CRÍTICA, *Revista Hispanoamericana de Filosofía*
Vol. XXXII, No. 94 (abril 2000): 43–88

MORAL PRINCIPLES AND AGREEMENT

FAVIOLA RIVERA
Instituto de Investigaciones Filosóficas
UNAM

Does agreement justify moral principles? Clearly, not any actual agreement could play this justificatory role since agreements are often influenced by power, bias, and prejudice. But, does a rational agreement (either possible or actual) justify moral principles; that is, one reached (or that could be reached) under some ideal conditions which exclude the influence of power, bias, and prejudice? The two most successful theories of justice and political legitimacy today, namely, John Rawls's[1] and Jürgen Habermas's,[2] assign this justificatory role to agreement under some ideal conditions. In his theory of justice, Rawls proposes to regard the principles of political justice as the focus of an agreement in what he calls an "original position", which is a point of view subject to constraints intended to guarantee the fairness of the agreement (*TJ* 11–21). Habermas, on the other hand, argues that only a rationally motivated agree-

[1] John Rawls, *A Theory of Justice*, Harvard University Press, Cambridge, Mass., 1971. Hereafter referred to in the main text as *TJ* followed by the page number; and *Political Liberalism*, Columbia University Press, New York, 1993. Hereafter referred to in the main text as *PL* followed by the page number.

[2] Jürgen Habermas, *Between Facts and Norms. Contributions to a Discourse Theory of the Law and Democracy*, trans. by William Rehg, MIT Press, Cambridge, Mass., 1996.

ment, one reached under the conditions of what he calls a "practical discourse", can support the universal validity of moral norms.[3] Rawls offered his social contract approach as an alternative to utilitarian accounts of justice. The success of his proposal has motivated some moral philosophers to extend the contractarian approach to the domain of *personal morality*[4] in order to provide an alternative to utilitarian —and more generally consequentialist— accounts of moral justification.[5] They propose to regard principles of personal morality as the focus of an agreement as well.

Although Rawls has emphasized that he intended the social contract approach for the solution of questions of political justice only, I will argue that he also understands the justification of principles of personal morality in terms of an agreement among persons.[6] Habermas, on the other

[3] J. Habermas, "Discourse Ethics: Notes on a Program of Philosophical Justification", in his *Moral Consciousness and Communicative Action*, trans. by Christian Lenhardt and Shierry Weber Nicholsen, MIT Press, Cambridge, Mass., 1993. Hereafter referred to in the main text as "DE" followed by the page number.

[4] With the expression "personal morality" I do not mean to imply that morality is something subjective. By personal morality I mean only the morality we think should not be coercively enforced by the political authority, that is, the moral principles and values by which we guide ourselves in our everyday lives. In contrast, the principles of political justice are principles for social and political institutions and are thus enforced upon us as citizens by the political authority. I mean this contrast between personal and political morality to be more or less rough and intuitive. I will refine it a bit more later.

[5] The best example is perhaps T.M. Scanlon's, *What We Owe to Each Other*, Harvard University Press, Cambridge, Mass., 1999. See also Ronald Milo, "Contractarian Constructivism", *The Journal of Philosophy*, 4, 1995, pp. 181–204.

[6] By "political justice" I mean the domain of political as opposed to personal morality. This is the domain of law enforced by the political authority. Rawls's principles of justice belong here since they are guides for the design of a legal system. It is important to keep in mind, however, that I am *not* using the term "political" in the

hand, explicitly presented his own moral theory as a theory about the justification of moral principles in general.[7] On his view, an agreement under ideal conditions justifies all kinds of moral principles, and not only those of political justice. Both Rawls's and Habermas's views on the justification of principles of personal morality go back to Kant. According to Kant's first formulation of the categorical imperative in the *Groundwork of the Metaphysics of Morals* (the formula of universal law) we ought to act on maxims that we can at the same time will as universal laws.[8] Rawls and Habermas share a widespread interpretation of this principle according to which to universalize a maxim is to check whether it could be willed by everyone. In other words a universalizable maxim, on this interpretation, is one that could be agreed upon by all agents. I will refer to this as "the interpretation of the universalizability test as an agreement test".

In this paper, I will argue for the following two claims. First, that the view according to which the justification of

technical way developed by Rawls. Thus, I call Kant's principles of justice "political" only to highlight the fact that they belong to the domain of legislation enforced by the political authority although they clearly are *not* "political" in Rawls's sense. I will say more about this in section 3.

[7] This claim might seem mistaken in light of a distinction between the democratic and the moral principle that Habermas makes in the 1994 Postscript to the English version of *Between Facts and Norms*. According to this distinction, the democratic principle applies to *legal* norms only (which, in my terminology, belong to the domain of political justice), whereas the moral principle applies to all kinds of moral norms. Despite this distinction, my claim still holds as I will explain in section 4.

[8] Translated by Mary Gregor, Cambridge University Press, Cambridge, 1991. Section two. Hereafter referred to in the main text as *G* followed by the volume and page number of *Kants gesammelte Schriftem* (published by the Preussische Academie der Wissenschaften, Berlin) and by the page number of the English translation.

principles of personal morality rests upon the agreement of everyone (either possible or actual) presupposes what I will call an "other-regarding morality". By this I mean the view according to which all moral questions are about how we ought to relate to each other, and all moral duties are duties to others. My first claim, then, is that we might understand the justification of principles of personal morality in terms of the agreement of everyone provided that we are committed to the view that morality is about the regulation of human interaction.

My second claim is that the agreement test interpretation of Kant's universalizability test implicitly attributes to him an other-regarding view of morality *and* that this attribution rests upon a confusion between ethics and justice, which are the two subdomains of morality clearly distinguished by Kant in *The Metaphysics of Morals*.[9] Kant's distinction between ethics and justice roughly corresponds to my distinction between personal morality and political justice. More precisely, I will argue that the agreement test interpretation presupposes an understanding of personal morality or ethics on analogy with political justice. As I will explain later, only in the domain of justice are moral duties exclusively other-regarding, according to Kant, whereas in the domain of ethics they are primarily self-regarding (though some ethical duties are duties to others). Although I cannot argue for this here, Kant's distinction between ethics and justice offers a way of understanding person-

[9] Translated by Mary Gregor, Cambridge University Press, Cambridge, 1991. Hereafter referred to in the main text as *MM* followed by the volume and page number of *Kants gesammelte Schriftem* (published by the Preussische Academie der Wissenschaften, Berlin) and by the page number of the English translation.

al morality that is an alternative to other-regarding morality.[10]

[10] Other-regarding morality cannot account for some central moral concepts such as respect for oneself, nor for the concern with one's own character which is at the center of Kant's view of ethics, as I will explain later. However, the complete treatment of the claim that Kant's ethics provides an attractive alternative to other-regarding morality would be the topic of another paper. My main aim now is to argue against an interpretation of the categorical imperative on the grounds that it presuposes a mistaken view of Kant's view of ethics.

Despite this claim concerning the limits of this paper I believe that, if successful, my argument might have a broader scope. The two main contenders in moral philosophy today are Kantianism and consequentialism. Consequentialists believe that morality is about maximizing quantities of some kind of value in the world: good consequences, happiness, utility, satisfaction, etc. They do not think that morality is about how we ought to relate to other people. Indeed, it is often said that what distinguishes consequentialism from Kantianism is that whereas the former takes morality to be about maximizing quantities of a certain kind of value in the world, the latter is about living up to certain values in our relations with other people. Those who think that morality is about the regulation of human interaction are "Kantian" in a broad sense, and this obviously includes those who endorse an other-regading morality. So, my second claim can also be taken in the following way. The "Kantian" view according to which morality is about the regulation of human interaction fails to take into account Kant's distinction between ethics and justice and does indeed conflate these two domains. At a different level, this kind of confusion is also present in the extension of Rawls's contractarian approach to the domain of personal morality as well as in Habermas's unified account of moral justification: in both instances agreement is taken to play a justificatory role in personal morality because personal morality has been understood on analogy with political justice.

I should mention that I think "Virtue Ethics" not to constitute a separate category alongside Kantianism and Consequentialism. A theory of virtue is not about the foundations of morality as Kantianism and Consequentialism are. Instead, it is a theory about what is involved in the development of moral character: about the interaction between reason and sentiment as well as the role of practice. Thus, a theory of virtue presupposes an account of the foundations of morality, and it comes as a second step after an account of what morality requires is available. Then, both Kantianism and Consequentialism have or ought to have their own theories of virtue. See Christine Korsgaard, I want to emphasize that my aim is *not* to question the understanding of *political* justification in terms of an agreement among citizens. Nor it is my aim to dispute Rawls's and Habermas's employment of some of Kant's moral insights for the purposes of theories of political justice and of political legitimacy.[11] My objection is *not* directed against neo-Kantian accounts of political justice and legitimacy, which I believe to be quite fruitful. Instead, my purpose is to challenge the widely held assumption that it is also quite right [within a Kantian view] to regard principles of *personal morality* as the object of an agreement. My objection is directed against the tendency to assume that a successful account of how to arrive at principles of political justice can serve as a model for arriving at principles of personal morality.

The paper is organized as follows: In section 1, I briefly introduce Kant's own distinction between personal morality and political justice. Section 2 contains a very brief characterization of Kant's universalizability test. In sections 3 and 4, I consider Rawls's and Habermas's interpretation of the universalizability test as an agreement test. I argue that both of them presuppose an other-regarding view of morality which reveals an understanding of personal morality on analogy with political justice. Concerning Rawls, I

"From Duty and for the Sake of the Noble: Kant and Aristotle on Morally Good Action", in Stephen Engstrom and Jennifer Whiting (eds.), *Aristotle, Kant, and the Stoics: Rethinking Happiness and Duty*, Cambridge University Press, Cambridge, 1997, fn. 20, pp. 232–234.

[11] Otfried Höffe has argued against Rawls's use of some aspects of Kant's moral philosophy for the purposes of addressing the question of justice. Höffe's objection is that Rawls proceeds without taking into account Kant's distinction between ethics and justice. I am not concerned with raising this kind of criticism which I believe not to pose a problem for Rawls's theory of justice. See Höffe, "Is Rawls's Theory of Justice Really Kantian?", *Ratio*, 2, 1984, pp. 103–124.

show how his interpretation of what he calls the "categorical imperative procedure" mirrors his own account of the original position, which he expressly designed for the justification of principles of political justice. Finally, I show that Habermas's intersubjective reformulation of the categorical imperative presupposes that moral questions are about how to solve conflicts that arise in human interaction. On his account, moral norms are a sort of mechanism for coordinating human interaction; they govern external acts and do not require us to act from moral motives. Thus, I argue that Habermas's other-regarding morality takes the norms of political justice as the model for all moral norms.

1. *Kant on Justice and Virtue*

It has become a commonplace to think that Kantian morality is concerned with universal rights and justice and to contrast it with views that are concerned with virtue and character. According to this widespread view, Kantian morality is about our duties and obligations *to others*, whereas theories of virtue are about the *agent-centered* cultivation of a good character. This other-regarding view of Kantian morality is, however, very much mistaken. This becomes clear as soon as we consider Kant's own distinction between justice and virtue in *The Metaphysics of Morals*. This is a distinction he draws *within* the domain governed by the categorical imperative. The details of his account are quite complex. For our present purposes, it will be enough to bring out the central aspects of the justice/virtue distinction in order to dismiss the other-regarding interpretation. As I will argue in detail in later sections, the agreement test interpretation of Kant's categorical imperative presupposes an other-regarding view of morality. The task of this section is to reject this other-regarding view. With this in place, the allegedly justificatory role of agreement in personal morality will become much less plausible.

Justice

Kant's *Metaphysical First Principles of the Doctrine of Justice*[12] contains his account of the justification of individual rights and of the political authority. A "right" according to Kant, is an authorization to use coercion. To say that you have a right is to say that there is an authorization to coerce another who obstructs the exercise of your right. The point of rights is to set limits to our actions in order to make possible the ordered coexistence of everyone's exercise of their freedom of action (the freedom to act in the world without being hindered by others). Kant argues that only the political authority can be authorized to use coercion in order to protect individual rights. So, he argues that insofar as we ought to live under conditions in which everyone's rights are protected, we ought to live under a political authority. According to Kant, rights have the following two central characteristics: their legislation is external, and they correlate with duties on the side of others, which duties are always other-regarding, perfect, and strict. I will take up these two features in reverse order.

Duties of justice are necessarily *other-regarding* because they are duties not to interfere with the exercise of other people's rights. To say that duties of justice are *perfect* is to say that they are *owed* to specific persons (who can be one person in particular —say, your partner in a contract— or every person —all other citizens of the same State).[13] And to say that the fulfillment of these rights is *strict* is to

[12] This is the first part of *The Metaphysics of Morals*.

[13] There has been a good deal of debate in the literature about how to understand Kant's twofold division of duties: perfect/imperfect and strict/wide. The interpretation that I am introducing here is my own.

say that the requirement is to perform (or to refrain from performing) a specific *outward act* (as opposed to adopting a certain maxim). Maxims are the principles on which we act, and they always contain our reasons or motives for action. Thus, since they do not require us to adopt any maxims in particular, duties of justice do not require us to act on any specific motives. In order to comply with your duties of justice, all you have to do is to perform (or to refrain from performing) certain outward acts from any motives whatever.

According to the other feature of rights (and of their correlative duties of justice) mentioned above, their legislation is external. This will bring out the internal connection between the justification of rights and the agreement of all.[14] That the legislation of duties of justice is external means that these duties are legislated for us by others. This might seem surprising because Kantian morality is supposed to be all about self-legislation. However, Kant also thinks that we can legislate duties to one another and that the content of this legislation is rights and their correlative duties of justice. How could this be possible? Duties arise from the will's legislation, and we can legislate duties to others when we share one common will.[15] We come to share one will when we make a contract or an agreement. If you and I agree to an exchange of goods, say, you and I share one will. In this case the content of our common will is to exchange these goods. Since we share one common will, we are both in a position to determine laws for it: both of us have the authority to legislate certain actions for ourselves and for each other. Now, according to Kant, the authority to legislate is the same as the authority to coerce: to

[14] As I will explain below, this connection is absent in the justification of duties of virtue.

[15] I am following here a suggestion by Christine Korsgaard made to me in conversation.

say that I have the authority to legislate laws for myself means that I have the authority to coerce myself to comply with them. Therefore, when we are in a position (e.g., authorized) to legislate certain actions to each other, we are also authorized to coerce each other to perform them. This is where the authorization to coerce another contained in claims of right comes from. According to Kant, then, rights and their correlative duties of justice presuppose the existence of a common will. This is why the establishment of a common will through the unanimous agreement of all is the foundation of all right and justice: we are entitled to make right claims against each other only under the presupposition that we all share one common will, the embodiment of which, according to Kant, is the political authority.

This extremely compressed presentation of Kant's account of rights and their correlative duties of justice has brought out two important facts: first, that duties of justice are always other-regarding, and that the foundation of individual rights is a common will which we establish through the agreement of everyone. Kant's view is not so much that we ought to enter into that agreement, but that insofar as we regard ourselves and each other as having certain rights, we ought to presuppose a common will.[16]

[16] See Christine Korsgaard, "Taking the Law into Our Own Hands: Kant on the Right to Revolution", in *Reclaiming the History of Ethics. Essays for John Rawls*, in Andrews Reath, Barbara Herman, and Christine Korsgaard (eds.), Cambridge University Press, Cambridge, 1997.

Virtue

Kant's *Metaphysical First Principles of the Doctrine of Virtue*[17] offers quite a different picture: it is not about the regulation of external conduct but about the acquisition of a morally good character. Duties of virtue are duties to adopt certain maxims, and their adoption necessarily involves the practice of the actions that they require. For example, virtue requires the adoption of a maxim of beneficence, and Kant is explicit that the way to adopt this maxim is by actually helping other people meet their (permissible) ends. That is, we can adopt this maxim only through the practice of beneficent actions[18] "Action" here does not merely mean an outward act; it also includes the motives or reasons for the act, which are obviously moral since the maxims that ethics requires are themselves moral.

According to Kant, these maxims required by ethics are maxims of *ends*. He identifies two general categories of ends: the happiness of others and one's own perfection. These ends can be further specified into a multiplicity of ends: the end of the happiness of others, for example, can be further specified into the ends of helping others where one can, helping one's neighbors, helping one's parents, and so forth; the end of one's own perfection can be further

[17] This is the second part of *The Metaphysics of Morals*. Hereafter referred to in the main text as *DV* followed by the volume and the page number of *Kants gesammelte Schriften* (published by the Preussische Akademie der Wissenschaften, Berlin) and by the page number of the English translation.

[18] The passage in *The Doctrine of Virtue* that is most relevant to this point is the one we find at 6:402/203, where Kant claims that through the constant practice of beneficent actions one "eventually comes actually to love the person [one] has helped". Kant's point here is about the acquisition of a feeling of love for others which, he says, cannot be a duty but accompanies the practice of beneficence, which is a duty. Practice, however, is equally important for the acquisition of the disposition to beneficence.

specified into the ends of cultivating one's talents, cultivating one's talent for languages, perfecting one's knowledge of Arabic, and so forth. The two general categories of ends are two different ways of making humanity one's own end, in one case in the person of others, and in the other case, in one's own person. In the *Groundwork* Kant has already put forward the claim that humanity is the only unconditional and independently existing end, and that moral conduct is about treating humanity always as an end and never as a mere means (*G* 4:437/44).[19] Of course, in the *Groundwork* Kant claims that to act morally is to act on the categorical imperative, and to those who tend to focus on the universal law formulation of this principle, it comes as a surprise that *The Doctrine of Virtue* is a doctrine of ends. I believe that Kant formulates the duties of virtue as maxims of *ends*[20] (and not as universalizable maxims, which they also are) for a reason that he gives in the *Groundwork*: because the formula of humanity brings the moral law closer to intuition (*G* 4:437/44).

The Doctrine of Virtue is not concerned with the foundation of morality. Kant carried out this task of grounding morality in the *Groundwork* and in the second *Critique*. Instead, the concern now is with the *application* of the

[19] All moral ends are unconditional such as the happiness of others, one's own perfection, a republican constitution and perpetual peace. But only humanity is an independently existing end because it does not need to be brought about through our actions as the other moral ends. Humanity is the basis of all moral ends in the sense that they are different ways of living up to the requirement of treating humanity as an end in itself. On this point see the discussion by David Velleman in "Love as a Moral Emotion", *Ethics*, 109, 1999, pp. 338–374.

[20] Taking *The Doctrine of Virtue* as the central text, Allen Wood has argued that the central formulation of the categorical imperative is the formula of humanity and not the formula of universal law. See his *Kant's Ethical Thought*, Cambridge University Press, Cambridge, 1999.

moral principle, which application must take into account "the particular nature of man" (*MM* 6:44/217). *The Doctrine of Virtue* takes into account some general facts about us regarded as natural beings: that we are susceptible to the influence of *inclination*, that we are passive beings with a capacity to *feel*, and that we always act with an *end* in view. In contrast with purely rational beings, finite beings like us are also passive, that is, susceptible to feelings that happen in us. These feelings work in us as incentives or deterrents, and are at the basis of all desire, aversion, and inclination. *The Doctrine of Virtue* takes account of the fact that we need incentives in order to act and that these incentives are feelings. Thus, the adoption of maxims necessarily involves the cooperation of feelings that are favorable to morality. The feeling of respect, which is the incentive to morality, arises in us through the influence of the thought of duty and of examples of moral conduct; the feelings of love of humanity and of sympathy with the needs of others arise in us through the constant practice of required actions. According to Kant, if we lacked the capacity for these moral feelings, we could not possibly be moved to moral action (*DV* 6:399–403/200–204). Also, as finite beings, we can only carry out over time what we set ourselves to do, that is, we act on ends. Thus, the representation of the end of humanity is more intuitive for us than the representation of the lawfulness of a maxim precisely because we are teleological beings; we pursue ends.[21] The

[21] This is not inconsistent with Kant's claim in *The Doctrine of Virtue* that ethics provides an end because "since men's sensible inclinations tempt them to ends (the matter of choice) that can be contrary to duty, lawgiving reason can in turn check their influence only by a moral end set up against the ends of inclination, an end that must therefore be given a priori, independently of any inclinations". (*DV* 6:186/380–381). We are susceptible to the influence of inclination because we are finite or passive, and it is this same feature that makes us need the representation of an end. Standing up against our incli-

pursuit of moral ends (the happiness of others and one's own perfection) necessarily involves the constant practice of morally required actions. Through this practice we come to acquire the settled disposition to act on the maxims of ends, a disposition which includes the presence of certain moral feelings. Although some ethical duties are duties to others (i.e., all the duties that fall under the end of the happiness of others), the acquisition of a moral disposition is primarily agent-centered. This suffices to reject the attribution of an other-regarding view of morality to Kant.

It is important to keep in mind that Kant's justice/virtue distinction holds *within* the moral domain. Since the categorical imperative is, as he argues in the *Groundwork*, the supreme principle of morality, these two sub-domains of morality must be in some way governed by this principle. Justice and virtue have each of them a governing supreme principle, which is, in each case, subsidiary to the categorical imperative. These principles are the universal law of justice and the supreme principle of virtue. I will not enter here into the debate about the sort of relation that exists between the categorical imperative and these two subsidiary principles.[22] The important point here is that if we want to trace back to Kant the claim that the justification of princi-

nations contrary to morality (not against *all* inclinations, of course) is part of the task of pursuing moral ends. On this point see Christine Korsgaard, "Morality as Freedom", in her *Creating the Kingdom of Ends*, Cambridge University Press, Cambridge, 1996, pp. 177–178.

[22] See Allen Wood, "The Final Form of Kant's Practical Philosophy"; Paul Guyer, "Comments: Justice and Morality", and Thomas Pogge, "Is Kant's Rechtslehre Comprehensive?", in *Kant's Metaphysics of Morals*, in Nelson Potter and Mark Timmons (eds.), *The Southern Journal of Philosophy. Spindel Conference 1997*. Supplement (1997), pp. 161–188. Also see Otfried Höffe, "Recht und Moral: ein kantischer Problemaufriss", *Neue Hefte für Philosophie*, 17, 1979, pp. 1–36, and "Kant's Principle of Justice as a Categorical Imperative of Law", in *Kant's Practical Philosophy Reconsidered*, edited by Yirmiyahu Yovel, Kluwer Academic, Dordrecht, 1989.

ples of personal morality (ethics) rests upon the agreement of everyone (either possible or actual), we will have made a mistake. Agreement plays a role in the justification of rights and their correlative duties of justice, but it plays no role whatsoever in the justification of duties of virtue. This second class of duties is derived from a version of the formula of humanity, and the justification of humanity as the fundamental end of ethics rests on its being the only unconditional and independently existing end. Those who trace back to Kant the claim that agreement plays a justificatory role in personal morality have transferred a feature of Kantian justice to ethics. This confusion is motivated, I believe, by the widespread and mistaken assumption that Kantian morality is exclusively other-regarding, concerned only with universal justice and rights. But, as we have seen, only justice is exclusively other-regarding, according to Kant; ethics or virtue is primarily self-regarding.

In the following section, I will briefly present Kant's categorical imperative in its universal law formulation followed by the agreement test interpretation of this principle. I will then argue in subsequent sections that Rawls and Habermas interpret the universalizability test as an agreement test because they assume that Kantian morality is exclusively other-regarding. They assume that moral questions are about how we ought to relate to each other because they understand personal morality on analogy with political justice.

2. *The categorical imperative and the agreement test interpretation*

In the *Groundwork*, Kant claims that universalizability is the criterion for determining whether a maxim is morally permissible, forbidden, or required (G 4:421/31). Though the formulas of humanity and autonomy (the other two formulations of the categorical imperative) also provide tests for checking the morality of our maxims, in the *Groundwork* Kant privileges the formula of universal law as a guide for moral appraisal.[23] This accounts, I believe, for the pervasive tendency among friends and foes of Kantian morality to focus almost exclusively on the universalizability test.[24] Yet, as we saw in the previous section, in his derivation of ethical duties in *The Doctrine of Virtue* Kant favors a version of the formula of humanity.[25] This other formula requires that we act on maxims that have a certain *content*, namely, the end of humanity. It tells us always to treat humanity as an end. Thus, regardless of whether it is correct to interpret the universalizability test as an agreement test, there is the prior question whether a purely *formal* test (one which does not make any reference to the *content* of the maxims of duty) should be our guide for the identification of ethical duties. Kant's account in *The Doctrine of Virtue* indicates that it should not. For the purposes of this paper, however, we need to consider the

[23] In the *Groundwork* Kant writes that "one does better always to proceed in moral *appraisal* by the strict method and put at its basis the universal formula of the categorical imperative: *act in accordance with a maxim that can at the same time make itself a universal law*". 4:436–7/44.

[24] For a strong and impatient condemnation of this tendency, see Allen Wood, "The Final Form of Kant's Practical Philosophy", in *Kant's Metaphysics of Morals*, edited by Nelson Potter and Mark Timmons.

[25] Introduction to *The Doctrine of Virtue*, 6:395/198.

interpretation of the universalizability test as an agreement test.

The formula of universal law of the categorical imperative commands us to act only according "with that maxim through which you can at the same time will that it become a universal law" (*G* 4:421/31). The agent who doubts the morality of his maxim is supposed to apply this test as follows. After having formulated the maxim of her action, she has to conceive of a world in which the maxim holds as a universal law, and then check whether there is a contradiction between the universalized maxim and her own maxim. If there is no contradiction the maxim is permissible. But if there is a contradiction the maxim is forbidden and its opposite is a duty. Hence, universalizable maxims are permissible, non-universalizable maxims are forbidden, and their opposites are moral duties. However, if there is no contradiction, the agent also has then to chech whether she could will the world with the universalized maxim. She could not will such a world if the universalization of the maxim would make it imposible for her to obtain certain things that she necessarily wills, such as the help of other people whenever she might need it. Kant calls the first kind of contradiction a "contradiction in conception", and the second one "contradiction in the will".

One of Kant's own examples in the *Groundwork* of a contradiction in conception goes as follows.[26] A man in

[26] I have picked this example because I believe that the application of the universalizability test works better here than in the other three examples which Kant gives in the second section of the *Groundwork*. There is a good deal of disagreement regarding the kind of contradiction that Kant claims we should try to look for (whether logical, teleological, or practical). For my own illustrative purposes here, I have followed the practical interpretation on to which the contradiction is between the agent's purpose in the maxim (or the conditions for the realization of all sorts of purposes) and the possibility of realizing the purpose in question (or having access to the conditions for the

need of money considers borrowing some though he knows that he will not be able to repay it. He also knows that, in order to get the money, he will have to promise that he will repay it. Kant formulates the maxim of this action as "when I believe myself to be in need of money I shall borrow money and promise to repay it, even though I know that this will never happen" (*G* 4:422/32). In the second step of the test, the agent conceives of a world in which the maxim holds as a universal law. According to Kant's example, a world in which the above maxim holds as a universal law would be a world in which nobody believed this kind of promise because everybody acts on the maxim of making money available to themselves whenever they need it by promising to repay it to a potential lender though they know that this will never happen. Kant argues that in such a world, it turns out to be impossible to act on the maxim under consideration: it is impossible to realize the purpose of obtaining money by making a false promise. In such a world, there is a contradiction between the maxim and its universalization. Therefore, the maxim is impermissible and its opposite is a moral duty.

According to the interpretation of the universalizability test as an agreement test, the command of acting only according to maxims that I can at the same time will that they become universal laws is just the command of acting only on maxims that could be agreed upon by all agents. On this interpretation, when we ask whether a maxim can be universalized, what we are asking is whether everyone could agree to it. So, in the example above, when you universalize the maxim you ask whether everyone could

realization of all sorts of purposes) in a world in which his own maxim has become a universal law publicly known. For an account of the different interpretations of the contradiction as well as a defense of the practical interpretation, see Christine Korsgaard, "Kant's Formula of Universal Law", in her *Creating the Kingdom of Ends*.

agree to live in a world in which people acted on this maxim. The obvious question is how to determine what people could agree to. Those who favor this reading take either of two options: either they retain the contradiction test or they claim that in order to determine what people could agree to we need to engage in an actual process of deliberation or a dialogue with others.[27] Rawls takes the first option and Habermas the second one. According to the first option, we do not need to ask other people what they can agree to. The contradiction test does this for us. According to the second option, for the test to justify moral duties we must carry out an actual process of deliberation with others. In the two remaining sections, I will consider these two options as they have been developed by Rawls and Habermas. I will consider Rawls's interpretation of the categorical imperative first.

3. *The original position and the categorical imperative procedure*

Rawls retains the contradiction test in his interpretation of the universalizability test, though in a modified way. The input of the procedure is a rational maxim, which the agent generalizes as a publicly known law of nature. Next, the agent has to ask herself whether she can intend to act on her maxim in the world with the new law. If she could, she then has to consider whether she can will the world with the new law. If the agent gives a negative answer at either step, she must reject the maxim as impermissible.

[27] Andrews Reath favors the second reading in "Self-Legislation and Duties to Oneself", in Kant's *Metaphysics of Morals*, Nelson Potter and Mark Timmons (eds.). A third alternative would be a thought experiment that differs from the contradiction test such as Scanlon's test of reasonable rejection (*What We Owe to Each Other*, ch. 4) but since he does not present his view as an interpretation of the contradiction test I will leave it aside.

The contradiction here is practical and would hold either, on the one hand, between intending to act on the maxim and certain features of the world in which the maxim holds as a universal law or, on the other hand, in the agent's own will. I will turn to the details of Rawls's procedure further below. I will first make some general remarks about his interpretation of Kant's categorical imperative.

Rawls presents the categorical imperative as a procedure for the justification of principles of right and of justice.[28] According to his classification of moral concepts in *A Theory of Justice*, principles of right can be principles for either individuals or institutions (*TJ* 109). Principles of justice are principles of right for institutions. Rawls's principles of right for individuals comprise the domain of personal morality.

The main difficulty with Rawls's interpretation of the categorical imperative is that he relies primarily on the *Groundwork* and the *Critique of Practical Reason*, but in these works Kant had not yet introduced his distinction between ethics and justice. When Kant introduces this distinction in *The Metaphysics of Morals*, it becomes clear that ethical duties cannot be arrived at in the same way as duties of justice. When Rawls presents the categorical imperative as a procedure for the justification of principles of right and of justice, he overlooks Kant's distinction between ethics and justice. I believe that the reason for this is that the categorical imperative is the supreme principle of these two moral domains. From Rawls's point of view, this means that Kant's theory of justice is comprehensive (e.g., part of a comprehensive moral doctrine). Very roughly, a moral doctrine is comprehensive when it

[28] John Rawls, "Themes in Kant's Moral Philosophy", in *Kant's Transcendental Deductions*, in Eckart Förster (ed.), Stanford University Press, Stanford, 1989, p. 97.

contains and organizes moral principles and values for all (or nearly all) areas of life, such as social justice, basic rights, ideals of virtue and character, and so on (*PL* 13–14). The contrast here is with a *political* conception of justice, such as Rawls's own, which contains principles of political justice for what he calls the "basic structure of society" only —that is, society's main social, political, and economic institutions. Since the categorical imperative is the supreme principle of a comprehensive moral doctrine, Rawls claims that this principle is intended to work for the justification of all moral principles. By contrast, the original position (which is the point of view from which to justify principles of political justice) is not intended in this way —that is, as a guide for moral judgment in general. Rawls explicitly designs the original position for the justification of principles of political justice only.[29] He has repeatedly emphasized this point and has warned his readers about using the original position as a guide for moral judgment as such.[30]

Rawls seems to think that because Kant's moral theory is comprehensive, we can arrive at all moral principles by following the same procedure. In other words, Rawls seems to assume that because the categorical imperative is the supreme principle of justice and of ethics we can arrive at duties of justice and of virtue by submitting our max-

[29] I am following here Rawls's considered views in *Political Liberalism*. In *A Theory of Justice*, before he called his conception of justice "political", he suggested that the original position could work for the justification of principles of right for individuals as well as for principles of justice for institutions (*TJ* 111).

[30] Rawls cannot present this point of view as a guide for moral judgment in general without making his own conception of justice part of a comprehensive moral doctrine. Part of the claim that justice as fairness is a political conception of justice depends upon not presenting the original position as a procedure for the justification of all kinds of moral principles.

ims to the universalizability test. But this conclusion does not follow. As I mentioned before, in *The Metaphysics of Morals* Kant derives the duties of right and of virtue from two principles that are subsidiary to the categorical imperative (the universal law of right and the supreme principle of virtue).[31] But why should this matter? That is, why should it matter that Rawls assumes that principles of personal morality and of justice can both be derived from the same categorical imperative procedure? I want now to argue that Rawls makes this assumption because he also assumes an other-regarding view of morality, which, I will argue, reflects an understanding of personal morality on analogy with political justice. The analogy in question is the following. Principles of justice, even when they are principles for institutions as they are for Rawls, tell us what we owe to each other as citizens, to use Scanlon's apt expression; they specify our rights and duties as citizens.[32] Now, Rawls appears to suppose that principles of right for individuals play the analogous role of governing human interaction, though they do so at a different level: they tell us how we ought to relate to each other more generally, not only as citizens. According to this analogy, principles of personal morality and of political justice differ primarily in the scope of their application. But as we saw in section one, Kant's own principles of personal morality (or ethics) play the primary role of telling us what kind of character we ought to have; on his view, only principles of justice are exclusively

[31] The distinction between ethics and justice, however, does not make Kant's moral theory any less comprehensive. So, Rawls could claim that Kant's theory of justice is part of a comprehensive moral doctrine and, at the same time, endorse Kant's distinction between ethics and justice.

[32] Principles of political justice, according to Rawls, govern the interaction among citizens only indirectly. These principles directly govern the institutions of the basic structure of society, and indirectly the conduct of, and relations among, citizens.

other-regarding. That Rawls assumes this analogy is clear from the fact that, as I will argue below, his rendering of the categorical imperative procedure mirrors his own account of the original position, which he expressly designed for the justification of principles of political justice. This also indicates that he interprets the universalizability test as an agreement test.

As is well known, the original position is a procedure for reaching agreement on principles of political justice. Rawls can interpret the universalizability test as an agreement test precisely because he assumes that, by analogy with principles of political justice, principles of personal morality play the role of telling us how we ought to relate to each other. If we suppose that the central question of personal morality is, as Scanlon suggests, "What do we owe to each other?", it might seem plausible to think that an agreement on principles can provide an answer.[33] But if we agree with Kant that principles of personal morality play the role of telling us what kind of character we ought to have and are therefore self-regarding, it is unclear how an agreement on principles might help. In other words, the interpretation of the universalizability test as an agreement test makes sense under the assumption that all moral principles are other-regarding. Of course, one might argue that personal morality is exclusively other-regarding and claim that the justification of its principles rests upon the agreement of everyone. But as far as our reading of Kant is concerned, we cannot interpret the universalizability test as an agreement test without wrongly attributing to him an other-regarding view of morality.

[33] Scanlon's morality of what we owe to each other (*What We Owe to Each Other*, ch. 4) provides an illustration of this point: as he emphasizes, justifiability to others makes sense as a test of wrongness when we want to know what we owe to each other but not when we want to know whether there is anything that we owe to ourselves.

In the remainder of this section I will argue for the claim that Rawls's rendering of the categorical imperative procedure mirrors his own account of the original position.

The original position

The original position is a procedure for reaching agreement on principles of political justice for a modern democracy.[34] These principles provide a solution to the question of justice, namely, the question how to assign individual rights and duties as well as how to distribute the advantages of social cooperation (*TJ* 4). Rawls proposes to regard the principles of justice as the focus of a fair agreement among citizens, and he argues that the original position provides a point of view from which to reach that agreement. He designs this point of view by organizing and modeling in it the features that characterize the question of justice in a modern democratic society. A central and universal feature is that the question of justice arises when individuals who participate in the same system of social cooperation make conflicting claims regarding their rights and duties as well as about the distribution of the advantages of social cooperation. Other equally central features that are specific to modern democratic societies are the following: the question of justice arises among persons who regard themselves and each other as free and equal citizens; they have different moral, religious, and philosophical comprehensive doctrines as well as different conceptions of the good (*PL* Lecture I);[35] they also regard their society as a fair system of social cooperation over time that works for

[34] Though my account of the original position is primarily based on Rawls's *A Theory of Justice*, I have also incorporated some elements which appeared only in later writings.

[35] Rawls, *A Theory of Justice*, Chapter III, "The Original Position". Here Rawls talks about a plurality of conceptions of the good. In *Political Liberalism* he focuses on the plurality of comprehensive

their reciprocal advantage. Furthermore, Rawls introduces the “liberal principle of political legitimacy”, which sets a condition that the principles of justice must meet: the principles must be acceptable in the eyes of each citizen (*PL* 137).

The only solution to the problem of justice that is compatible with the features of the question of justice and that also meets the condition set by the liberal principle of political legitimacy is a fair agreement among citizens on principles of political justice for their society. The combination of these elements (the idea of free and equal citizenship, the conception of society, the plurality of comprehensive doctrines and conceptions of the good, and the condition set by the liberal principle of political legitimacy) leads to the view that only an agreement among citizens can justify principles of political justice for a modern democracy.

Rawls models the features of the question of justice by making the original position a case of pure procedural justice: the parties to the agreement deliberate under certain constraints such that whatever the outcome of their deliberation, the principles agreed upon are fair. The parties are rational deliberators who are concerned with maximizing a share of primary goods for the citizens they represent. Primary goods are those which it is rational for any person in society to want because they are necessary all-purpose means for being a fully cooperating member of society —that is, a citizen. The rational deliberation of the parties models citizens’ interest in the realization of their own ends. Since the primary goods are all-purpose means for the realization of their interests and ends, citizens have a rational interest in securing for themselves a share of these goods.

religious, philosophical, and moral doctrines, which include conceptions of the good.

Although the parties are only concerned with the advantage of those they represent, their deliberation is subject to limits on information that guarantee the fairness of their agreement. The information they lack would allow them to tailor principles in order to suit the interests of the citizens they represent. The parties do not know which comprehensive doctrine and conception of the good citizens affirm, nor do the parties know citizens' social position, natural talents and abilities, race, gender, and so forth. This condition, which Rawls calls "the veil of ignorance", secures that no matter who those they represent turn out to be, the parties choose principles that guarantee a fair share of primary goods for each citizen. The veil of ignorance forces them to deliberate from a general point of view; it models the conception of society as a *fair* system of social cooperation as well as the condition of acceptability set by the liberal principle of political legitimacy.

The important point here is that the design of the original position comes after having properly characterized the question of justice. Since Rawls designs this point of view specifically to provide a solution to the question of justice, any attempt to extend it to the solution of other moral questions must show that they share some central feature with Rawls's question of justice. An agreement on principles turns out to be an adequate solution to the problem of justice given the kind of problem that this is. Nevertheless, Rawls himself uses some features of the original position in his interpretation of the categorical imperative. This suggests that he believes that non-political moral questions share some central features with the political question of justice he addresses. Let us now turn to his interpretation of the categorical imperative.

The categorical imperative procedure

Rawls's account of the categorical imperative procedure, like the original position, has two parts: the agent's rational deliberation and the limits on information that constrain it. The content of the deliberation is a sincere rational maxim, one that expresses the interests of the agent.[36] The maxim contains the intended action, the reason why the agent intends it (the purpose), and the relevant circumstances. The point of the procedure is to check whether a rational maxim is also reasonable, that is, whether it could be accepted from the standpoint of everyone as a law for a common social world. The agent generalizes the maxim: she transforms it into a publicly law of human nature. By "universalizing" a maxim Rawls understands something like proposing a new public law to the social world in which we live.[37] The agent asks herself whether she can intend to act on her maxim in this "perturbed" social world (the world with the new law); and even if she can, whether she can will the social world with the new law. If the agent cannot intend to act on her maxim in the "perturbed" social world, she ought to reject the maxim as impermissible. This rejection is in order when the agent's purpose would be thwarted by the new law. But, as I mentioned, even if the agent can intend to act on her maxim, she still has to ask herself whether she can will the social world with the new law. In order to give content to the idea that there may be some laws that the agent cannot will, Rawls introduces the idea of "true human needs."[38] The role of these needs is similar to the role of primary goods in the original position. The parties in the original position cannot agree

[36] Rawls, "Themes in Kant's Moral Philosophy", pp. 82ff.

[37] Rawls, "Themes in Kant's Moral Philosophy", p. 86.

[38] Rawls, "Themes in Kant's Moral Philosophy, p. 85.

to any principles that would not secure an acceptable share of primary goods (all purpose means) for the citizens they represent. Similarly, the agent cannot will any law that would deny her access to the necessary means for the satisfaction of her true human needs. So, if the new public law obstructs her ability to meet her true human needs, she ought to reject it, and its opposite is a duty. But if the new public law does not hinder her ability to meet her true needs, the law is at least permissible.[39] The outcome of the procedure are either permissible maxims (reasonable), or forbidden maxims (unreasonable), or maxims that contain required ways of acting (also reasonable).

Rawls introduces limits on information in order to rule out the possibility that the agent may tailor principles to suit her own purposes. Since these limits mirror the veil of ignorance in the original position, let us refer to them as "the veil of ignorance". This veil of ignorance excludes information about the particular features of persons, including herself, as well as the specific content of their and

[39] Rawls's interpretation follows Kant's distinction between contradictions in conception and contradictions in the will. A contradiction in conception arises when the agent cannot act on her own maxim in a world in which the maxim has become a universal law: the maxim cannot even be conceived as a universal law. A contradiction in the will arises when the universalization of her maxim does not make it impossible for the agent to act on it, but she cannot will the world of the universalized maxim. The reason that Kant gives is that in such a world the agent would be deprived of some means necessary for the realization of all sorts of purposes (*G* 4:423/33). These means are the other's help and cultivated talents and capacities. This suggests that human agents have some basic needs, the satisfaction of which they cannot renounce. Though Kant simply assumes that we need the help of others as well as cultivated talents and capacities for the realization of our own purposes, his application of the test leaves it open for the agent to deny that he has those needs. Rawls rules out this possibility by making the true human needs part of the procedure.

her final ends and desires. The agent deliberates as if she did not know her place in the social world.[40]

In the original position, the parties do not reject some principles of justice because the principles are unfair or unreasonable but because it would be *irrational* to accept them: in a society governed by those principles, the citizens they represent would not have a guaranteed access to a fair share of primary goods.[41] Similarly, the agent who applies the categorical imperative procedure does not reject an unreasonable maxim because it is unreasonable (because not everyone would accept it) but because it would be irrational for her to will the maxim and its universalization at the same time. The transformation of the maxim into a new public law for our social world would either make it impossible for the agent to intend to act on her maxim, or it would obstruct the satisfaction of her true needs. The irrationality in question here is practical: we contradict our own will when we will a maxim and its universalization at the same time if the new law would make it impossible for us to intend to act on the maxim. In such a situation we would not be able to achieve our purpose in the social world with the new public law. Similarly, it would be irrational to will a maxim and its universalization at the same time if the new law canceled the possibility of satisfying our true needs.[42]

[40] Rawls, "Themes in Kant's Moral Philosophy", p. 86.

[41] I am following here Rawls's distinction between being rational and being reasonable. According to his distinction, a person is rational when she takes the necessary means to advance her own ends, but she is reasonable when she is "willing to listen to and consider the reasons offered by others", that is, when she takes into account the other's views. See his "Themes in Kant's Moral Philosophy", pp. 87–88.

[42] Although Kant does not say that the reason for rejecting a given maxim is that it would be *irrational* for the agent to will the maxim and its universalization at the same time, it is clear that this is the relevant reason. In Kant's own account of the universalizability test,

Rawls's interpretation has the advantage of solving some difficulties with the universalizability test. He gives us a procedure that is not only clear and easy to follow but that also avoids leading to the wrong result. A major problem with Kant's own presentation of the universalizability test is that the agent who applies it gets different results depending on how she formulates her maxim. Rawls avoids this problem by giving a detailed account of how to formulate a maxim that limits significantly what can count as a well-formed maxim. Because she has placed herself behind a veil of ignorance, the agent cannot tailor principles that suit her own purposes. And the account of true human needs gives content to her deliberation. But the introduction of the veil of ignorance makes it clear that Rawls interprets the universalizability test as a procedure for reaching *agreement*. In the original position, the veil of ignorance plays two main roles: by limiting the information available to the parties, the veil of ignorance secures the *fairness* of the agreement and guarantees that the parties *unanimously* agree on the same principles of justice. These two roles are also present in the categorical imperative procedure. Rawls explicitly asserts the second one. He tells us that "The point is simply that all persons affected must apply that procedure in the same way both to accept and to reject the same maxims."[43] But why do we want unanimity? Presumably because the principles are meant to function as public laws for our common social world, and they can perform this function provided that everyone accepts them. This takes us to the first role. If the aim is

the agent does not reject a maxim because its universalization would be unfair or unacceptable to other people. Instead, the agent rejects a maxim when, if universalized, she could either not achieve the purpose in her own maxim or would be deprived of some means necessary for the realization of all sorts of purposes.

[43] Rawls, "Themes in Kant's Moral Philosophy", p. 90.

to reach an agreement on principles, it only makes sense to design the procedure so as to guarantee that it be fair.

An agreement among persons seems to be a plausible solution for the justification of principles of personal morality assuming that the central problem of personal morality is some sort of disagreement. This point is clearer if we consider the following. In Rawls's theory of justice, the problem in the circumstances of justice is that people make conflicting claims about what they are entitled to expect from each other: they disagree about their rights, duties, and distributive shares. The solution to this conflict is, according to Rawls, an agreement on principles that specify what they are entitled to expect from each other: the role of the principles is to govern their interaction as citizens by determining their rights, duties, and a scheme of distribution. In other words, Rawls's proposal is that if disagreement is the problem, an agreement might be the solution. Whether this is the right way of understanding questions of personal morality is far from clear, however.[44]

[44] It is not at all clear that all questions of personal morality are about how we ought to relate to each other, and that all principles of personal morality are principles of fairness. However correct it may be to understand justice *as* fairness (that is, to regard the principles of justice as the focus of a fair agreement), it does not seem right to understand *rightness as fairness*. The expression *justice as fairness* means that "the principles of justice are agreed to in an initial situation that is fair" (*TJ* 12). Accordingly, *rightness as fairness* means that the principles of right are also the object of a fair agreement (see *TJ* section 18). This way of understanding rightness might be adequate for the principle of keeping one's promises, say, for it does not seem implausible to claim that the parties in the original position would agree to it. However, why would the parties agree to the principle of not killing oneself, for instance? It seems rather odd to claim that what it is wrong with killing oneself is that the principle of not doing so is the object of a fair agreement. And, although we might also suppose that the parties would agree to the principle of not killing others, it also seems odd to say that what it is wrong with killing others is that the principle of not doing so is the object of a fair agreement.

But at least one thing is clear: agreement might seem to play a justificatory role assuming that questions of personal morality are about how we ought to relate to each other. If we need to determine how we ought to relate to other people, it might be plausible to think that we can do this through an agreement with others. If the question is about which public and shared laws should govern human interaction, it seems plausible to say that such laws are those to which everyone could agree. In other words, the agreement test interpretation of the categorical imperative presupposes an other-regarding morality. But according to Kant, as we have seen, the central question of personal morality is *not* how we ought to relate to each other. Instead, the central question in this domain is about the character that one ought to have, that is, about the kind of person that one ought to be. If the problem of personal morality is about the ends that I ought to have, it is unclear how an agreement with others could provide a solution. Let us now consider Habermas's interpretation of the categorical imperative.

4. *Practical discourses*

Habermas's reformulation of the categorical imperative as a procedure for reaching agreement stems a different motivation from Rawls's. Whereas Rawls's account can be seen as an extension of the Contractarian approach to the justification of principles of personal morality, Habermas's reformulation is motivated by a rejection of the contradiction test. He rejects the strategy of checking the morality of maxims through thought experiments ("DE" 65–67).[45] He argues that a solitary agent can never be sure whether

[45] Habermas, "Remarks on Discourse Ethics", in his *Justification and Application. Remarks on Discourse Ethics*, trans. by Ciaran P. Cronin, MIT Press, Cambridge, Mass., p. 64.

she has applied the test in a way that suits her own interests. Since he believes that a mere thought experiment can lead to the wrong result, he argues that in order to determine whether a maxim can be universalized, we need to engage in an actual deliberation or dialogue with others. By taking their point of view into account, he argues, we are forced to take an impartial perspective. This deliberation, which he calls "practical discourse", is subject to ideal conditions that guarantee the rationality of the agreement. Thus, he presents Discourse Ethics as an intersubjective reformulation of the categorical imperative. According to this reformulation, the validity of a moral norm can only be established in an actual argumentation with others or practical discourse.[46]

Habermas puts his objection to Kant as follows:

> It is not a foregone conclusion that maxims generalizable from my point of view must also be acknowledged to be moral obligations from the perspective of others, let alone of all others. Kant could disregard this fact because [...] he

[46] At the beginning of this paper I mentioned that in the 1994 Postscript to the English version of *Between Facts and Norms*, Habermas makes a distinction between a democratic and a moral principle. He claims that these two principles are specifications of the discourse principle (the principle of discourse ethics). This distinction does not undermine my claim that on Habermas's view all moral norms are arrived at by following the same procedure, though it does call for a minor qualification in my claim. According to this distinction, the moral principle applies to all types of moral norms, whereas the legal principle applies to legal norms only. Habermas also tells us that these two principles specify two kinds of practical discourses: in a purely moral discourse "*only* moral reasons are decisive" (p. 460), whereas in a legal discourse non-moral reasons might be taken into account. These considerations indicate that on Habermas's view different kinds of moral norms might be arrived at by following two different kinds of procedures. However, my central claim still holds, namely that according to Habermas the justification of all moral norms rests upon the agreement of everyone in a practical discourse (which might be either legal or purely moral).

> assumed that all subjects in the Kingdom of Ends share the *same* conception of themselves and of the world. Once we abandon [this premise] it becomes imperative to submit all norms [. . .] to a *public*, discursive, generalization test that necessitates reciprocal perspective taking.[47]

According to this objection, Kant secures the intersubjective validity of moral norms by assuming that the agent regards himself and all others as purely rational beings who are members of a Kingdom of Ends.[48] Purely rational beings do not have different interests and needs. It is worth noticing at this point that in Rawls's version of the categorical imperative procedure, in order for the test to work, we need something like a veil of ignorance that brackets particular interests and needs. Habermas's objection seems to be that if we drop this assumption, namely, that individuals share the same conception of themselves and of the world, it is a mistake to suppose that an individual alone could determine what can be a universal law. Thus, his objection is not really that a thought experiment can lead to the wrong result. Instead, the objection appears to be that the thought experiment works under an assumption which he thinks we should reject. But why should we reject the assumption that in order to determine what morality requires we should regard ourselves and others as ends in themselves? Habermas takes it to be obvious that this assumption is implausible. I believe that the reason has to do with his own conception of the role of moral norms and of the nature of moral questions. According to this conception, moral questions are about how to solve conflicts of interests among individuals, and the role of moral

[47] Habermas, "Remarks on Discourse Ethics", in his *Justification and Application*, p. 64 (emphasis in the original).

[48] Habermas, "Remarks on Discourse Ethics", in his *Justification and Application*, p. 51.

norms is to tell us how to solve these conflicts. So, if the question is how to solve a conflict of interests, in order to do so, we cannot just bracket the particular interests that people have. If we did that, instead of offering a solution, we would just be dissolving the problem. Further below, I will explain and challenge this conception of moral norms and of the nature of moral questions.

At first glance, it looks like Habermas's objection to Kant's test is only epistemological: we cannot have an insight into what counts as morally permissible, required, or forbidden except through a dialogue with others. In order to gain an insight into what morality requires we cannot rely on our own point of view; we need the input of others. Habermas argues that taking into account the perspective that others have on the same question and presenting my point of view to their criticism plays the role of enlightening my own judgment. He tells us that real argument "makes moral insight possible" ("DE" 57). But, as I will argue later, Habermas's point is not only epistemological —that is, his objection is not just that we cannot determine what's morally required in a thought experiment. His point is also metaphysical, since he also argues that only the actual agreement of all is *constitutive* of moral norms. More precisely, his claim is also that the justification of a moral norm cannot rest on what a *single individual* can will. On his view, even if you could, by making certain ideal assumptions, determine what everyone could will, the justification of moral norms can only rest upon a *collective will*.[49]

[49] Thomas McCarthy puts the point as follows: "Rather than ascribing as valid to all others any maxim that I can will to be a universal law, I must submit my maxim to all others for purposes of discursively testing its claim to universality. The emphasis shifts from what *each can will* without contradiction to be a general law, to what *all can will* in agreement to be a universal norm", *The Critical Theo-*

It is important to notice the following ambiguity. Habermas's claim is either that Kant's test already is an agreement test or that it should be conceived of in this way (though Kant himself did not). In the quotation above, Habermas makes it sound as if Kant is concerned with agreement which he (Kant) secures by assuming that all members of the Kingdom of Ends share the same conception of themselves and of their world. In other passages, however, Habermas seems to be saying that Kant's moral test is not at all about agreement since the test is about what a single individual can will. The source of the ambiguity is that an agreement test can be carried out by a single person in a thought experiment (as in Rawls's original position). However, this assumes that the agreement can be *hypothetical*, which, as the quotation above indicates, Habermas also wants to reject in favor of an actual agreement. He thinks that we cannot determine what everyone can will in a thought experiment; only the *actual agreement of all* can justify a moral norm. The question for us now is why he thinks that an agreement among persons justifies moral norms. To answer this question we need to turn now to Habermas's conception of moral questions and of the role of moral norms.

On his view, moral questions are about how to solve in a legitimate manner conflicts that arise in human interaction.[50] Accordingly, moral norms play the role of solv-

ry of Jürgen Habermas, MIT Press, Cambridge, Mass., 1978, p. 326 (emphasis added). Habermas quotes this passage approvingly in "Discourse Ethics. Notes on a Program of Philosophical Justification", in his *Moral Consciousness and Communicative Action*, p. 67.

[50] Habermas endorses a distinction between morality and ethics that is normally construed as a distinction between the domain of our obligations to others, on the one hand, and the domain of values which guide the good life or the life worth living, on the other. Among those who subscribe to this distinction are some philosophers who, one might think, disagree on almost everything else, such as Charles

ing these conflicts and of coordinating human interaction ("DE" 66). He writes:

> Moral judgment [. . .] serves only to clarify legitimate behavioral expectations in response to interpersonal conflicts resulting from the disruption of our orderly coexistence by conflicts of interests.[51]

On Habermas's view, moral norms tell us what expectations of behavior are legitimate; they tell us how we ought to relate to other people, so that when conflicts arise among individuals about what they owe to each other, moral norms provide a standard for the adjudication of conflicting claims. The function or role that he attributes to moral norms comes out clearly in his views on what a moral question is about, how it arises, and how to solve it. Let us consider these points briefly.

In his theory of social action, Habermas distinguishes between strategic and communicative action. An action is *strategic* when the agent aims at the realization of his own purposes and proceeds by calculating the actions of other people and by influencing their decisions.[52] We might say that the agent adopts the third person perspective of an observer who makes calculations about the actions of others in order to achieve his own ends. The use of language in this type of action is strategic because the agent uses it to get information from others and to influence their decisions in a way that is conducive to the realization of his

Taylor, *Sources of the Self. The Making of the Modern Identity*, 14; Bernard Williams, *Ethics and the Limits of Philosophy*, chs. 1 and 10; and Jürgen Habermas, "On the Employments of Practical Reason", in his *Justification and Application: Remarks on Discourse Ethics.*

[51] Habermas, "On the Employments of Practical Reason", in his *Justification and Application*, p. 9.

[52] Habermas, *The Theory of Communicative Action*, volume 1, p. 85

own ends. By contrast, an action is *communicative* when at least two agents aim at the coordination of their action by way of agreement.[53] The agents involved use language in a communicative way because they do so in order to reach a shared or common understanding of the situation. They do not manipulate or influence the behavior of each other, but coordinate their plans of action on the basis of a shared view of the circumstances. We might say that the agents alternate in taking the first and second person perspectives. Both of them take what Habermas calls the perspective of the "participant". They may not be pursuing the same end, but each knows what the other aims at, and the way in which they pursue their respective ends is known and accepted by both of them.

The distinction between strategic and communicative action turns on the way in which people relate to each other: in the former, at least one agent takes the third person perspective of an observer who regards other people as the object of his calculations and manipulations; in the latter, at least two agents take the first and second person perspective of participants who interact by giving each other reasons. Let us now consider communicative action more closely. In order to act together, Habermas claims the agents must presuppose a background agreement on legitimate expectations of behavior, knowledge of causal connections, and trust in each other's sincerity. The participants can coordinate their plans of action because they assume acceptance of the same norms that govern interaction, they agree on a description of the relevant states of affairs, and each assumes that the other is sincere. Communicative action goes awry when this background consensus is broken. This happens when either of the participants challenges any

[53] Habermas, *The Theory of Communicative Action*, volume 1, p. 86.

of the assumptions that make communicative action possible: the validity of a norm of behavior, or the description of the relevant causal connections, or the sincerity of the other. At this point, the participants may decide not to act together or they may try to restore their agreement. It is the second possibility which Habermas is interested in, when, as he puts it, communicative action becomes “reflexive”. This means that now the agents engage in a discourse or argumentation in order to settle the conflict and restore their agreement. When the conflict is about what they are entitled to expect from each other, the agents engage in a practical discourse in order to determine the validity of the moral norm that has been either proposed or called into question. Thus, a practical discourse serves the purpose of restoring a background agreement on legitimate expectations of behavior that has been broken. Without such an agreement the agents could not coordinate their plans of action.

This is the way in which Habermas conceives of moral questions: they arise against a background of shared expectations of behavior when someone challenges the validity of an accepted norm or claims validity for a new norm. On his view, a moral question arises when someone breaks with her action, or explicitly challenges, a shared understanding about the way in which we ought to relate to each other. Since the problem is the disruption of a normative consensus, the solution appears to be the restoration of consensus. The aim of moral deliberation, on his view, is to restore agreement on legitimate expectations of behavior. Such agreement is necessary for the coordination of human interaction because if individuals disagree on what they owe to each other, their interaction will be ridden with conflict and they will not be able to act together.

It is worth noticing at this point that norms which work as a sort of mechanism for coordinating human interaction come closer to the legal norms enforced by the political authority than to the norms of personal morality. Traditionally, moral philosophers have thought that a characteristic of moral action is that it is done from a moral motive. But norms whose only function is to help us get along with each other need not make any demand on our motives. And Habermas thinks that indeed they do not. How is it, one might ask, that the issue of moral motivation, which has been of central concern in modern moral philosophy, has completely dropped out from Habermas's moral theory? The answer is that, as I have already mentioned, he has taken legal norms as the model for all moral norms. In other words, in Habermas's other-regarding morality, personal morality is understood on analogy with political justice.

It should be clear by now that Habermas's view of agreement as justificatory of moral norms presupposes an other-regarding view of morality. On his view, moral questions arise when individuals make conflicting claims regarding what they owe to each other. Since disagreement is the problem, agreement appears to be the solution. That is, an agreement on what individuals are entitled to expect from each other seems to be the solution to the conflict in question. There are difficulties with the attribution of a justificatory role to agreement, however: if we assume that the role of moral norms is to tell us which reciprocal expectations of behavior are legitimate, it does not follow that the justification of such norms must rest upon the agreement of all. One might claim, say, that what justifies moral norms is their correspondence to moral facts. But even if we say this, it is at least clear why Habermas thinks that agreement has something to do with the justification of moral norms: in order to coordinate their actions individuals need a shared conception of how they ought to relate to each other. This shared conception may be provided by norms which correspond to moral facts. But for coordination to be possible, individuals must agree that correspondence to moral facts is what justifies moral norms and also agree upon some way of determining this correspondence. In other words, correspondence to an independent order of moral facts could play a justificatory role only if individuals accept that they ought to govern their conduct by norms that exhibit such correspondence.

This is, of course, not Habermas's view since he claims that agreement by itself justifies moral norms. We need not worry here whether he successfully argues for the justificatory role of agreement or not. The main point for us is that the claim according to which agreement justifies moral norms presupposes the view that morality is about the coordination of human interaction. If we challenge this presupposition, the role of agreement may become implausible.

I want now to go back to my earlier claim that Habermas's objection to Kant's universalizability test is metaphysical and not only epistemological. These two ways of taking Habermas's objection turn on an ambiguity in his "intersubjective reformulation" of the categorical imperative. By an "intersubjective reformulation" he may mean either of two things: either he means to replace the thought experiment in the universalizability test with a dialogue with others; or he means that an individual alone cannot legislate universal laws; that only the general will can. Both alternatives are implicit in his writings, but they contain two quite different objections to Kant. According to the first alternative, Habermas's objection is that a solitary agent may go wrong in determining which maxims can be laws. The alleged need of a dialogue with others appears to be motivated by this worry: deliberation with others keeps the agent from applying the test in a way that suits his own interests.[54] On this reading, Habermas's objection to Kant is epistemological: an individual alone cannot discern what can count as a universal law. There is no doubt that this is part of Habermas's complaint. But this objection leaves intact Kant's view that the legislator of universal laws is the individual will. On this reading, Habermas's intersubjective reformulation only replaces a thought experiment by a dialogue with others.

According to the second alternative, Habermas's objection is that the legislator of universal laws cannot be an individual will. On this view, universal laws must be legislated by the general will. This objection to Kant is metaphysical, for the point at issue now is about the source of moral authority. Against Kant, Habermas claims that only the general will can be this source. Discourses do not just play the role of reassuring us that we have deliberated correctly: they are constitutive of the validity of moral norms. The problem is not just that the individual might go wrong in the application of the universalizability test. The problem is that even if he gets the right result, his individual will cannot be the source of authority of moral norms.[55] This

[54] Albrecht Wellmer also considers this possibility in "Ethics and Dialogue: Elements of Moral Judgment in Kant and Discourse Ethics" in his *The Persistence of Modernity*, trans. by David Midgley, MIT Press, Cambridge, Mass., 1993, pp. 142–143.

[55] This is the core of Habermas's objection to Rawls's original position. In the debate with Rawls, it seems sometimes that Habermas is objecting that the problem with the original position is epistemological: that we cannot, in a thought experiment, have insight into the principles of justice for a liberal society. There is no doubt that this is part of the objection but there is more to it than that. Habermas is also saying that principles of political justice can only be justified on the basis of an actual rational agreement among citizens because only a general will can be the source of authority of these principles. Even if by using the original position as a device of representation an individual arrives at the principles that we have good reasons to suppose would be agreed upon by all citizens in a rational dialogue,

second version of Habermas's objection expresses more fully what he means by an intersubjective reformulation of the categorical imperative.

Let me propose the following reconstruction of Habermas's objection to Kant. The categorical imperative tells you to act on that maxim that you can will as a universal law. Universal laws can be willed by everyone, but Kant does not say that what makes a maxim into a law is that everyone can will it. The first step in Habermas's intersubjective reformulation of the categorical imperative is the claim that the *agreement of all* is constitutive of moral laws. In other words, the first step is to correct Kant's understanding of universal laws: whereas Kant claims that an individual will legislates universal laws, Habermas claims that universal laws rest on a common will. This leaves it open whether the agreement is hypothetical or actual: if hypothetical, it is possible to say that an individual alone can establish the content of the common will in a thought experiment (such as the contradiction test). Habermas's second step is his claim that the content of the general will is the outcome of a practical discourse. That is, on Habermas's view, only the *actual* agreement of all justifies moral norms. The crucial step in his interpretation of the universalizability test as an agreement test is the first one. The second step already presupposes that the justification of moral principles rests upon what everyone can will. Nevertheless, the debate around Habermas's intersubjective reformulation has focused on the second step. Defenders and critics have debated whether, in order to determine what everyone can will, we need to engage in a dialogue with others, or whether a thought experiment, such as the original position, will do. In other words, the debate has centered

individual inside cannot ground the legitimacy of principles of political justice.

on the issue whether the agreement must be actual or hypothetical, but nobody asks whether the agreement of all is what justifies moral principles in the first place. And as I have been arguing, the view that agreement justifies moral norms presupposes an other-regarding view of morality.

Defenders of other-regarding morality might, of course, want to resist Kant's views. But then the question that presses upon us is how to understand personal morality. As I have suggested, Kant offers us a view of personal morality according to which the principles of personal morality are answers for the individual facing the question about what kind of person he ought to be, or as I have also put it about the ends he ought to have. This is a question that is forced upon us by our own freedom. Free individuals must decide how to shape their own characters.

In this paper I have argued for two main claims. First, I have argued that the view according to which the justification of moral norms rests upon the agreement of everyone presupposes an other-regarding view of morality. And second, I have also argued that although this view on moral justification has been traced back to Kant, it is a mistake to attribute it to him because he rejects the presupposition that morality is exclusively other-regarding. I argued that both Rawls and Habermas interpret Kant's universalizability test as an agreement test. In developing both of my claims, I argued that this view of morality reflects an understanding of personal morality on analogy with political justice: I showed how Rawls's interpretation of the categorical imperative procedure mirrors his own account of the original position, and I argued that Habermas takes the norms of the legal system as the model for all moral norms.

Rawls himself has taught us that before using an agreement test for the justification of moral principles we must first carefully describe the central questions in this domain. As we saw, he turns to the design of a procedure for the justification of principles of political justice only after having identified the central aspects of the question of justice. This approach has proven very fruitful in his theory of justice, and I think that we should follow him on this: in order to address questions of personal morality we must begin by carefully describing what they are about before proposing any criterion for the justification of principles in this domain.

Recibido: 18 de enero de 2000

RESUMEN

De acuerdo con una postura familiar, la justificación de los principios morales se basa en un acuerdo. También es familiar considerar a ésta como una postura neo-kantiana, ya que la prueba del acuerdo para la justificación de principios morales es, supuestamente, una interpretación de la prueba de universalización que I. Kant propone en la *Fundamentación de la Metafísica de las Costumbres*. En este artículo rastreo esta interpretación en la obra de John Rawls y de Jürgen Habermas y sostengo que está basada en una confusión.

La prueba del acuerdo para la justificación de principios morales tiene sentido siempre y cuando afirmemos una concepción de la moralidad según la cual los deberes morales son siempre hacia los demás. Sin embargo, la atribución de esta concepción a Kant no toma en cuenta su distinción entre la ética (o moral personal) y la justicia (o moral política). Dicho de manera más concreta, la interpretación de la prueba de universalización como un procedimiento que exige el acuerdo introduce aspectos de la concepción kantiana de la justicia en su concepción de la ética. Los deberes de justicia son, efectivamente, hacia los demás, pero la ética, aunque comprende deberes hacia los demás, está centrada en la adquisición de un carácter moral.

El propósito del artículo no es criticar la apropiación que hacen Rawls y Habermas de tesis kantianas en sus teorías de justicia y de legitimación, sino criticar la extensión de sus teorías de justificación moral hacia el dominio de la ética o moral personal como parte de una postura kantiana. Esta extensión presupone que las preguntas morales son exclusivamente acerca de cómo gobernar las relaciones entre las personas.