**On the Expressive Limits of Kant's Universalizability Tests**

Abstract: My goal in this piece is to show that there is a problem lurking in the shadows of recent attempts to derive positive duties from Kant's so-called universalizability tests and, further, to show that the most obvious way of fixing these attempts renders them unable to fulfill their function. I shall begin by motivating and explaining such an attempt.
Keywords: universalizability tests; Formula of Universal Law; Formula of a Law of Nature; positive duties

My goal in this piece is to show that there is a problem lurking in the shadows of recent attempts to derive positive duties from Kant's so-called universalizability tests and, further, to show that the most obvious way of fixing these attempts renders them unable to fulfill their function. I shall begin by motivating and explaining such an attempt.

If we confine ourselves to the results of the universalizability tests ("so act as if the maxim of your action should through your will become a universal law of nature")[1] as applied to a single maxim, then it looks like a problem arises: a maxim either will pass the test or it will fail the test, yielding only two options even though we need (at least) three (permissible, obligatory, and forbidden). In chapter 5 of Onora O'Neill's justly famous *Acting on Principle*, she proposes a solution. O'Neill argues that we can determine the deontic status of an action in the following way.

Start with a maxim and its contrary. O'Neill takes maxims to be subjective principles of volition, such as "I will A if B," and she says that two maxims are contraries if but only if "their UTC's are contraries."[2] 'UTC' is shorthand for universalized typified counterpart. The UTC of "I will A if B" is "everyone will A if B." The universalizing means that the "I" of a maxim becomes an "everyone" in the counterpart; the typifying means that the "willing"/volition of the maxim becomes "willing"/happening-according-to-a-deterministic-law-of-nature in the counterpart. Thus, because "everyone will A if B" and "everyone will omit A if B" are contraries as UTC's (they cannot both be true even though they both can be false), O'Neill says that "I will A if B" and "I will omit A if B" are contrary maxims.

---

[1] The german is as follows: "handle so, als ob die Maxime deiner Handlung durch deinen Willen sum allgemeinen Naturgesetze werden sollte" (GMS, AA 04: 421.18-20, emphasis omitted; translation above is my own).

[2] O'Neill, Onora: *Acting on Principle*. Cambridge, 2013, 162n25.

Now when an agent is deliberating about whether to A, s/he can adopt one but only one of a maxim/contrary pair. This, according to O'Neill, enables us to derive obligatory maxims from the universalizability tests because, if an agent is deliberating about whether to A if B and "I will A if B" is universalizable whereas "I will omit A if B" is not, then the agent has only one permissible option, from which it follows that that option is obligatory.

More formally, O'Neill's proposal is as follows. Test a maxim and its contrary to determine whether either can be conceived or willed as a universal law without contradiction. There are then three options:

1. If a maxim is universalizable and its contrary is not, then acting on the maxim is obligatory and acting on its contrary is forbidden and, thus, there is a positive duty to act on the maxim and a negative duty not to act on its contrary.
2. If a maxim and its contrary are both universalizable, then acting on either is merely permissible.
3. If neither a maxim nor its contrary is universalizable, then acting on either is merely permissible.

Thus, for example, the maxim "I will to help others" is universalizable but its contrary, "never to help anyone," is not, whence it follows that there is a positive duty to act on the former and a negative duty not act on the latter. By way of contrast, the maxim "to buy clockwork trains" and its contrary "never to buy clockwork trains" are both universalizable, whence it follows that acting on either is merely permissible.

The third option is a bit trickier, but O'Neill illustrates it by appeal to what she calls "non-reciprocal action maxims," like "to buy clockwork trains but never to sell them" and "to sell clockwork trains but never to buy them." These two maxims are contraries, and neither is universalizable—but both are intuitively permissible, and so O'Neill fills out option 3 as above.[3]

O'Neill's account is quite popular. For instance, consider the following short selection of philosophers who follow suit:

---

[3] O'Neill, Onora: *Acting on Principle*. Cambridge, 2013, 165.

1. "Kantian ethicists typically classify actions in three ways: morally permissible acts (whose maxim and its contrary do not contradict the moral law), morally obligatory acts (whose contrary maxim conflicts with the moral law), and morally forbidden acts (whose maxim contradicts the moral law)."[4]
2. "...if I must reject the maxims of letting all my talents rust or never helping anyone else, then I must accept their logical contraries, namely, maxims of cultivating at least some of my talents and helping at least some other people some of the time."[5]
3. "A maxim is fit to be a law in one sense, the sense corresponding to permissibility, if it could function as a law. It is fit to be a law in a stronger sense, the sense corresponding to obligation, if it not only can but must be a law. The way we ascertain this is by showing that the maxim of doing the opposite is unfit to be law, and must be rejected."[6]
4. "If the maxim [to which the first formulation of the Categorical is applied] fails the universalizability test, the action is forbidden. An action is required if the negation of its maxim fails the test. Maxims that otherwise pass the test are permissible."[7]

The problem that arises now, however, is that a contradiction can be derived from this account quite easily. To see how, consider the following three maxims: (i) never to help anyone; (ii) to help white supremacists to kill nonwhites; and (iii) to help nonwhites not to be killed by white supremacists. These three maxims are pairwise contraries. I suppose that (i) and (ii) are not universalizable whereas (iii) is. Someone might reject my supposition, but that only would create further problems for the account. But if my supposition is accepted, it follows immediately that actions performed on maxims (i) and (ii) are both merely permissible and forbidden.

Perhaps someone will contend that option 3 should be revised: if neither a maxim nor its contrary is universalizable, then acting on either is forbidden. This would be an interesting revision, not least because of what it would require us to say about non-reciprocal action maxims. While it certainly does

---

[4] Hernandez, Jill: „Impermissibility and Kantian Moral Worth". In *Ethical Theory and Moral Practice*, 2010, 403–419, at 403.

[5] Guyer, Paul: *Kant*. New York, 2006, 194.

[6] Korsgaard, Christine: *Self-Constitution*. Oxford, 2009, 16.

[7] Stohr, Karen: „Kantian Beneficence and the Problem of Obligatory Aid". In *Journal of Moral Philosophy,* 2011, 45-67, at 50.

seem plausible that acting on *some* non-reciprocal action maxims is impermissible (e.g., "to rape but not be raped"), if acting on "to buy clockwork trains but never to sell them" is impermissible, then we have to give up on what can be called (following Williams) the principle of permission agglomeration: if it is permissible to A (in this case: to buy clockwork trains) and it is permissible to B (in this case: not to sell clockwork trains), then it is permissible to A and B. That would be an interesting result.

But if we are going to be pushed toward rejecting the principle of permission agglomeration, it will not be from the theory embodying these three (now revised) conditions. It will not be from this theory because a contradiction still can be derived from it quite easily. To see how, consider the following three maxims: (I) to become a doctor and not an engineer; (II) to become an engineer and not a doctor; and (III) to become a serial child rapist and neither a doctor nor an engineer. These three maxims are pairwise contraries. I suppose that (I) and (II) are universalizable whereas (III) is not. Again someone might reject my supposition, but as before that only would create further problems for the account. And if my supposition is accepted, it follows now that actions performed on maxims (I) and (II) are both merely permissible and obligatory.

The problem here, I think, is pretty clear. It is that the deontic status of acting on a maxim is not being determined solely by something about the maxim itself; it is being determined by something about the maxim in conjunction with something about its contrary, and because a maxim can have many (many, many) contraries, that second piece of the puzzle is variable. So if the deontic statuses that result can conflict, problems will arise.

But, you might say, all is not lost yet! Why not modify option 2: why not say that if a maxim and its contrary are both universalizable, then acting on either is permissible rather than *merely* permissible? An action that is obligatory is (*a fortiori*) permissible (at least if we buy into a principle commonly accepted in deontic logic). So this modification would eliminate the contradictions.

Well, yes and no. It would eliminate the contradictions I was deriving above. But it would not eliminate other contradictions. Consider maxims (I), (II), and (III) again. From what already has been said, it follows that acting on both (I) and (II) is obligatory, whence it would follow that any agent trying to decide whether to become a doctor or an engineer is going to be in a serious quandary. Moreover, if we buy into another principle of deontic logic (sometimes just "the" principle of deontic logic), $\Box (A \rightarrow B) \rightarrow (OA \rightarrow OB)$, further problems might arise because becoming a doctor and not an engineer necessarily entails not becoming an engineer and not a doctor.

So maybe the solution is to scale back on the scope of option 1. Maybe option 1 should be: if a maxim is universalizable and *all* of its contraries are not, then acting on the maxim is obligatory and acting on any of its contraries is forbidden. This, I think, avoids contradiction (provided we do something about option 3, too). But it renders the tests too unwieldy for use: any maxim is going to have infinitely many contraries, and there is no way for any human agent to test all of them. So if there are positive duties, there will be no way for us to know that, at least not by using the additions to the universalizability tests proposed here. And given that that is exactly what these additions are supposed to enable us to do, this failure, I think, is a serious problem.

Alternatively, perhaps the solution is not to appeal to maxim contraries. As seen above, O'Neill defines maxims contraries in terms of their associated (propositional) UTC's. To make things more manageable we might consider simply paired maxims of commission and omission for any given action. The idea would be that if an agent is deliberating about whether to A and if "to A" is universalizable whereas "not to A" is not, then A is obligatory (and conversely). The other parts of O'Neill's account (conditions 2 and 3) can be rewritten in a similar fashion.

The problem that arises now is that not all agents, not even all agents involved in deliberation about whether to A, will adopt one of these two maxims. In fact, this reveals a further problem with O'Neill's original account (one that has wider application than O'Neill's account, although I cannot explore these wider applications here). Even appealing to the infinitude of maxim contraries is not sufficient to capture the complexity of deliberation; appealing to a single maxim of commission and a single maxim of omission only magnifies the issue. An example will illustrate my point.

Suppose I am deliberating about whether to help a friend who is in the process of moving from one apartment to another. Then my deliberative options will include maxims like (c1) "to help friends" and its corresponding maxim of omission, (o1) "not to help friends." But I also might act on the maxim (c2) "to help friends move" or its corresponding maxim of omission, (o2) "not to help friends move."

Notice that the four maxims in the previous paragraph are not pairwise contraries. It would be possible for an agent to act on both maxims of commission, (c1) and (c2), or both maxims of omission, (o1) and (o2). It also would be possible for an agent to act on some combination of the maxims of omission and commission. (For example, I might will to help friends in general but not when they are moving, acting on (c1) and (o2).) And I have listed only four options for space constraints; plainly there are others. From this it may be seen that this alternate solution does not obviate the difficulties associated

with the one it was intended to replace: infinitely many maxims still must be tested in order to determine whether a given maxim is obligatory.

It is tempting at this point, I think, to assert that acting on *any* maxim of commission when deliberating about whether to help friends entails acting on (c1). And one might think that this could be used as a reply to the objection in the previous paragraph. The idea would be that if acting on any maxim of commission entails acting on (c1), then we shall be able to simplify things so that only one or two maxims have to be tested after all. But there are two problem with this.

The first problem is that acting on a maxim by entailment is not a thing. It is true that any agent is going to help or not to help. But so is any non-agent. What this overlooks is that maxims are agents' principles of volition. What sets agents apart from non-agents is that we are capable of governing ourselves in accordance with the representation of a law, and that representation is what a maxim is. From the fact that an agent's behavior can be described as Aing it does not follow that s/he is acting on the corresponding maxim, nor does it follow that an agent is acting on some maxim(s) because the propositional content of his/her actual maxims entails the propositional content of the other one(s). The issues here should be familiar from discussions of the deductive closure of belief (from the fact that I believe P it follows that I would be justified in believing P or Q *if* I deduce the latter from the former; it does *not* follow that I *actually* believe P or Q).

The other problem is that even if these ideas about maxim entailment could be made out, there still will be infinitely many maxims to test. In the helping example, acting on o2 (or some other similar maxim of omission) does not entail acting on o1, so even if acting on c2 entailed acting on c1 (it does not but even if it did) the test being proposed still would be too unwieldy for (human) use.

To be clear, I am not disputing the following conditional: if an agent has only two options and only one of those options is permissible, then the permissible option is obligatory. What I am arguing is that the antecedent of this conditional never will be satisfied in a way that will enable us to derive positive duties from Kant's universalizability tests as envisioned in the foregoing.

One way in which one *could* argue for this would be to focus on the moral part of the antecedent, to argue that agents always will have more than one *permissible* deliberative option. I suspect that that is true. But what I have argued here is slightly different: I have argued that agents are never faced with only two deliberative options, and there does not seem to be a non-problematic way to limit the number of

maxims that need to be tested. To borrow the words of Razumikhin, “You can’t skip over nature by logic. Logic presupposes three possibilities, but there are millions!”