From Supervenience to "Universal Law": How Kantian Ethics Become Heteronomous

Scott Forschler

Abstract

In his *Groundwork of the Metaphysics of Morals*, Kant's desiderata for a supreme principle of practical reasoning and morality require that the subjective conditions under which some action is thought of as justified via some maxim be sufficient for judging the same action as justified by any agent in those conditions. This describes the kind of universalization conditions now known as moral supervenience. But when he specifies his "formula of universal law" (FUL) Kant replaces this condition with a quite different kind of universality: the judgment that some agent could rationally (i.e., without willing the frustration of his own valued ends) will his adoption of some maxim under the condition that this would cause all agents in his world to adopt it as well. Our wills typically lack this efficacy, so requiring that our wills conform to what would be rational for a hypothetical agent in this situation to will is a heteronomous requirement. Several intuitively wrong maxims pass Kant's test but fail the test of supervenience, because they generate no contradiction in a world of universal compliance but do so in non-ideal worlds, demonstrating the inadequacy of the FUL and the logical superiority of moral supervenience.

1. Moral Supervenience and Kantian Ethics

Moral supervenience[1] is a commonly-accepted contemporary view. R. M. Hare[2] pioneered the philosophical use of this idea, which was central to his argument from universal prescriptivism to utilitarianism. Simon Blackburn[3] used the apparent obviousness of the view as an attack on moral realism, understood as the claim that moral properties exist inde-

1 The general concept of supervenience was later taken up by philosophers of mind including Donald Davidson and Jaegwon Kim, and is also used in other areas of philosophy, but I will not be concerned with these applications here.

2 Hare (1952, 145).

3 Blackburn (1993, 122).

pendent of non-moral properties of the world. Michael Smith[4] concurred in the view, claiming with little argument that "everyone agrees" with it, and that it is true *a priori*, on the ground that any judgment violating it would thereby fail to count as a *moral* judgment.

Hare defined the concept as follows: a fact r was supervenient on some other facts just if "Necessarily, if r, then there is a valid inference" of the form "p: for all x, if Gx then Fx; q: Ga; So r: Fa", and the two premises p and q are true.[5] We can then say, for any object x, that r (Fx) supervenes on q (Gx), or that in general F-ness supervenes on G-ness. The premise p might be called the supervenient principle which connects r and q. This formulation has been translated into the more contemporary formulation that A supervenes on B just if there is no A-difference without a B-difference, where A and B play the role of Fx and Gx in Hare's analysis. Such abstract formulations leave it completely open just which other particular facts moral facts might supervene upon. Hare argued that if we rule out supervenience on facts described essentially with reference to individuals, then we are left with supervenience upon the natural properties of situations describable using only universal terms. Blackburn and Smith likewise agreed that moral properties supervene upon natural properties. Blackburn even diagnoses the temptation to think otherwise as "mis-identifying a caprice as a moral opinion."[6]

Centuries before moral supervenience was so defined, Immanuel Kant agreed with its basic claim when he insisted that the validity of a maxim of action, specifying the subjective motivational and external conditions under which an agent will perform some action, must be governed by a moral law which appears to us in the form of a categorical imperative.[7] "Categorical" just means "not hypothetical," that is, not depending upon contingent conditions such as the subjective desires of the agent. Morally valid maxims must conform to this law "which contains no conditions to which it could be limited" except for "the universality of a law as such." (*Groundwork of the Metaphysics of Morals* (= *G*), AA 4:421) Any other condition would be "heteronomous,"

4 Smith (1994, 21).
5 Hare (1989, 69–70).
6 Blackburn (1993, 122).
7 I omit here any part of Kant's argument that moral laws must indeed be categorical and hence (in some sense) universal in nature. I take this claim to be true, but make no attempt to support it directly.

making the validity of a maxim depend upon some contingent fact, such as but not limited to the satisfaction of the current agent's desires (*G*, AA 4:433). Fully rational, moral willing must on the other hand be autonomous, whereby the agent acts on laws of his own rational making, i. e. contingent only upon their conformity to universal law, not to any contingent condition. None of this means, of course, that a moral agent cannot act to satisfy his subjective desires, or the desires of others, nor does it even rule out at this stage the possibility that morality consists of nothing but the maximal satisfaction of such desires, though Kant later tries to show that it does do this. Rather, it means that any maxim which describes the existence of some desire(s) in its subjective conditions, and the satisfaction of such desires in its specified action, may only be followed if the maxim as a whole can pass the test of this moral law which itself permits of no subjective variation in *its* application.

To tie this back into supervenience: for Kant, the moral permissibility of some action supervenes upon the subjective conditions of the maxim which the agent is following while performing it. For some such conditions, the action is permissible, while for others it is not, but there can be no case of two agents following the same maxim yet it being permissible for one but not for the other. O'Neill conveniently describes a Kantian maxim schema as "M: In circumstances C, I will do action A, to achieve end E."[8] The conjunction of C and E (the agent a's external circumstances and subjective motivation) constitute Ga. Fa is the property of action A's being permissible for agent a to perform.[9] To take a maxim to be morally valid is, then, to assume that some true supervenience principle makes Fa true if Ga is true (i. e., that it is permissible to A to pursue E if I am in C). The moral law or categorical imperative which we seek, then, is a law which guarantees that for all agents x, if Gx holds, Fx also holds, with no exceptions. Kant forbids the contingent, subjective determination of Fx's permissibility, once Gx is established, on the basis of contingent facts, like the original agent a's

8 O'Neill (1975, 37).

9 Alternatively, we might map Kant's schema onto Hare's by identifying Fa with "a permissibly pursues E with A" and Ga with "a is in C." But I would suggest we evaluate the permissibility of adopting end E separately from the permissibility of using any particular action as a means of pursuing it. Presuming the permissibility of pursuing E to be established via a logically prior test, I will take the intention to pursue E as part of the background conditions, along with C, purportedly justifying the permissibility of A via some maxim.

subjective desires and a's relationship to x (identical or non-identical). Kant repeatedly castigates any instance of doing so as amounting to making a special exception of oneself. We may be tempted, for example, to think it permissible for us to tell a lie in circumstances when we would not want others to tell a lie, but this is just an instance of admitting that for some x and y, Gx and Gy both hold, arbitrarily accepting Fx while denying Fy, yet taking Gx to be the reason for Fx being true (for Gx, again, is by hypothesis the conjunction of being in C and pursuing E in our subjective maxim). Such reasoning gives the lie to the claim that we can consistently maintain Gx to be a sufficient reason for Fx, revealing a contradiction between the instances of our F-judgments.

Again, it is necessary to be careful here. It is not that subjective desires play no role in determining when an action is permissible; misunderstanding on this point has often confused both Kantians and their opponents about the role which desire-satisfaction can play in Kantian ethics. Subjective motivations have their role in maxims; indeed, every maxim *must* have some such motivation. But that's their *only* role; they play no *further* role in practical reasoning, and in particular they do not enter into the test for the moral legitimacy of maxims – the categorical imperative, here described as moral supervenience. But of course a subjective motivation for some action is wholly unobjectionable *just if* it motivates action via some maxim which passes that test.[10]

2. The Formula of Universal Law

Kant next claims that the conditions so far specified for the categorical imperative are completely captured by his Formula of Universal Law (FUL), namely that each agent should "act only in accordance with that maxim through which you can at the same time will that it become a universal law." He immediately equates this with a formula requiring that agents should "act as if the maxim of your action were to become by your will a universal law of nature." (*G*, AA 4:421) There are several

10 I agree with Henry Allison (1990, 39–40) that Kant's test seeks to determine when we can properly incorporate our desires into our wills via morally-acceptable maxims, which is not at all the same thing as doing without desires altogether. Some Kantian passages may be, or appear to be, inconsistent with this claim, but I don't think any alternative can be seriously defended.

cases in which one could not rationally will the maxim to be followed universally, as illustrated by his examples which follow.[11] A maxim of falsely promising to repay a loan in case of need cannot be willed as a universal law because if everyone followed the same practice, then "all would laugh at such expressions as vain pretences" and the needed loan could not be obtained (*G*, AA 4:422). A maxim of failing to develop one's talents could be a law of nature, but could not be willed as a universal law, for this amounts to willing that some of the necessary means to one's future ends – namely, the developed talents which would further those ends—be absent, which again amounts to willing that one's ends be frustrated. Finally, a maxim of failing to help others in need would, if universalized, frustrate one's own ends in the "many cases… in which one would need the love and sympathy of others" so any agent willing this would will a world in which he would "rob himself of all hope of the assistance he wishes for himself." (*G*, AA 4:423) In summary, Kant says that willing the universalization of these maxims generates a "contradiction in our own will" (*G*, AA 4:424), whereby what one tries to will is either a wholly inconceivable situation, or a conceivable situation which would frustrate our ends. The contradiction lies in the fact that as a purposeful agent, one wills the satisfaction of some ends or other, but at the same time through willing the universalization of the maxim, wills that circumstances be such that some such ends *not* be satisfied. To avoid a contradiction in one's will, the agent must abandon or modify the maxim which creates it when universalized.

I will note here a curious tendency amongst many scholars and teachers of Kantian ethics to focus on examples, like that of the promise to repay a loan, in which the end being frustrated by the universalization of some maxim M is the specific end E of the very maxim in question.[12]

---

11 I omit his first example, involving suicide for the relief of pain, as it seems to me to present special problems which would only confuse the points at hand.

12 For instance, Korsgaard suggests that the best interpretation of Kant's idea of a contradiction in will is that the contradiction is neither logical nor teleological (contrary to a natural end in nature), but "practical," meaning just that the end of this very same maxim cannot be satisfied if the maxim is universally practiced. She then admits difficulty in dealing with counter-examples in which the end of a clearly bad maxim is not so frustrated, if its universalization does not frustrate its *own* end. (Korsgaard 1996, 100) She weakly concludes that such maxims are not so problematic since they are rarer than the ones which

Perhaps the ease of grasping these sorts of examples (where lies cannot attain their purpose in a world where lying is the norm, and you can't get to the front of the line by budging if everyone budges, and so forth) is partly responsible for their pervasiveness. Yet this is only one possible way Kant describes of generating a contradiction in will. In his last two examples – failing to develop your talents and failing to help others – the maxim in question is not fully specified, but we can imagine their ends being something like the temporary ease or convenience attained by not bothering oneself with developing one's talents or helping others in need. Kant never suggests that *these* ends will be frustrated by their maxim's universalization. Rather, it is some *other* ends – the other "possible purposes" one might have, for which one "needs the love and sympathy," and practical assistance, of other agents, which would be so frustrated. (*G*, AA 4:423)

Of course, to make his case complete, Kant would need to show that it is not possible (or not rational) for an agent to have just the one purpose of attaining immediate ease and comfort. I find this claim plausible, but shall not try to defend it here. But clearly Kant assumed it was true, and used it to identify contradictions in will as arising from willing the frustration of *any* ends you posit as justified, not just the one you are aiming at in your current action.[13] In many cases, then, it is not merely a single maxim, but the conjunction of two or more maxims, which generates the contradiction.[14]

---

do cause problems, but this is clearly an inadequate analysis for a principle purporting to be the "supreme" principle of morality. (*G*, AA 4:411)

13 I show in my (2007) that if we evaluate our maxims or principles not one by one, but as sets, and also observe that the choice of each particular maxim pragmatically implies commitment to more general higher-order maxims with more general ends – in particular, the end of adopting or conforming to the lower-order maxim – we strengthen the universalization test because these higher-order maxims must also pass it. Many maxims of the sort which troubled Korsgaard (see previous note), or which are taken to be counter-examples to similar moral standards like the golden rule, fail to pass such tests, showing that otherwise troublesome "counter-examples" to Kantian or other universalization tests (the golden rule, etc.) fail those tests at higher-order levels, bringing the judgments of such tests in line with our intuitions. See also Wattles (1996, 6) and Reinikainen (2005, 159) for versions of this argument.

14 This is also suggested by his Formula of a Kingdom of Ends, which permits us to will the end of each of our maxims just if it can be part of a rational "whole of all ends in systematic connection." (*G*, AA 4:433) I take this to mean that each of our contingent ends is rational just if it is compatible with the rationally necessary ends of all beings (like preserving their life, developing their talents,

3. Universal Law vs. Supervenient Universality

All of this is important, but we need to go back and notice something even more important: when Kant specifies the FUL, he abandons the condition of supervenience, which as noted earlier is equivalent to avoiding heteronomy. The test given by the FUL asks if a contradiction in will arises just in the case where *everyone follows the same maxim*. But this is, rather obviously, a contingent condition which does not always hold; it is one of many possible conditions an agent might find himself in, and one of the more unlikely ones at that. Knowing that a maxim generates no contradiction in will in *that* condition by no means guarantees that it will not generate a contradiction in will in any other, and to suppose otherwise commits a subtle logical mistake.

I suspect that the above paragraph will be well-nigh incomprehensible to many Kantians on its first reading, and might appear to completely misunderstand the Kantian formula of universal law, and its concept of universalization. But I wish to show, rather, that it is Kant and many of his followers who have misunderstood the concept of universalization for over two centuries. For right at the beginning of his analysis – at *G*, AA 4:421 to be precise – Kant mixed up two completely different kinds of universality. He substituted the wrong one – universalization *within a world* – for the right one – supervenient universalizability, or universalizability of judgment.

There are crucial differences between the two kinds. Supervenient universalizability ranges not over a set of agents who happen to inhabit a world, but over the moral qualities of any acts of any possible agent, or perhaps more precisely over the judgments a practically rational agent makes about any possible agents acting in response to the antecedent

---

etc.) and with the other contingent ends of all beings (including your own) which pass the same test of rationality. There may be some circularity involved here, if the rationality of a's end E is contingent on whether or not it conflicts with b's rational end F, but F's rationality is likewise a function of whether it conflicts with E, should E be a rational end. I suspect this problem is solvable if we adopt general meta-principles for coordination and adjudication between potentially conflicting ends which are themselves universalizable, but the issue is too large to discuss further here. Assuming that there is a solution to this problem, I will call ends which pass this test our "rational ends," and say that an end is rational just if its universalization does not conflict with other agents' rational ends. Of course, some conflict is unavoidable; the relevant goal is more precisely its minimization, via willing the maximal probability that rational ends as such are to be satisfied; see §4 below.

conditions of some maxim.[15] In terms of the elements of maxims, it ranges over all possible instances of agents in C and pursuing E by means of A, and requires that whatever judgment I make of one such case, I must repeat in all such cases, if I think of (being in C and pursuing E) as indeed being the practical justification for (doing A). It is crucial that the universality here ranges over *all possible* instances of a maxim's instantiation, not just all cases in a particular world. For if there were some possible case where the antecedent conditions were present but the permissibility judgment were denied, this would show that the permissibility judgment was limited by some contingent conditions. This is precisely what Kant ruled out when defining the concept of a categorical imperative.

Both before and after introducing the FUL, Kant makes clear that the categorical imperative he is seeking must range over acts of our will, our evaluations of maxims, for doing so in conformity with such an imperative would alone give "moral worth" to our maxims. The principle he seeks is a "principle of volition" or a "principle *of* the will" (*G*, AA 4:399–400), "which alone is to *serve* the will as *its* principle" (*G*, AA 4:402; my emphasis on both quotes). We are clearly not talking here about a law whose universality consists in commanding other agents about, but one which guides *my will alone*, now and into the future. Naturally, if it is a law that is always valid for my will, regardless of who I am or in what situation I am in, then it is valid for all other agents as well. But we must not confuse the question of the law's validity for an agent with the question of what *form* the law takes, including what it ranges over or how it operates if it is adopted by some agent for whom it is valid. Christine Korsgaard describes another aspect of Kant's requirement by suggesting that the function of such a principle is to provide unity to our wills, without which we lack either full practical rationality or a coherent set of values. To attain such unity

> every rational agent must will in accordance with a universal law […] [which] ranges over all rational beings, that is, it commands you to act in a way that any rational being could act, because you could find yourself in anybody's shoes, anybody's at all, and the law has to be one that would enable you to maintain your integrity, in any situation, come what may.[16]

15 Note that Kant shares with Hare and Blackburn the view that moral facts derive from facts about willings, prescriptions, or judgments, and not from some perceptual detection of substantive moral facts or qualities lying outside of us.

16 Korsgaard (2009, 214).

Again, the supervenience principle seems to describe this, by requiring that the principles we act on, and which determine when we will apply moral predicates to various acts and situations, be ones that we can continue to find applicable in any situation whatsoever.

But both Kant and Korsgaard are mistaken in thinking that the FUL fulfills the promised roles. For the universalization of the FUL is not one that governs my will or acts of volition; rather it ranges over a set of agents, and only those in a specific (and highly imaginary) world. In fact, it doesn't even test for supervenience within that world. It does not ask whether we can make the same moral evaluation of each agent following M in such a world – let alone of any agent in any world, or equivalently of myself in a situation I may find myself in. Rather, it confronts us with a completely different choice, asking if we can will without contradiction our performance of the following act: doing A in C_1 to pursue E, where C_1 = the conjunction of C *plus the fact that if you follow this maxim so will everybody else in your world*. But as C_1 is palpably not the same as C, this change in conditions makes the proposed test almost completely irrelevant to the original question of whether it is rational to do A in C to pursue E. In fact, C_1 contains a limiting condition of the acceptability of a maxim of precisely the sort that Kant claimed to rule out of bounds just sentences earlier. By assuming without argument that if action A is acceptable in C_1+E, then it must be acceptable in C+E (for any A, C, and E) he has violated his own stipulation that a maxim's validity must be contingent upon *no* limiting conditions whatsoever. The universalization present in the FUL is not a law of volition, or a law of my will; it is a law imposed upon a set of agents within a world as a result of a hypothetical volitional act of one person.

It may help to briefly represent the difference in formal symbolism, although I will make no attempt to formalize the rest of my argument involving the distinction.[17] If we take $\mathbf{W}\varphi$ to be a modal operator mean-

17 I attempted this in my (2010), but am now skeptical of my initial trial, which involves some weighty and subtle issues in formal logic. The failure is certainly largely due to my lack of mastery of the required symbolism, but I will also note that there is fairly little scholarship to build on in this area, and it is possible that the requisite formal notation and systems for representing Kantian ethical formulas have simply not yet been devised. The sets of Kantian ethicists and of deontic logicians seem to have few common members. This is particularly odd considering that Kantians frequently employ terms like "universal" and "necessity" which are part and parcel of logical notation, which one would have thought would facilitate the formalization of Kantian ethics, or at least at-

ing "I will φ to be the case," and Mx to mean "x follows maxim M," then we can distinguish the following:

FUL Universalizability: **W**(x)Mx
Where (x) ranges across all agents in a given world.

Supervenient Universalizability: (x)**W**Mx
Where (x) ranges across all possible agents in all possible worlds.

The problem with substituting the question of the rationality of an agent following M with that of an agent following a different maxim, which we might call M_1, where C is replaced with C_1, is not a new one; it was pointed out a century ago by Broad (1916), and doubtless has occurred to many people before and after this time. Frankena (1964) was also alive to the distinction, noting that the supervenience championed by R. M. Hare was importantly different from the kind of universalization used in the moral theories of M. G. Singer and rule utilitarianism – and he could have added, Kant. However this distinction has been made far too seldom in moral philosophy, and too often ignored thereafter. Even Hare, in a surprisingly late essay (2000), assumed that the two kinds of universalizability were equivalent, and hence argued that Kant should have derived consequentialist norms from his conception of universalization. But Kant's universalization test is not the same as the requirements of moral supervenience, and the former is certainly incompatible with a consequentialism which takes the moral properties to an act to be a function of its *actual* consequences. Of course this leaves open the question of whether moral supervenience might lead to consequentialism; more on that anon.

Kant's confusion may be partly based on a false analogy between moral laws and physical laws; the latter indeed govern the behavior of a class of objects within a world or universe. But that by no means suggests that moral laws must resemble these; indeed they palpably do not in at least one respect, for they can be disobeyed. Kant does not, of course, think that disobeying moral laws is impossible, yet by suggesting that the possibility of willing a world where some maxim is invariably followed is a sufficient condition for its rational adoption by any agent in any other world he may reveal that the idea of a physical law as a "typic" for moral ones held too great a grip on his imagination (*Critique of Practical Reason* (= *CPrR*), AA 5:67–71). But this is speculation.

tempts to do so. A deeper investigation into the possibility of doing so is long overdue.

A more important factor may be that any M's passing the FUL universalizability test is indeed a *necessary* condition of M's passing the supervenience test. This is because if I satisfy the supervenience requirement that I be able to will of *any* possible agent that she follow M (i. e., that she be permitted to do A in C to pursue E), then I can of course will this for all agents in a single world; if the latter fails then by *modus tollens* we know the former does as well. As a result, the FUL, if used properly, should not produce false negatives where it rejects a valid maxim; any maxim which fails it will also fail the supervenience test, though as I will show below, the reverse implication does not hold.[18] Most of Kant's examples are indeed of maxims which fail the FUL test, but this focus tends to obscure its deeper problems, which are more apparent when we try to treat it as a *sufficient* test for a maxim's permissibility.

Some readers may still be baffled by my suggestion that the test of willing M in a world where all agents will also follow M can be a "limiting condition" – for doesn't this precisely mean that M's applicability is not limited, by covering *all* agents in a world? But this again reveals a confusion between two different conceptions of what kind of conditions we're talking about, and for what kind of result, which equivocal words like "applicability" can gloss over. Within such a world, there is of course no condition limiting *which rational agents actually follow M*; they all do. But that wasn't what we were supposed to be after. We wanted to know if my following M, in whatever situation I happened to be in – in other words, in any situation whatsoever, categorically or universally – is *rational to will*, and hence has moral worth or the properties of rightness or permissibility in just that situation. And if we replace this with the question of whether M is rational to will just in a world *where everyone else follows it if we do* – ignoring what may happen in other worlds with different behavioral laws – or with the question of whether it is possible for a maxim to be endorsed or acted on "univer-

18 Christine Korsgaard (2008, 122–123) argues that the FUL essentially forces us to modify many apparent false-negative maxims with common-sense qualifications, after which they pass its test in ways which match our intuitions. I agree, but this response cannot apply to false positives; since these already pass the FUL test, it cannot force us to qualify them further. Herman (1993, 139) rejects this maxim-modifying strategy as it would seem to apply equally well to (acceptable) coordination maxims and (unacceptable) free-rider ones; I show below how the present analysis puts this problem to rest.

sally" by all agents within a world at once,[19] then we are not only asking the wrong question, but the wrong *kind* of question entirely.

This substitution raises any number of problems, but one particular problem occurs if the rational permissibility of the maxim depends precisely upon these limiting conditions being satisfied. To prove that this is possible, I will now present several such maxims which pose no risk to our rational ends if everyone follows them, but which disastrously frustrate the same if followed by some people in many situations where not everyone is following the same maxim. For some reason this particular class of "false positive" cases have rarely been considered in the context of Kantian ethics, although there is occasional discussion of the more general classes of maxims of confronting evil or solving coordination problems, of which these are sub-classes. Consider the following:

> Maxim of Left-Hand Driving (MLHD):[20]
> When I want to drive somewhere, I will drive on the left side of the road, to arrive safely.

19 Much is made out of the distinction between a contradiction in will and the supposedly stronger contradiction in conception, which are described by these last two phrases respectively. But unless one can will a contradictory state of affairs, the latter is surely a subclass of the former. For this and other reasons I think the significance of the distinction has been greatly exaggerated.

20 I owe this example to Hardin (1988, 67), who applied it against a certain interpretation of Marcus Singer's generalization argument, showing that it led to the "absurd – murderous and suicidal" result that since the consequences of everyone driving on the left (or right) would be "desirable," it would be morally correct for me to drive that way *now*, even if not everyone else is doing so. Singer (1961) actually made the same mistake as Kant, for his "generalization principle" that "what is right (or wrong) for one person must be right (or wrong) for any (relevantly) similar person in (relevantly) similar circumstances," is simply another version of the supervenience principle, while his "generalization argument" that if the consequences of *everyone's acting in a certain way* would be undesirable, then no one ought to act in that way without a reason, switches again to the wrong kind of universalization. Much of my argument will also apply with equal force, *mutatis mutandis*, against other proposed fundamental ethical principles such as rule utilitarianism, or Habermas's principle U (1990, 65), which base the test of whether any agent may perform some action, or adopt some maxim or principle, on whether it is consistent to will (or is desirable, utility-maximizing, unreasonable for inhabitants of to reject, etc.) some world in which all or most people do the same.

Maxim of Absolute Pacifism (MAP):[21]
In all situations, I will refrain from violence against another human, to promote peace.

Maxim of Divisible Helping (MDH):
When there is some set of persons in need and another set of persons who can help them, divide the amount of help needed by the number available to help (when and to the extent that this division is meaningful), and provide exactly this much help.

When I have presented these maxims to Kantians convinced that the FUL is the supreme principle of morality, I have gotten a variety of strange and clearly *ad hoc* responses. Some have proposed that these are not proper maxims, which is obviously false since they fit all the requirements Kant gives for one, as well as the tripartite maxim schema suggested by O'Neill.[22] Others have suggested that they are obviously stupid maxims and that no normal person would consider acting on them. One commentator's primary argument centered around the point that no one in England, for example, actually follows MLHD. Of course they don't; in practice people instead follow much saner maxims, like that of driving according to local convention. But that just shows that ordinary common sense is smarter, and evaluates principles according to more stringent criteria, than the FUL, for since MLHD can be willed as a universal law without contradiction the FUL cannot convict it of irrationality. I have also heard people seriously propose maxims like MAP and MDH as solutions to problems of war, poverty, or climate change, despite the reasons to think that when others can be counted on to *not* follow such maxims, a single person following them will often not only fail to achieve their ends, but may even

21 Suggested by Harrison (1985, 252). Cases of this sort have occasionally been discussed by Kantians, but never adequately as far as I know; resort to the Formula of Humanity as an End is often used to dodge the bullet here.

22 O'Neill (1989, 87) distinguishes between "specific" and "underlying" intentions, supposing that only the latter are true "maxims" subject to the FUL test. I do not find this distinction clear or compelling; the difference is relational rather than predicative, for one maxim can underlie another, but that doesn't mean that any given maxim is either specific or underlying *simpliciter*. In any case, again, both types of intention could be described in ways that conform to O'Neill's basic schema for a maxim, and so applying the name of "maxim" to some such intentions but not others is arbitrary. Finally, one can simply stipulate that some agent follows one of the above maxims not as a specific instantiation of some other intention, but as a fundamental "underlying" intention, which again will pass the FUL while being obviously immoral.

invite disaster. Following them can certainly prevent us from taking more substantial action to fix the world and achieve the obviously legitimate ends specified in these maxims.

A surprisingly frequent response is to simply ignore the counter-examples and the rest of the argument, and insist that once we accept Kant's claim that universalizability is constitutive of rational, autonomous willing, the FUL – which requires precisely that maxims be universalizable – just *must* be the supreme principle of practical reasoning and morality, in spite of the evidence to the contrary. Such responses simply show lack of attention to the distinction between the two kinds of universalizabilty. Confusing the two, or pretending that they are mutually entailing, is actually a version of the fallacy of division: assuming that if something is true of some whole, the same thing is true of any of its parts. Kant fallaciously assumed that if we can rationally will that *all* agents in a world follow a given maxim simultaneously, then we can rationally will that any agent can follow it in any situation, apart from what the others are doing. Logic assures us that this needn't be true, and the false positive cases described above confirm this.[23]

The history of scholarship on the FUL assures me that there are many desperate attempts and ad hoc strategies one could try to use to salvage the FUL.[24] But I urge my readers to resist such temptations,

23 The slide between "any" and "all" in commentators on the FUL is pervasive; for a few instances where the slide occurs on a single page of text see Engstrom (2009, 125 and 158), and Herman (1990, 170) (Herman does not say "any," but uses "unconditioned" to express the same idea). An equivalent slide from the idea that passing the FUL test is a necessary condition of a maxim's rationality to claiming that it is sufficient is made equally often, such as when Kitcher moves from claiming that an agent must consider it "possible" for everyone to follow her maxim (i.e., the latter is a necessary condition for the maxim's moral acceptability), to the claim that "the test of the moral acceptability of your action" is that "you could will that everyone follows such a law" of adopting the maxim for moral reasons," i.e. that it is a sufficient condition (2004, 571). On 578 she is even candid enough to call this shift a "trivial" inference, showing how deeply ingrained this mistake is in the Kantian tradition.

24 Another common response to similar problems is to resort to the Formula of Humanity (FH), either to replace or supplement the FUL when the latter seems to give the wrong answers. Some may think this formula works better because it embodies a substantive value not mentioned by the abstract and formal FUL. But perhaps the more crucial difference is that, unlike the FUL, it does not specify the wrong kind of universalization. It says we should *always* respect humanity as an end in itself, including respect for the rationally-chosen ends of others – not just that we must act in ways that would respect others and

and instead admit that Kant simply made a mistake. The FUL as stated is heteronomous by asking us to guide ourselves according to what would be rational to will for a magical agent who could cause others in his world to follow his lead. We should abandon it for the supervenience principle which Kant originally specified as the supreme requirement of moral and autonomous practical willing. Indeed, any other attempt to modify the FUL in the face of the problems I have given must either make its test equivalent to the requirement of supervenience, or it will necessarily fail to conform to Kant's initial desiderata for a supreme moral principle, will not succeed in unifying an agent through its practical reasoning as Korsgaard suggested it should, and will remain open to some counter-examples like the ones I have described.

4. Supervenience and Probability

Once we replace the FUL with some kind of supervenience test, some further changes in the test are also required. At a minimum, it is more obvious than ever that we cannot identify a contradiction in the will, as Kant appeared to, with willing the denial merely of any necessary means to the satisfaction of our rational ends, or to put it another way, with willing conditions which make such satisfaction impossible. For there will often be many possible combinations of persons following a certain harmful maxim which do not make the satisfaction of my rational ends completely impossible. Nor will it work to say that if there is *any* combination or number of persons following the maxim which would make it impossible for me to satisfy my rational ends, then the maxim generates a contradiction in will. Consider:

> Maxim of Random Wrongdoing (MRW):
> When I wake up each morning, I will roll six dice in a row; if they are all

their ends just if everyone else was acting as we were. This, rather than the fact that the FH specifies a substantive end, may more importantly explain why the second formulation often leads to intuitively better results, but I will not otherwise comment on the relative merits of the two formulations here. As noted earlier, the third formula, of a "kingdom of ends," has the advantage of suggesting that willing the frustration of other ends besides that of the maxim in question can generate a contradiction in will. But by requiring us to act as if total harmony with all agents' maxims had already been achieved, even when this is not so, the formula again invites logical and practical disaster.

> sixes, I will kill one person that day if I also believe I can get some personal advantage out of it.[25]

We can replace "kill one person" with "tell a lie" or any number of other pernicious activities to get a similarly bad result. There is no number of persons following MRW which would make it *impossible* for me to satisfy all my rational ends, including the end of getting occasional advantage out of MRW should its antecedent conditions for action occur. Of course, this entails that the maxim also passes the FUL test, revealing that the "possibility" condition was already far too weak under this principle, and only appeared to be a strong test for rationality if we artificially limited the choice of maxims submitted to it. In fact, many of the maxims which Kant said failed his test actually don't. It's *possible* that my lying promise would be believed in a world of universal lying, for everyone *might* be extremely stupid, or cosmic rays might hit someone's brain to make him believe my lie against all reason just for a moment. Of course these things are not *likely* to happen, and I would be irrational to trust in such remote possibilities. Likewise, I am increasingly unlikely to satisfy my rational ends the more other agents follow MRW, starting mildly with one agent doing so and getting worse from there. But this merely shows that Kantians have all along been surreptitiously relying upon unspoken assumptions as they tried to make the possibility condition bear more weight than it could logically support. We should instead openly admit that the *probability* that our ends will be satisfied, under the whole range of conditions which supervenient universality requires us to will, is a crucial factor in determining the rationality of our maxims and ends.

---

25 To any who might protest that this is not a maxim anyone would seriously follow, I submit that many people *do* follow maxims of this general form, by engaging in behavior which causes more marginal harm to the environment or economy than the gain obtained from it, but where the harm is incremental, probabilistic, or diffuse. It is not only *possible* for them to satisfy many of their ends if many or even all others do the same: this too often describes the *actual* state of affairs. Nevertheless it may be irrational to will that others follow such maxims, and hence for oneself to follow them, because our total set of rational ends could be *better* satisfied if people behaved differently. In a world where many of us use new technology which causes diffuse or probabilistic harm affecting billions of people, accumulating more than ever before, we are in desperate need of an ethics which can ground our duties to change the relevant behaviors.

This does not, of course, lead to a purely instrumental reasoning which only maximizes the probability of satisfying fixed, intuited, unchallenged, or otherwise heteronomous ends. Rather, we should check the instrumental rationality of willing that other agents (*any* others) follow the maxims we propose to act on as a test of the practical rationality and hence morality of the maxims and their ends. I have no doubt that this will still seem a radically unKantian idea to some, but there is no good reason it should. Indeed, willing the impossibility of the satisfaction of our ends is merely the limiting case of willing an increased probability of their being frustrated; impossibility simply means that such probability has reached 100 %. If the reason we thought that willing the impossibility of satisfying our ends is a problem for practical reasoning is that this is to will the frustration of our ends, then there is no reason not to consider willing the decreased probability of satisfying those same ends to be proportionately irrational.

This principle has additional benefits, for it allows us to distinguish between, on the one hand, maxims of taking advantage of a benefit obtainable by a limited number of persons, when doing so causes no or extremely minimal harm to others, and on the other hand maxims of free-riding, lying for personal gain, or otherwise taking advantage of others' weaknesses or moral scruples.[26] Both could be universalized given qualifications like: I will take the advantage as long as it's still available, otherwise not. In the latter case, those persons who lost out on the benefit can still follow the maxim vacuously, for its antecedent conditions have just not been met in their case. But the obvious moral difference between these two kinds of maxim are that any number of people acting on the first kind tends to *increase* the likelihood of agents at large satisfying their rational ends, while second kind has the reverse effect. The standard FUL questions of whether it is possible for all to act on the maxim, or whether it is possible to satisfy your ends if they did so, both fail to distinguish these two types of maxims; they are also almost completely irrelevant to the practical rationality of following them in the quite different situations we are typically in.

Engstrom[27] charged consequentialists with using an "attenuated conception of practical reasoning" when they base their own ethics on, or understand the Kantian test as involving, some "prudential or consequentialist considerations," i.e. considerations of the effect that

26 A problem which stumped both Herman (1993, 139) and Wood (1999, 106).
27 Engstrom (1993, 165).

certain actions or situations have on the probability of achieving certain ends. But it seems that it is actually Kantians who have attenuated their conception of practical reasoning, by ignoring instrumentalist considerations of probability and relative harm, and artificially restricting their supreme principle of practical reasoning to only consider certain necessary means for achieving our goals. Of course, some will wonder if this move will turn Kantian ethics into a kind of consequentialism, as Hare would have it. Perhaps it will, but I refrain from describing further normative implications of supervenience, as I trust I have already been sufficiently provocative.[28]

Bibliography

Allison, Henry (1990): Kant's Theory of Freedom, New York.
Blackburn, Simon (1993): Essays in Quasi-Realism, New York.
Broad, Charlie D. (1916): On the Function of False Hypotheses in Ethics, in: International Journal of Ethics 26, pp. 377–397.
Engstrom, Stephen (2009): The Form of Practical Knowledge: A Study of the Categorical Imperative, Cambridge, MA.
Forschler, Scott (2007): How to Make Ethical Universalization Tests Work, in: The Journal of Value Inquiry 41, pp. 31–43.
Forschler, Scott (2010): Willing Universal Law vs. Universally Lawful Willing: What Kant's Supreme Principle of Morality Should Have Been, in: Southwest Philosophy Review 26, pp. 141–152.
Frankena, William (1964): C. I. Lewis on the Ground and Nature of the Right, in: Journal of Philosophy 61, pp. 489–496.
Habermas, Jürgen (1990): Moral Consciousness and Communicative Action, Cambridge, MA.
Hare, Richard M. (1952): The Language of Morals, Oxford.
Hare, Richard M. (1989): Supervenience, in: Essays in Ethical Theory, Oxford.
Hare, Richard M. (2000): Could Kant Have Been a Utilitarian?, in: Sorting Out Ethics, Oxford, pp. 147–165 [revised from first appearance in: Utilitas 5 (1993)].
Harrison, Jonathan (1985): Utilitarianism, Universalization, Heteronomy, and Necessity, or UnKantian Ethics, in: N. T. Potter and M. Timmons (ed.): Morality and Universality: Essays on Ethical Universalizability, Boston, pp. 237–266.
Herman, Barbara (1990): Morality as Rationality: A study of Kant's ethics, New York.
Herman, Barbara (1993): The Practice of Moral Judgment, Cambridge, MA.

28 I wish to thank Dr. Ronald Glass for his feedback on earlier versions of this paper.

Kant, Immanuel (1785): Groundwork of the Metaphysics of Morals, in: Immanuel Kant: Practical Philosophy, New York. [cited as *G*]
Kant, Immanuel (1788): Critique of Practical Reason, in: Immanuel Kant: Practical Philosophy, New York. [cited as *CPrR*]
Kitcher, Patricia (2004): Kant's Argument for the Categorical Imperative, in: Nous 38, pp. 555–584.
Korsgaard, Christine (1996): Creating the Kingdom of Ends, New York.
Korsgaard, Christine (2008): The Constitution of Agency: Essays On Practical Reason and Moral Psychology, New York.
O'Neill (Nell), Onora (1975): Acting On Principle: An Essay in Kantian Ethics, New York.
O'Neill (Nell), Onora (1989): Constructions of Reason: Explorations of Kant's Moral Philosophy, New York.
Reinikainen, Jouni (2005): The Golden Rule and the Requirement of Universalizability, in: The Journal of Value Inquiry 39, pp. 155–168.
Singer, Marcus George (1961): Generalization in Ethics. An Essay in the Logic of Ethics, with the Rudiments of a System of Moral Philosophy, New York.
Smith, Michael (1994): The Moral Problem, Malden, MA.
Wattles, Jeffrey (1996): The Golden Rule, New York.
Wood, Allen W. (1999): Kant's Ethical Thought, New York.