*Consent and the Formula of Humanity*

Japa Pallikkathayil

Kant famously argued that one ought never to treat others merely as a means. It is unclear, however, what exactly treating someone merely as a means comes to. In this paper, I explore one prominent approach to explicating this idea. This approached, advanced by Christine Korsgaard and Onora O'Neill, suggests that one treats another merely as a means if one treats the other in a way the other could not possibly consent to being treated. Call this the 'possible consent interpretation'. In Section I, I argue that the possible consent interpretation involves attributing to Kant an implausible view, one that we have good evidence he did not espouse. In Section II, I argue that Korsgaard's attempt to address the implausibility of the view being attributed to Kant is inadequate. In Section III, I argue that when the motivation behind the possible consent interpretation is made clear, the view has implausible implications that have thus far gone unnoticed. These implications suggest that the possible consent interpretation articulates a view that is fundamentally misguided. Finally, I offer a suggestion regarding how we might go about trying to develop an alternate interpretation.

## *I. Interpreting Kant*

Kant's Formula of Humanity reads: "So act that you use humanity, whether in your own person or in the person of any other, always at the same time as an end, never merely as a means" (*G* 429).[1] By humanity, Kant here means rational nature, i.e. the capacity to set ends.[2] The prohibition on treating someone as mere means is thus directed at how one interacts with rational nature.

Some of Kant's interpreters, Allen Wood for example, suggest that the application of the Formula of Humanity in moral deliberation always requires combining it with contentious claims

about what constitutes proper respect for rational nature. There is no straightforward criterion for treating someone as a mere means that falls out of the Formula of Humanity itself.[3] Others, like O'Neill and Korsgaard, are more optimistic about the prospects for such a criterion.[4] O'Neill and Korsgaard's views are very similar. Since Korsgaard's view is a bit more developed, I will focus on it.

Drawing on Kant's treatment of the case of the lying promisor, Korsgaard suggests that "you treat someone as a mere means whenever you treat him in a way to which he could not possibly consent (G 430)."[5] The lying promisor does just this. If I ask you for a loan with no intention of repaying it, when you hand over the money you haven't agreed to participate in bringing about the realization of my true end, namely the permanent acquisition of your money. You think my end is something entirely different. So, agreement that your action be used to further my true end is impossible.

There is something very intuitive about this characterization of treating someone as a mere means. If your actions are put in service of an end that it is impossible for you to agree to further, there seems to be a clear sense in which your agency has been co-opted. But when we consider the concrete implications of the possible consent criterion, problems begin to emerge. Korsgaard claims:

> Kant's criterion most obviously rules out actions which depend on force, coercion or deception for their nature, for it is of the essence of such actions that they make it impossible for their victims to consent. If I am forced, I have no chance to consent. If I am deceived, I don't know what I am consenting to. If I am coerced, my consent itself is forced by means I would reject.[6]

The condemnation of force generates a problem for the possible consent criterion because the condemnation is absolute.[7] Kant, however, acknowledges at least four cases in which the use of force is permitted and, in the latter three cases, perhaps even obligatory. First, I may use force

to defend myself and the objects in my possession in the state of nature if the establishment of a civil condition is not possible.[8] Second, I may use force to compel others to leave the state of nature and enter the civil condition (*MM* 312). Third, agents of the state may use force to enforce its laws, as when the police interfere in the commission of crimes.[9] Finally, agents of the state may use force to punish criminals (*MM* 331-337).

It is important not to be distracted by the role of the state in these cases (direct in the third and fourth cases, indirect in the second case and, perhaps, still lurking somewhere in the first case). Kant was, of course, working in the social contract tradition. So, it might be tempting to think that, insofar as the state is involved, one can consent to the treatment one receives. This, is a complicated matter for Kant because, although one might read Kant as attributing to the idea of consent an important role in legitimizing the coercive power of the state (*MM* 315-316), at least with respect to the fourth case, he explicitly denies that one can consent to the treatment – one cannot consent to being punished (*MM* 335). I won't attempt to work through these issues here, though, because doing so won't actually help the possible consent interpretation.

In order to overcome the objection that Kant allows for cases of permissible force, a defender of the possible consent interpretation needs to deny either that in these cases the use of force really is permissible or that in these cases the use of force makes consent impossible. The first direction would involve abandoning all of Kant's political philosophy and result in an extraordinarily implausible view. The second direction seems clearly preferable. But, this direction is a non-starter given the conception of the possibility of consent we've been working with. The very nature of force, coercion and deception are supposed to make consent impossible because, at the time of the interaction, one has no real opportunity to consent. So, we'd need another conception of when consent is possible in order to make the suggestion work.

There are two natural ways to rethink the idea of possible consent. First, one might say that although one cannot consent to being the subject of, say, force at the time one is forced, one can consent to it *prior to* the interaction. So, I might say to you, for example, "If it ever seems like I am going to hurt myself or others, restrain me." But, prior consent seems like it can at most do work in cases in which the consent is actual. If the mere *possibility* of this kind of consent is all that is needed to make uses of force permissible, the possible consent criterion will no longer condemn force at all because it is not impossible for someone to make a declaration allowing *any* future use of force for *any* reason. Hence, no use of force you undertake involves negating the possibility of my consent.

Perhaps, then, we might want to consider what I could *rationally* consent to. Notice, however, that this interpretation ends up making the Formula of Humanity at least partially circular. The Formula of Humanity is a principle of practical reason. So, we need to know how it governs our actions before we can determine what we could rationally consent to. But, on this view, we need to know what we could rationally consent to in order to interpret the second half of the Formula of Humanity. Consider the matter in concrete terms. Jamie wants to know whether she can push Theodore out of her way. She wonders, "Could Theodore rationally consent to my pushing him out of the way?" Well, one of the questions we must answer to determine whether Theodore could rationally consent to this treatment is whether, in doing so, he would be treating himself as a mere means. But, on this interpretation, the only way to determine whether Theodore would be treating himself as a mere means is to determine whether Theodore could rationally consent to the treatment. Thus, we end up stuck. In the end, then, it's not clear how a defender of the possible consent interpretation could claim that one can

sometimes consent to being forced. This leaves the possible consent interpretation with no way to make sense of Kant's political philosophy.

Korsgaard does not acknowledge that her interpretation saddles Kant's view with a serious internal tension. She does, however, attempt to deal with the intuitive implausibility of the complete condemnation of force, coercion and deception by suggesting a substantial alteration of Kant's view. I consider this attempt in the next section. Here I simply want to emphasize that taking such a strategy involves interpreting Kant not only as having a view that strikes many as counterintuitive, but also as having a view that is internally inconsistent. Thus, even apart from the issues I discuss in the next two sections, we have good reason to look for an alternate interpretation.

## *II. Revising Kant*

Setting aside interpretive issues, intuitively it seems as though the use of force, coercion and deception is not always impermissible. Consider the infamous case of the murderer at the door. Someone has come looking for a friend who is in your home in order to murder him. You know what the would-be murderer is up to and you must decide whether or not to lie to him. Unlike his more moderate view on force, Kant's view on deception seems to be just as extreme as it has been made out to be. Kant claims that you must not lie, not even to the murderer at the door. This, however, strikes many as the wrong conclusion.

The reasoning behind Kant's claim here is a matter of some debate.[10] It is worth noting that Kant discusses this matter within the context of his political philosophy. But, one might try to understand the claim as a straightforward moral claim. If so, one runs into a bit of a puzzle. The Universal Law Formulation of the Categorical Imperative does not seem to condemn the murder's maxim. As Korsgaard argues, the maxim of lying to the murderer at the door can pass

the universalization test, which requires that we consider whether the maxim would be efficacious if it were adopted by everyone. In most cases of deception it would not – if everyone lied to get quick cash, no one would make loans. But, if the murderer isn't aware that you know he's a murderer, even if everyone had the maxim of lying to murderers the lie would still be efficacious.[11] If, however, we accept the possible consent interpretation of the Formula of Humanity we can see why Kant condemns lying to the murderer at the door – lying would involve treating the would-be murderer's humanity as a mere means.

In addition to delivering what Kant regarded as the correct verdict on lying to the murderer, Korsgaard suggests that the possible consent interpretation allows us to understand the motivation behind Kant's view in a way that lessens its apparent implausibility. Behind the condemnation of lying to the murderer at the door lies an attractive ideal of human relations. We must reason together with anyone we want to play a role in our plans. In other words, our relationships should be cooperative rather than manipulative.

Korsgaard suggests, however, that sometimes the ideal for human relationships that the Formula of Humanity suggests is impossible to realize. There seem to be two ways of understanding this claim. First, she might mean that sometimes the Formula of Humanity itself gives us conflicting directions and hence that the Formula is impossible to follow under certain circumstances. Second, she might mean that, although it is clear what the Formula of Humanity directs us to do under these circumstances, doing so will not enable us to realize the ideal of cooperative relationships. Although there is some textual evidence in support of attributing the first position to Korsgaard, I don't think that she has the resources to make that view work. And, the second interpretation makes best sense of the solution she offers. So, for reasons of space, I'm just going to consider the second interpretation here.

Let's consider, then, the claim that, although it is clear what the Formula of Humanity directs us to do in the murderer case, doing so will not enable us to realize the ideal of cooperative relationships, which the Formula of Humanity is supposed to be directing us toward. Cooperative relationships require the cooperative participation of all parties. I cannot be in a cooperative relationship with you if you refuse to cooperate. This is essentially what is going on in the case of the murderer. Because of the murderer's attitude with respect to both you and his potential victim, your relationships already fail to realize the ideal of cooperation. For this reason, doing what the Formula of Humanity requires you to do, namely not lying, won't actually lead to the realization of the ideal the Formula of Humanity is meant to capture. Moreover, it seems in a certain way perverse to continue to treat the murderer as a real participant in a cooperative relationship when that is clearly not what he is.

Korsgaard's proposed solution involves what she calls a 'double-level' theory. On this view, the Formula of Universal Law sets a minimum standard for the permissibility of our actions that must never be violated – we must never act on a maxim that cannot be universalized. The Formula of Humanity gives us more robust standards for our actions. When, however, the point of the Formula of Humanity is not effectively realizable, the prohibition against treating others in ways that eliminate the possibility of consent gives way. In these non-ideal situations, we must work toward a situation in which the point of the Formula can be effectively realized even though that may involve doing something that the Formula itself forbids. In other words, the Formula represents a goal that guides us rather than a norm that constrains us.

As we have observed, in the murderer case, the point of the Formula of Humanity cannot be realized because the murderer refuses to cooperate. On Korsgaard's proposal, since the Formula of Universal Law permits deception in this case, one may lie to the murderer. It seems,

then, that whether or not one should is settled by whether lying will best limit damage to the potential for future cooperative relationships and facilitate the development of those cooperative relationships, i.e. whether lying would further being able to non-perversely follow the direction of the Formula of Humanity in the future. Here the Formula represents a goal, something that we should strive to be able to follow non-perversely, rather than a norm, something that we should follow here and now. Although this is not how Kant understood his view, Korsgaard suggests that he might have been sympathetic to the modification.

There is, however, a serious problem fitting this approach into an overall Kantian outlook. On Kant's view, since the will is a kind of causality it must operate in a law-like manner and, since the will is free, it must operate according to a law it imposes on itself rather than according to a law that is imposed on it externally (*G* 446-447). So, if we accept the suggestion that the Formula of Humanity should sometimes function as a goal rather than a norm, an agent must have some principled way of determining how to regard the Formula of Humanity under different circumstances. What principle could we point to in order to supply the needed guidance? The Formula of Humanity itself cannot tell us when it should be regarded as a goal rather than as a norm. If it could, the Formula of Humanity would continue to function as a norm – a norm that simply implies something different in different circumstances. So, the Formula of Humanity would never actually function merely as a goal.

Could the Formula of Universal Law direct us with respect to the status of the Formula of Humanity under particular circumstances? It is not at all clear how it could do this. And, if it did, we'd cease to have a double-level theory. Instead, the Formula of Universal Law would be the norm that governs all of our actions – a norm that tells us to do different things in ideal and non-ideal conditions. The idea of a norm that sometimes functions as a goal would cease to

reflect anything fundamental about the structure of moral reasoning but would rather be at most a convenient way of describing the implications of a norm that is always a norm. The same will be true if, instead of focusing on the Formula of Universal Law, we try to identify and explicate another principle of practical reason that can do the work of directing us with respect to the Formula of Humanity. If one had such a principle in hand – a principle that indicated what proper respect for humanity comes to in both ideal and non-ideal conditions – that new principle would serve as a norm that governs all our actions. The idea that a norm that plays any fundamental role in moral reasoning could cease to function as a norm and function instead as a goal thus seems confused.[12]

It seems, then, that the idea of a double-level theory is at odds with the idea of action as governed by principles of practical reason. A fundamental principle of practical reason cannot cease to function as a norm. So, if the Formula of Humanity has the wrong implications for how we should treat the humanity of others in non-ideal situations, this would not indicate that the Formula of Humanity should sometimes be understood as a goal but rather that we are working with an inadequate rendering of the principle of practical reason that directs us with respect to treating humanity in our own person and that of others. Whatever we say about deception, the complete condemnation of force is clearly a mistake – consider a police officer forcibly stopping a rapist. So, the Formula of Humanity on the possible consent interpretation unquestionably directs us incorrectly with regard to how to treat the humanity of others in certain cases.

Perhaps we might rework Korsgaard's argument and present it as articulating a new way of formulating the norm that governs our treatment of humanity, something like: When cooperation is possible, treat others always as ends and never merely as means and, when cooperation is not possible, work toward making it possible through actions that are

universalizable. This would be a pretty radical departure from the command to always treat others as ends and never merely as means and, as such, it might be hard to square with the ways in which Kant argues for *his* Formula of Humanity. So, it might involve rethinking much of Kant's moral philosophy from the ground up. Perhaps, in the end, if we can find no better interpretation of what it is to treat someone as a mere means, this kind of massive overhaul would be called for. But, before taking up a Korsgaardian direction in such a project, it is worth considering just how plausible the ideal of cooperation Korsgaard articulates is.

### *III. A Problem with the Ideal*

Earlier we saw that the possible consent criterion is supposed to be plausible as a norm because it articulates an attractive ideal of relationships – relationships should be cooperative rather than manipulative. In this section, I want to consider more closely what exactly this emphasis on cooperation amounts to. We will see that the focus on cooperation has implications that have not been noticed, implications that are strongly counterintuitive.

It is supposedly important that our relationships be cooperative because people should be the determiners of how their own agency is used. As Korsgaard puts it:

> On Kant's view, the will is a kind of causality (G 446). A person, an end in itself, is a free cause, which is to say a first cause. By contrast a thing, a means, is a merely mediate cause, a link in the chain. A first cause is, obviously, the initiator of a causal chain, hence a real determiner of what will happen. The idea of deciding for yourself whether you will contribute to a given end can be represented as a decision whether to initiate that causal chain which constitutes your contribution. Any action which prevents or diverts you from making this initiating decision is one that treats you as a mediate rather than a first cause; hence as a mere means, a thing, a tool.[13]

If you are a first cause, you decide for yourself whether you will contribute to a particular end. This seems to suggest that in order to avoid treating you as a mere means, you need to be aware that I am using you and have the option of refusing. This suggests a criterion that is a bit

stronger than what we have had in mind. Force, coercion and deception all supposedly require making consent impossible in order to be successful – the deceiver can only successfully deceive if the other party is unaware of what is going on and hence unable to consent. But, now it seems that the reason for condemning such actions applies equally to cases in which consent is not, strictly speaking, impossible.

An example will make this clear. Maggie takes a stranger's picture as he is walking down the street in order to use that picture in a collage for a school project. Here she is using his action (walking down the street) to further her end (creating an image for her project). She makes no attempt to hide what she is doing and does nothing to prevent the stranger from telling her that it is okay for her to use the picture. But, she neither asks the stranger's permission nor would she desist if he asked her not to use his image. (It is the perfect picture for her purposes.) Now, there is a sense in which Maggie has done nothing to make consent impossible. The stranger could inquire about her activities and respond very positively. Maggie, however, has made consent irrelevant. She's going to proceed whether he likes it or not. So, the stranger is not in control of whether his actions will further her ends. If he doesn't object, he is simply lucky that the end he is being used to further is one he endorses. But his endorsement in no way affects what will happen. Hence he is being treated simply as a mediate cause. If he objects, his attempt to direct his own contribution to what will happen is thwarted and he is again treated simply as a mediate cause.

So, when we consider the ideal that is motivating the possible consent criterion, we see that something a bit stronger than the possible consent criterion is implied. Consent must not only be possible but meaningful. Let's call the norm according to which consent must not only be possible but meaningful the 'meaningful consent criterion'.

It might perhaps seem that the meaningful consent criterion suggests the right result – that what Maggie is doing is objectionable and it is objectionable precisely because of the way she is using the stranger. But, regardless of what one says about this case, it is clear that one doesn't always need to give others a veto over uses in order to avoid wronging them. Suppose that you are trying to decide whether it is cold enough outside to wear a coat. You look outside your window to see whether people on the street are wearing coats. Here you use their actions (walking around with or without a coat) to further your end (determining what to wear). In this case, just as with Maggie, you need not do anything that prevents others from seeing what you're up to. Hence, some passerby might notice that you're looking out the window at the people on the street and attempt to signal you to stop. Unlike with Maggie's stranger, though, it seems much less intuitive to think that the passerby's wishes must be decisive. It seems as though one may look out one's window at the activities on the street even if some particular person on the street objects.

Both looking out the window and taking a picture of someone on the street seem to be activities that are governed by privacy norms. At least in contemporary American society, the latter seems more like a violation of privacy than the former. Could a defender of the meaningful consent criterion make use of this consideration to limit the applicability of the criterion in the window case? It seems not. One would need to show that the passerby is somehow consenting to participate in your plans despite his apparent objection – otherwise he is not being treated as a first cause but rather simply as a mediate cause. I argued in Section I that the possible consent criterion is not concerned with prior or rational consent but rather with consent in the interaction in question. This seems to be even more clearly true of the meaningful

consent criterion and this suggests that arguing that the passerby is consenting after all is an untenable strategy.[14]

This suggests a serious problem with the meaningful consent criterion. The very idea of a public/private distinction rests on the thought that sometimes one should be sensitive to others' wishes about how and when they are observed but that there are cases in which such sensitivity is not required. The possible consent criterion is committed to denying this – sensitivity is always required because our relationships should always be cooperative. There is much to be said about why we should affirm the need for a public/private distinction, which I won't be able to discuss here. Here I shall simply note that this distinction is a significant aspect of our commonsense view of the moral landscape and hence that denying it makes the ideal that motivates both the possible consent criterion and the meaningful consent criterion less plausible.

Why, then, did the emphasis on cooperative relationships seem so attractive as an ideal? When cooperative relationships are contrasted with manipulative ones, the former clearly seem better. But, what we've just seen is that relationships can fail to be cooperative without being manipulative. When you observe people through your window, you are neither working together with them nor are you intervening in their activities to get them to do what you want. Instead, you are simply making use of what they are already doing. It is far from obvious that this way of interacting with another is always inappropriate. Indeed, the discussion of the public/private distinction above gives us reason to think that it is not. The emphasis on cooperative relationships thus seems plausible only when we fail to see that cooperative and manipulative relationships are not exhaustive of the kinds of relationships we can be in.

To summarize, then, the ideal that motivates the possible consent criterion actually implies the stronger meaningful consent criterion. Once we appreciate the full implications of

the ideal, it seems much less plausible. It seems, then, that the possible consent criterion is an inadequate reflection of an implausible ideal.

## *IV. A New Strategy*

I have argued that (1) the possible consent criterion faces serious problems as an interpretation of Kant, that (2) the apparent implausibility of the possible consent criterion cannot be mitigated by embedding it in a double-level theory, and that (3) when the full implications of the ideal for relationships that motivates the possible consent criterion are understood the ideal is implausible. I'll conclude by briefly suggesting how we might go about trying to develop an alternate interpretation.

Kant asserts a close connection between the transgression of rights and treating another as a mere means and he also suggests that the sense in which another can be treated as a mere means is clearest in cases in which their freedom or property is assaulted (*G* 430). These remarks indicate that thinking through Kant's political philosophy should not be an afterthought when considering the appropriate interpretation of the Formula of Humanity but may rather be a helpful starting place.

As I noted in Section I, the idea of consent plays an important but complex role in Kant's political philosophy. So, taking Kant's political philosophy as a starting point need not involve abandoning any connection between the idea of consent and the Formula of Humanity. This starting point would, however, provide us with more guidance regarding how Kant understood the idea of consent and its significance. This may enable us to develop a more nuanced account of the role of the idea of consent in the Formula of Humanity as well as to identify other factors that may bear on treating humanity always as an end and never merely as a means.

---

[1] I abbreviate *Groundwork of the Metaphysics of Morals* as *G* and *The Metaphysics of Morals* as *MM*. For all citations of Kant's work, I give the page numbers of the relevant volumes of *Kants gesammelte Schriften* (published

by the *Preussische Akademie der Wissenschaften*, Berlin), which appear in the margins of most translations. All quotations from Kant's work are taken from Immanuel Kant, *Practical Philosophy*, ed. and trans. Mary J. Gregor (Cambridge: Cambridge University Press, 1996).

[2] See *G* 437 and *MM* 392. For a helpful discussion, see Christine M. Korsgaard, *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996) 110-114.

[3] Allen Wood, *Kant's Ethical Thought* (Cambridge: Cambridge University Press, 1999) 150-155. Wood rejects the Korsgaardian interpretation I'm about to consider on p.153. John Rawls also seems skeptical of placing too much emphasis on consent in interpreting the Forumula of Humanity. See John Rawls, *Lectures on the History of Moral Philosophy*, ed. Barbara Herman (Cambridge, MA: Harvard University Press, 2000) 190-191.

[4] See Korsgaard, 106-132, 137-140, 295-296 and Onora O'Neil, "Between Consenting Adults," *Constructions of Reason* (Cambridge, Cambridge University Press, 1989).

[5] Korsgaard, 295.

[6] Korsgaard, 295.

[7] It is worth mentioning that *any* criticism of force is a bit mysterious given the Formula of Humanity's focus on using another's rational nature. Suppose I shove you into my enemy in order to hurt him. Here I have used your body but there is no straightforward sense in which I have used your capacity to set ends. Yet Kant clearly thinks that some instances of the use of force run afoul of the Formula of Humanity (*G* 430). Clarifying how Kant understands the relationship between using another's body and using her capacity to sets ends requires, I believe, an investigation of Kant's political philosophy, which I won't be able to undertake here. As I go on to argue above, though, the possible consent criterion doesn't square with fundamental aspects of Kant's political philosophy and hence the possible consent interpretation may have trouble explaining the connection between using another's body and using her rational nature.

[8] Kant writes: "[I]f a certain use of freedom is itself a hindrance to freedom in accordance with universal laws (i.e., wrong), coercion that is opposed to this (as a *hindering of a hindrance to freedom*) is consistent with freedom in accordance with universal law" (MM 231). Kant uses the term 'coercion' (*der Zwang)* to include physical force. Kant indicates that interfering with another's body is inconsistent with freedom in accordance with universal laws (MM 250). So, one is permitted to stop others from interfering with one's body. Kant also indicates that one may defend one's provisional property (MM 265).

[9] Kant does not say much about enforcement as opposed to punishment. But, he acknowledges the need for a police force to provide for, among other things, security (*MM* 325).

[10] Korsgaard, 134.

[11] Korsgaard, 135-137.

[12] In developing her double-level theory, Korsgaard draws on the work of John Rawls. In *A Theory of Justice*, Rawls describes what he calls the 'general conception' of justice, which indicates necessary features of just institutions in both ideal and non-ideal circumstances. He also describes a more detailed conception of justice, which he calls the 'special conception' of justice. The special conception of justice indicates the necessary features of just institutions in ideal conditions. So, in non-ideal conditions, institutions should still embody the general conception of justice and the special conception serves as a goal. Rawls's view is not subject to the same worry I described above because the conceptions of justice tell us what institutions ought to look like but not directly what we ought to do. The worry I suggest above arises because we are trying to think about the fundamental principles that govern what we ought to do.

[13] Korsgaard, 140-141.

[14] At this point, one might one to object that the people on the street *tacitly* consent to being observed. I do not have the space here to fully respond to this objection. But, here is a quick gloss on an answer. To say that one tacitly consents to something has to mean more than that one has reason to expect that to happen. If I lock my bike up outside of my building in a high crime neighborhood, I may have good reason to expect that it will be stolen. But, of course, I do not tacitly consent to its theft. So, what is doing the work in the talk of tacit consent is, I suspect, a more moralized conception of consent than we have thus far been considering. The meaningful consent criterion will, I think, get itself into muddles if it tries to lean on tacit consent so understood.