that it is a concern with moral significance. A concern for justifiability to hypothetical others seems to be at most a concern for (maximal) rational consistency, but it lacks a specifically moral force.[28]

But once it becomes doubtful that the antirealist strategy is actually able to account for the moral significance of consent, nothing seems left to counteract the most counterintuitive features of the claim that reasonable agreement constitutes moral rightness. From a cognitive point of view, the equation of reasonableness and infallibility entailed in this claim seems highly questionable. Even if only in light of the pragmatist's idea that what applies to inquiry in general applies to moral inquiry in particular,[29] it is hard to understand why moral reasonableness should guarantee moral rightness in particular, if cognitive reasonableness does not guarantee truth in general. Again, the persistence of hard cases in moral discussion seems to speak against such an assumption. But even from a strictly moral point of view, the claim that what makes a norm just is that reasonable people

[28] In *What We Owe to Each Other*, Scanlon suggests as much, when he claims: "Actual agreement with those around us is not only something that is often personally desirable; it is sometimes morally significant as well. There are many cases in which morality directs us to seek consensus or to secure the permission of others before acting. But where actual agreement is morally significant this reflects a particular substantive judgment within morality, and the significance of this kind of agreement should be clearly distinguished from the ideal of hypothetical agreement which contractualism takes to be the basis of our thinking about right and wrong" (Scanlon 1975, 155). Later in the book, he also draws an equivalent distinction with regard to the "others" to whom justification is owed (ibid., 202). To the extent that these distinctions are taken seriously, it seems clear that no account of the moral significance of agreement and consent (of actual "others") should be expected from an approach with these characteristics. But given that this is usually seen as the most appealing feature of such approaches, it seems to me worth showing that they cannot in fact provide such account.

[29] I take this formulation from Putnam 1994, 175. See also Putnam 2002, 104.

could agree to it has something counterintuitive as well. To paraphrase the usual objection in terms of Russell's concern with emotivism, it just seems hard to believe that all is wrong with wanton cruelty is that I cannot justify it.[30] There are two senses to this objection. At the more superficial level, it seems just false to claim with Scanlon that moral judgments of right and wrong are "judgments about reasons and justification" (Scanlon 1998, 4). Our moral judgments seem hardly ever to be about cognitive disagreements, but first and foremost about violated interests and conflictive actions. Moral judgments may *ask for* justification, but they clearly are not themselves *about* justification. But there is a deeper sense to the objection. Given the holism about justification that these authors accept (see Habermas 1998a, 239–46; Scanlon 1989, 214), it seems clear that what can be justified in a specific context is not only a function of the reasonableness of the participants, but it is necessarily also a function of the substantive beliefs about the world, criteria of valid justification, admissible arguments, etc., that they happen to share at a given time. Depending on how unfortunate such constellations of beliefs turned out to be in a specific community, virtually any norm could seem justified to its reasonable members. This is why ideologies can be so powerful in justifying injustices even in the victim's own eyes. Just a short look at the recent feminist literature on the views on women and other minorities of major figures of the history of philosophy (from Aristotle to Rousseau, Kant, Hegel, etc.) shows how reliable the agreement of reasonable people may actually be.

Of course, this problem can be avoided by strengthening the conditions of reasonableness as to include perfect knowledge. This would make the antirealist claim surely uncontroversial. For moral rightness would no longer be constituted merely by the agreement of reasonable people, but by the agreement of infallible people (i.e., of those who have the right reasons).[31]

[30] Russell famously objected to subjectivist views in ethics with the following remark: "I cannot see how to refute the arguments for the subjectivity of ethical values, but I find myself incapable of believing that all that is wrong with wanton cruelty is that I do not like it" (Russell 1960, 310–311). For an objection to Scanlon's approach along the same lines see Thomson 1990, 30.

[31] This is actually the reading that Scanlon explicitly suggests in his book, when he distinguishes a weaker and a stronger sense of the notion of "good reasons" and claims that the stronger sense is the one relevant for his approach. According to his distinction, a reason is good, not just if it could be convincing for someone (so that it could be her *operative* reason), but only if it is "a consideration that *really* counts in favor of the thing in question" (Scanlon 1998, 19, my italics). This mirrors the standard distinction in epistemology between a weak and a strong sense of justification. In the weak sense, someone is justified in believing something if her reasons are good in the sense that they could be convincing for everyone who is reasonable (i.e., epistemically responsible), whereas in the strong sense, someone is justified only if her reasons actually track the truth (i.e., if they are the right reasons). In the first case the term "justified" is understood in an operational sense, whereas in the second case it is used as an achievement word. For a useful discussion of these two senses of justification see Fogelin 1994. I discuss in much greater detail the implications of this distinction for the antirealist strategy in the last chapter of Lafont 1999. See also footnote 34.

And this is a claim that can hardly be false. But, unfortunately, the moral principle that would result from this transformation would be not only uncontroversial, but also necessarily empty.[32] Whereas it seems possible to give an account of what makes people reasonable in the sense of epistemically responsible,[33] it does not seem possible to give an account of what makes people infallible.[34] Reasonableness so understood would no longer be a characteristic that actual human beings could possibly have. And this, of course, would also have immediate consequences with regard to the moral significance of agreement and voluntary consent. If the only consent that matters for moral rightness is the consent of those who have the right reasons, the claim that the validity of norms is a function of the agreement of those to whom the norms apply would turn out to be just false.

But is there really a way out of this impasse? Is the latter claim compatible with the claim of objectivity after all? If, as it seems, both claims pull in opposite directions, how can a realist strategy that accounts for the objectivity of our normative judgments precisely by denying the claim that agreement constitutes moral rightness, be any more able to account for the moral significance of agreement and consent? Although I cannot address all the difficult issues related to these questions here, in what follows I would like to point very briefly in the direction that a realist strategy could follow in order to account successfully for the truth behind each of the two claims.

[32] In a nutshell, the dilemma facing the antirealist strategy could be stated as follows: A moral principle that requires an agreement on the right reasons would be empty, whereas a principle that merely requires an agreement on the most convincing reasons (even if possibly wrong) would be blind. I analyze this difficulty in much more detail in Lafont 1999, chap. 7.

[33] See footnote 31.

[34] Scanlon admits as much in his book. After spelling out the idealizations involved in the notion of an "ideally rational agent" (namely, "(1) possession of full information about one's situation and the consequences of possible lines of action, (2) awareness of the full range of reasons that apply to someone in that situation, and (3) flawless reasoning about what these reasons support"; Scanlon 1998, 32), he claims: "it seems to me very unlikely that there could be such a thing as a theory of reasons in this sense" (ibid.). Accordingly, when he then moves on to explain the notion of "reasonableness," he offers first of all an operational definition of this notion, according to which "judgments about what it is or is not reasonable to do or think are relative to a specified body of information and a specified range of reasons, both of which may be less than complete" (ibid.). However, he then oscillates once again towards the achievement sense of the notion in the exact same way he did earlier in his account of "good reasons" (see footnote 31), so that in the end it is not clear at all in which of the two senses he wants to interpret the notion of reasonableness that he introduces in his moral principle to explain moral rightness. Later, in chapter 5, he seems to opt once more for the achievement sense of the notion, when he claims: "A claim about what it is reasonable for a person to do presupposes a certain body of information and a certain range of reasons which are taken to be relevant, and goes on to make a claim about what theses reasons, *properly* understood, *in fact* support" (ibid., 192, my italics). In the earlier version of his moral principle that he offered in *Contractualism and Utilitarianism* (Scanlon 1982) he explicitly interpreted the notion of reasonableness in an operational sense, but this move was immediately counteracted through the addition of a further condition of "full information." Accordingly, the agreement that constitutes moral rightness should be not only reasonable and uncoerced, but also "informed." He explained this further condition as follows: "The idea of 'informed agreement' is meant to exclude agreement based on superstition or *false* belief about the consequences of actions, even if these beliefs are ones which it would be *reasonable* for the person in question to have." (Scanlon 1982, 272, my italics).

III. Finding the Right Balance between Realism and Antirealism in Kantian Constructivism

As seen so far, the problem with the antirealist strategy is that it has no resources to motivate the assumption of single right outcomes entailed by the claim of the objectivity of our moral judgments other than by appeal to the notion of reasonableness. However, the inference from reasonableness to single outcomes seems entirely ungrounded. If reasonableness is understood as a possible characteristic of actual human beings, it is just a fact that on difficult moral issues the judgments of reasonable people fail to converge. Alternatively, if reasonableness is idealized as to include correctness as a condition, convergence is just *stipulated* with the very assumption of a single right outcome, but no account is given as to what justifies the stipulation in the first place. The assumption that moral questions have single right answers seems left unaccounted for in both cases. Moreover, both lines of arguments have equally undesirable results with regard to the claim that voluntary agreement is constitutive of the validity of norms. If agreement is understood as the reasonable agreement of actual persons, the claim turns out to be true only in cases of absolute unanimity. But given that those are precisely the cases where agreement is most irrelevant, the moral significance of the claim in light of all other cases seems clearly lost. If agreement is understood as the hypothetical agreement of infallible persons, the claim turns out to be false of actual persons and actual norms. Here its moral significance for all actual cases seems lost.

By comparison, the realist strategy I have sketched in the prior section clearly has the resources to motivate the assumption of single right outcomes entailed by the claim of objective validity of our moral judgments. For, according to it, what we evaluate with our moral judgments is whether the social situation that would result from the general observance of a specific norm is one in which the generalizable interests of all those affected by it are equally protected. If it is, the norm is just. If not, it is not. In light of the realist sense of the claim, this strategy has no problem accounting for the lack of convergence of the judgments of reasonable people in difficult moral cases. Given that our moral judgments are about a social circumstance, whose obtaining is logically independent of any agreement, our moral judgments may be mistaken in difficult cases as much as any other cognitive judgments may.[35] However, the difficulty in this case may seem to

[35] There are many factors in our assessments of the justice of norms that explain why the correctness of our moral judgments can transcend the given epistemic situation of reasonable people. First of all, reasonable people may disagree about which of our interests and needs are really basic, rational, generalizable, etc. But even under the presupposition of an ideal transparency towards *our own unrenounceable interests*, this provides no guarantee concerning knowledge of the unrenounceable interests of others. Yet such knowledge is equally necessary for a correct assessment of the justice of a norm. Moreover, even if all of us could never possibly be mistaken concerning our own unrenounceable interests, this would still not guarantee infallible knowledge concerning the objective consequences that a norm in the long run and under

come from the other direction: If the claim that reasonable agreement constitutes moral rightness is false, how can the claim that the validity of social norms depends on the voluntary agreement of those to whom these norms apply be true? In order to answer this question, though, some further distinctions must be introduced.

According to the realist strategy, the reasonable agreement of all possibly affected by a norm does not constitute its moral rightness. It just offers the best epistemic support for the supposition that the norm at issue is in fact morally right. Reasonable agreement cannot *guarantee* the moral rightness of a norm, simply because *nothing* can. But it can *entitle us* to claim moral rightness for a norm, as long as no counterarguments appear (whether on the basis of new experiences, consequences, side effects or any learning processes in general). Thus, although reasonable agreement is not a necessary condition for the moral rightness of a norm, it is a necessary condition *for us to tell* whether a specific norm is morally right. This is one sense, in fact a *cognitive* sense, in which reasonable agreement matters for the validity of social norms. But precisely the logical gap between agreement and moral rightness opens up the possibility that reasonable people fail to converge in their judgments in cases of difficult moral conflicts. Given this possibility and the need to reach some decision as to which norm to implement in order to avoid those conflicts, a further distinction is necessary to evaluate the validity of social norms. One aspect of the validity of norms is their justice or moral rightness. Another aspect is the legitimacy of their implementation. With regard to the latter, reasonable agreement and voluntary consent matters for the validity of social norms in another sense, namely, a *volitive* sense.

In this context, it is important to keep in mind that although the notions of justice and legitimacy are internally related, they express two genuinely different senses in which norms can be considered valid or invalid. On the one hand, even if justice is considered to be a necessary condition for legitimacy, it is surely not a sufficient one. That is, the fact that a norm is just does not make it legitimate. Thus, between two equally just norms or regulations, only the one voluntarily agreed upon by a specific political community is legitimate, according to this view. On the other hand, this

changing (i.e., currently unpredictable) circumstances would have for all those who are possibly affected. Discrimination is not always a consequence of the repressed will of those affected by it. It can also result from our incapacity to foresee the side effects of a norm in the long run, or even from our inability to imagine a more satisfactory norm, despite all our best intentions. It does not seem meaningless to say that we might find out that a norm, despite our general agreement (based on our prior epistemic situation) as to its moral rightness, turned out to be morally wrong (i.e., *in fact* unfair, discriminatory or the like). But once we recognize that factual knowledge is an essential component of the reasons employed in moral discourse to determine the justice of norms there seems to be no obvious reason left to insist on the only claim that distinguishes the realist from the antirealist strategy, namely, the claim that our epistemic attitudes under ideal conditions are infallible. We can just drop this dubious claim and maintain a fallibilist attitude towards our moral claims as much as we do with any other cognitive claims.

dimension of factuality gives the notion of legitimacy an operational sense that the notion of justice does not have: a norm is de facto legitimate if it is met with the assent of the participants of a given political community under reasonable conditions of deliberation (on the basis of their given epistemic situation). Thus, if the same community decides to revoke the norm in the future because it turns out to be unjust in an unpredictable way (e.g., some of its consequences or side effects turn out to be discriminatory, a much better norm is found, etc.), it still makes sense to say that it was *de facto* legitimate, but it is no longer so. By contrast, the notion of justice as we use it does not have such an operational sense. If a norm turns out to be unjust, it was always unjust.

In my opinion, these differences in our use of both notions can only be accounted for by combining the realist and the antirealist elements of Kantian constructivism, instead of reducing one to the other. An antirealist strategy seems appropriate for explaining the notion of legitimacy. It is *prima facie* plausible to claim that reasonable agreement constitutes legitimacy, precisely because legitimacy does not require single right outcomes and thus can be understood as a purely procedural notion. Accordingly, the question of the legitimacy of a norm does not have a single right answer because it cannot have *any* answer prior to or independently of carrying out the deliberative procedure in which a *factual* reasonable agreement among the participants is reached. But this antirealist strategy is inappropriate for explaining the notion of justice. The fact of reasonable agreement cannot make a norm any more or less just than it actually is, whereas it surely makes an essential contribution to the legitimacy of its implementation.[36]

Now the interesting question is whether this combination of strategies allows for a better account of the moral significance of consent. To the extent that the antirealist strategy is still employed in order to explain legitimacy, it may seem that the structural problems of that strategy would now reappear. How can the claim that social norms can only be valid to the extent that those to whom these norms apply have voluntarily agreed to submit to

[36] Here there may seem to be an additional difficulty involved in following this strategy. How can legitimacy be a purely procedural notion, if justice is a necessary condition for legitimacy and it is not itself a procedural notion? I think that the difficulty is only apparent, though. As the example mentioned above shows, due to its operational sense legitimacy is constrained not by injustice *per se* but by *perceived* injustice. Once a norm is perceived as unjust, that is, once some participants can provide reasons to show the specific way in which the norm is unjust, its legitimacy will be undermined. But this is precisely a situation in which the norm would no longer meet with the assent of all reasonable participants. And it is for this reason that the norm will no longer be legitimate, and not because of its putative injustice. Otherwise legitimacy would be constrained already by *possible* and not only by *real injustice*, as it should be. This explains the different significance that agreement and disagreement have with regard to the notions of justice and legitimacy. On the one hand, the logical space between possible and real injustice creates the logical space for rational disagreements. But, on the other hand, the need to decide here and now which one of the two obtains with regard to the norms we need to enforce creates the space for the distinction between legitimate and illegitimate ways of deciding, even in the face of unavoidable disagreements.

them be true in cases of reasonable disagreements? Why should those who disagree give their voluntary consent to norms they think are wrong? It is with regard to this question[37] that in my opinion the combination of realist and antirealist strategies pays its highest dividends.

The major problem in following an exclusively antirealist strategy is that reasonable agreement has to account for both dimensions of the validity of norms. As a consequence, justice and legitimacy become indistinguishable. By contrast, a realist strategy can maintain their logical independence. For according to this view the validity of a norm depends not only on its legitimacy (on whether it could meet with the reasonable assent of the participants), but also on its justice (on whether it would in fact be equally in everyone's interest). And unless participants see themselves as omniscient, there is no way to reduce the second question to the first. This distinction in turn provides the necessary resources to account for the different senses in which the reasonableness of an agreement and the voluntariness of consenting to it matter for the validity of social norms.

Given that the assumption of a single right answer with regard to the justice of norms by no means implies (or even suggests) that we will be able to find that answer regardless of how lucky our epistemic conditions may be, reasonable disagreements in difficult moral cases are just a natural consequence of the fact that, even among reasonable participants, anyone can get the answer wrong as much as anyone else. Thus, those participants in a reasonable process of deliberation that on a given occasion disagree with the agreement reached by the majority may still give their voluntary consent to it for the *cognitive* reason that they failed to convince the majority that the norm is actually unjust and not only putatively so. And this could be an indicator of their being wrong in that case, if anything is. Precisely to the extent that participants consider reasonable agreement to be a condition for legitimacy, the minority's failure to provide convincing arguments to the majority here and now requires them to accept the factual outcome of the deliberation process even by their own lights and thus *voluntarily*. But precisely because participants do not regard reasonable agreement as constitutive of justice, the conditional agreement of the minority by no means makes the norm thereby any more or less just than it actually is. Thus, the minority's success in finding convincing arguments at a future time to show the specific way in which the norm is actually unjust would be all it takes to undermine the prior, majoritarian agreement, even by the majority's own

[37] Needless to say, here I cannot address all issues related to this difficult question. My present aim is only to point very briefly to the way in which a realist strategy can provide a distinctive account of the notions of legitimacy and justice and of their internal relationship without collapsing one into the other. I explain in much greater detail the specific conception of legitimacy that follows from this strategy in my discussion of Habermas's approach to deliberative democracy in Lafont 2003b. For a very illuminating discussion of conceptions of legitimacy on the basis of distinctions similar to the ones I employ here see Estlund 1997. For an excellent overview of the current debates on these issues see Bohman and Rehg 1997.

lights.[38] Only under the assumption that the justice of a norm is logically independent of the reasonableness of the agreement that brought its implementation about is it understandable why in cases of reasonable disagreement the minority may give its voluntary consent to the outcome of their deliberation process without thereby having to change their minds about the justice of the norm, and also why the majority has to admit that the norm may still be unjust in spite of the reasonableness of the agreement reached so far.

Along these lines it is possible to see how the different elements contained in the antirealist strategy get reorganized, if a realist strategy is followed. On the one hand, the moral significance of voluntary consent is accounted for, precisely to the extent that, according to this view, it is the *actual* voluntary consent of all *actual* participants in a reasonable process of deliberation of any *actual* political community that matters for the legitimacy of any norm, as it should be. On the other hand, precisely because the reasonableness of the agreement does not guarantee that the norm is just, reasonableness can retain its operational sense. Accordingly, the reasonableness of an agreement is a function of the best epistemic resources available at a given time and not of any (supposedly) infallible ones, as it should be. Only on this basis can the moral constraint of mutual justifiability turn from a hypothetical constraint towards hypothetical others (who can always be thought of as sharing our favored arguments or beliefs) into a real constraint towards all actual members of the community to which the norms would apply, who reasonably disagree with our arguments and views. Our moral obligation is to find the specific arguments that would succeed in bringing *them*, *as they actually are*, from their reasonable but different views to ours. But, moreover, given that our success at reaching a unanimous agreement at a given time does not make our norms any more just than they actually are, our moral obligation cannot stop there. We still need to be vigilant to the (ever-present) possibility of undetected injustices and powerful ideologies that such agreements may contain. For we may always discover in the future that for all our reasonableness, we were nonetheless mistaken about the justice of any of our social norms. This is precisely the moral significance of rejecting the antirealist claim that reasonable agreement is all that justice requires.

Northwestern University
Department of Philosophy
1880 Campus Drive
Evanston, IL 60208
U.S.A.

[38] However, in cases with these characteristics the need to revoke the norm by no means implies necessarily that the decision in their prior epistemic situation was illegitimate or the agreement unreasonable after all. It can just mean that by virtue of their new epistemic situation (based on new arguments, experiences, etc.) they are now able to see the specific way in which the norm is actually unjust.

References

Barry, B. 1989. *Theories of Justice*. Berkeley, CA: University of California Press.
———. 1995. *Justice as Impartiality*, Oxford: Clarendon.
Bohman, J., and W. Rehg. 1997. *Deliberative Democracy*. Cambridge, MA: MIT Press.
Brink, D. 1989. *Moral Realism and the Foundations of Ethics*. Cambridge, MA: Cambridge University Press.
Darwall, S., A. Gibbard, P. Railton. 1997. Toward Fin de Siècle Ethics: Some Trends. In S. Darwall, A. Gibbard, P. Railton, *Moral Discourse and Practice: Some Philosophical Approaches*. Oxford: Oxford University Press.
Estlund, D. 1997. Beyond Fairness and Deliberation: The Epistemic Dimension of Democratic Authority. In *Deliberative Democracy*. Ed. J. Bohman and W. Rehg, 173–204. Cambridge, MA: MIT Press.
Fogelin, R. 1994. *Pyrrhonian Reflections on Knowledge and Justification*. Oxford: Oxford University Press.
Habermas, J. 1975. *Legitimation Crisis*. Trans. Th. McCarthy. Boston, MA: Beacon.
———. 1990. *Moral Consciousness and Communicative Action*. Trans. C. Lenhardt and S.W. Nicholsen. Cambridge, MA: MIT Press.
———. 1993. *Justification and Application*. Cambridge, MA: MIT Press.
———. 1996. *Between Facts and Norms*. Trans. W. Rehg. Cambridge, MA: MIT Press.
———. 1998a. *On the Pragmatics of Communication*. Cambridge, MA: MIT Press.
———. 1998b. A Genealogical Analysis of the Cognitive Content of Morality. In J. Habermas, *The Inclusion of the Other*, 3–48. Cambridge, MA: MIT Press.
———. 1999. Richtigkeit versus Wahrheit. Zum Sinn der Sollgeltung moralischer Urteile und Normen. In J. Habermas, *Wahrheit und Rechtfertigung*, 271–318. Frankfurt: Suhrkamp.
Hill, Th. E. 1989. Kantian Constructivism in Ethics. *Ethics* 99: 752–70.
Kant, I. 1964a. *Groundwork of the Metaphysics of Morals*. Trans. H.D. Paton, New York, N.Y.: Harper & Row. (1st ed. 1785.)
———. 1964b. *Metaphysics of Morals*. Part 2. Trans. M. Gregor as *The Doctrine of Virtue*. New York, N.Y.: Harper & Row. (1st ed. 1797.)
Krasnoff, L. 1999. How Kantian is Constructivism. *Kant-Studien* 90: 385–409.
Lafont, C. 1998. Pluralism and Universalism in Discourse Ethics. In *A Matter of Discourse: Community and Communication in Contemporary Philosophies*. Ed. A. Nascimento, 55–78. London: Avebury.
———. 1999. *The Linguistic Turn in Hermeneutic Philosophy*. Cambridge, MA: MIT Press.
———. 2002. Realismus und Konstruktivismus in der Kantianischen Moralphilosophie. Das Beispiel der Diskursethik. *Deutsche Zeitschrift für Philosophie* 50: 39–52.
———. 2003a. Procedural Justice? Implications of the Rawls-Habermas Debate for Discourse Ethics. *Philosophy and Social Criticism* 29: 167–85.
———. 2003b. Justice and Legitimacy: The Intricate Relation of Morality to Politics. Manuscript.
Mackie, J.L. 1977. *Ethics: Inventing Right and Wrong*. London: Penguin Books.
Milo, R. 1995. Contractarian Constructivism. *The Journal of Philosophy* 92: 181–204.
O'Neill, O. 1989. *Constructions of Reason. Exploration of Kant's Practical Philosophy*. Cambridge, MA: Cambridge University Press.
———. 2003. Constructivism in Rawls and Kant. In *The Cambridge Companion to Rawls*. Ed. S. Freeman, 347–67. Cambridge, MA: Cambridge University Press.
Putnam, H. 1994. Pragmatism and Moral Objectivity. In H. Putnam, *Words and Life*, 151–81. Cambridge, MA: Harvard University Press.
———. 2002. *The Collapse of the Fact/Value Dichotomy and Other Essays*. Cambridge, MA: Harvard University Press.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
———. 1993. *Political Liberalism*. New York, NY: Columbia University Press.
———. 1995. Reconciliation through the Public Use of Reason. *Journal of Philosophy* 92: 132–80.
———. 1999a. Kantian Constructivism in Moral Theory. In J. Rawls, *Collected Papers*. Ed S. Freeman, 303–58. Cambridge, MA: Harvard University Press. (1st. ed. 1980.)
———. 1999b. Themes in Kant's Moral Philosophy. In Rawls, *Collected Papers*. Ed. S. Freeman, 497–528. Cambridge, MA: Harvard University Press. (1st. ed 1989.)
———. 1999c. Justice as Fairness: Political not Metaphysical. In J. Rawls, *Collected Papers*. Ed. S. Freeman, 388–414. Cambridge, MA: Harvard University Press. (1st. ed 1985.)
———. 2000. *Lectures on History of Moral Philosophy*. Ed. B. Herman. Cambridge, MA: Harvard University Press.
Rousseau, J.-J. 1994. *The Social Contract*. Oxford: Oxford University Press. (1st 1762.)
Russell, B. 1997. Notes on Philosophy. In B. Russell, *Last Philosophical Testament 1943–68*. Ed. J. Slater. London: Routledge. (1st ed. 1960.)
Sayre-McCord, G. 2000. Contractarianism. In *The Blackwell Guide to Ethical Theory*. Ed. H. La Follette, 247–67. Malden, MA: Blackwell.
Scanlon, T. 1975. Preference and Urgency. *The Journal of Philosophy* 72: 655–69.
———. 1982. Contractualism and Utilitarianism. In *Utilitarianism and Beyond*. Ed. A. Sen and B. Williams, 103–28. Cambridge, MA: Cambridge University Press.
———. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.
Thomson, J. 1990. *The Realm of Rights*. Cambridge, MA: Harvard University Press.