(2012). *Southwest Philosophy Review*

# Unifying the Categorical Imperative*

Marcus Arvan
University of Tampa

> [T]he concept of freedom…constitutes the *keystone* of the whole structure of a system of pure reason…[and] this idea reveals itself through the moral law.[i]

This paper demonstrates something that Immanuel Kant notoriously claimed to be possible, but which Kant scholars today widely believe to be impossible: unification of all three formulations of the Categorical Imperative.[ii] §1 of this paper provides an intuitive reading of Kant's theory of practical reason and morality according to which the three[iii] formulations of the Categorical Imperative (the Universal Law Formulation, the Humanity Formulation, and the Kingdom of Ends Formulation) are identical. §2 then provides clear textual support for each premise in a formal argument for this Unifying Interpretation.

## §1. Kant's Theory of Morality and Pure Practical Reason

Kant argued that human beings differ from other animals in one monumental respect. Non-human animals act on their desires and inclinations. They are "pushed around the world" by whatever it is that they are inclined to do. For example, if my dog wants to go outside, he will stand in front of the door and look outside longingly. Then, if he gets himself outside and wants to come back inside, he will stand by the door and look inside longingly. Dogs and other animals, in other words, seem not to have any real choice whether to act upon their inclinations.

We human beings seem to behave very differently. We are ordinarily capable of refusing to act upon our desires or inclinations, and indeed, seem capable of willing ourselves to act independently of them. I may have a strong desire to tell a lie, but can will myself not to do so. I can say to myself, "I will not tell the lie," and then act accordingly. This capacity not only seems

to be what makes us distinctly human; it seems to be at the very root of what we admire most in good human beings. Consider a person who is tempted to lie but chooses not to because they judge that it would be wrong. We admire this person because they "overcame" their personal temptation and chose to do what is right. We admire their strength of will to do what is right. Now, of course, we might think better of the person if they were never tempted at all – due to, perhaps, training themselves not to lie – but even then, we would admire how the person consciously willed themselves to develop better inclinations, by working to become a better person.

What we admire about people, then, seems to be their capacity to rise above their animal nature – their capacity to will themselves to do the right thing, or to become better people. This is even true of things like friendship, something which Kant's theory is often said to get wrong. Consider Michael Stocker's famous criticism that Kant's theory gives a person the wrong reasons to visit a friend in the hospital.[iv] According to Stocker, Kant's theory entails that one ought to visit a sick friend in the hospital because a maxim to visit the friend would satisfy the Categorical Imperative. That, however, seems like the wrong kind of reason to visit: according to Stocker, one should visit a sick friend because one cares about them, not because some principle of action would satisfy an abstract principle such as the Categorical Imperative. Yet, I submit, this objection is a mistake, and for two reasons. First, Kant gives us the contradiction-in-willing test as a test of "imperfect duties" – duties that state the kind of person we have a duty to become (i.e. the virtues we have a duty to develop).[v] One could not will the maxim, "I will visit friends in the hospital out of duty" as a universal law of nature, as the universal law "Friends will visit each other out of duty" contradicts what we take the value of friendship to be (e.g. caring for one another, not merely acting on abstract principle). Because friends visit one another in

hospitals at least in part because they care about the sick person, the maxim, "I will visit sick friends because I care for them", could be intuitively willed as a universal law—in which case, according to Kant, it follows that we have an imperfect duty to be the kind of people who visit sick friends because we care (which is exactly what Stocker thinks we should be like).[vi] Second, I believe that Kant's theory gets at a fundamental aspect of friendship that the "caring" account of friendship leaves out. True friends care for each other, but they do not *merely* care. They choose to treat each other well *even when they do not want to*. Indeed, there is even a common saying to express this point: the saying that "true friends are there for each other no matter what." True friends, in other words, at least when they behave like true friends, choose to treat each other well *categorically*. Indeed, this is intuitively what separates true friends from "friends of convenience." Friends of convenience only care for you when it suits their interests. True friends choose to act in caring ways *categorically, even when it doesn't suit their interests*. Indeed, people often heard people say things like, "I really don't want to see so-and-so today, but I really should," or "I have to do my friendly duties," etc. This is a crucial part of what makes people true friends. True friends choose to treat each other well as a matter of principle even when they do not want to.

Kant's basic point, then – that morality is a matter of our capacity to overcome our desires and *choose* to do right by others categorically – seems intuitively right. It is this capacity that moves us when we see, for example, firefighters rush into burning buildings knowing that they may die. They don't want to die, but they *choose* to risk their lives nonetheless because they know it is right. We consider them heroes because they overcome their fear based on principle. Contrary to common objection that his theory fails to make appropriate room in moral life for things like love, friendship, and emotion, Kant's theory seems capable of getting at the

true depth of these things: the fact the best sort of love and friendship are *chosen*, and involve categorical commitments to *care* for the other even when caring is difficult or inconvenient. Nobody is perfect. Nobody – not the truest friend, not the most faithful spouse, not the courageous person – is *always* inclined to be true, faithful, or courageous. We are all beset (more often than most of us would probably like to admit) by desires and inclinations to behave badly. The thing we admire about the true friend, the truly faithful spouse, and the truly courageous person, is their choice to do the right thing even when they do not want to. The true friend chooses to "be there" for the other person, even when being there is hard. The faithful spouse chooses to remain faithful, temptations be damned. And the truly courageous person chooses to act in the face of danger, however great their fear might be.

What makes us distinctly human, then – and what makes us *moral beings* – is the capacity to will ourselves to act, not only on any inclinations we might have, but on the matter of mere principle. And what is the capacity to act on principle? To say that a person can act on a principle *despite their inclinations* is to say that they can act on the principle *as a matter of absolute law* (i.e. unconditionally). The capacity for freedom, then – "humanity" – simply *is* the capacity that makes it possible to act on laws of practical agency. Respecting humanity, then, would seem to involve respecting the capacity to act on *laws*. And that is precisely the Universal Law Formulation of the Categorical Imperative. Thus, the Humanity Formulation seems to say nothing more than the Universal Law Formulation. Acting only on universal laws of practical agency simply *is* respecting humanity. Finally, however, the capacity to act on universal laws is the capacity to act *independently of any sensible wants or inclinations*. Thus, we respect humanity not merely by acting on universal laws of practical reason; we respect humanity (and act on universal laws of practical reason) *only insofar as we act abstracting away from any*

*sensible wants or inclinations* – which is exactly what the Kingdom of Ends Formulation says.[vii] In short, respecting humanity *just is* acting on universal laws, and acting on universal laws *just is* acting in a way that abstracts away from sensible wants and inclinations. Thus, the Humanity Formulation, Universal Law Formulation, and Kingdom of Ends Formulation all seem identical. Each formula can only be properly understood and expressed in terms of the others.

**§2. The Formal Argument for the Unifying Interpretation**

Let us begin with,

(1) *The Humanity Formulation*: For Kant, our fundamental moral-practical obligation is to respect *humanity-insofar-as-it-is-capable-of-morality*.

This is, obviously, a decidedly non-standard statement of the Humanity Formulation of the Categorical Imperative. The canonical statement of the Humanity Formulation is: "*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."*[viii] However, does Kant really mean to say that humanity has unconditional value? Kant is surprisingly inconsistent here. He repeatedly insists (less than a page after giving the "canonical" statement of the Humanity Formulation) that it is *not humanity* that has dignity or unconditional value, but only *humanity-insofar-as-it-is-capable-of-morality* that has such value. For example, he writes:

> Now, *morality is the condition under which alone* a rational being can be an end in itself, since only through this it is possible to be a lawgiving member of the kingdom of ends. *Hence morality, and humanity insofar as it is capable of morality, is that which alone has dignity*.[ix]

These textual inconsistencies pose a difficult interpretive dilemma. What exactly has absolute worth for Kant: humanity, or merely humanity-*insofar-as-it-is-capable-of-morality*? For my

part, I do not think there is consistent textual support either way. Again, Kant appears to explicitly endorse both (contradictory) claims in different places. There are, however, three reasons to favor the view that, whatever Kant might have actually thought, he *should* say that it is not humanity but only humanity-insofar-as-it-is-capable-of-morality that has unconditional worth and is worthy of respect as an end-in-itself. First, it is independently implausible to suppose that bare "humanity" – which for Kant is the mere capacity to set ends – has unconditional worth, for why should we respect the capacity of the murderer or thief to set ends (given that their particular ends may be to commit murder or theft)? It is surely not the bare capacity to set ends that has moral value, but rather the capacity (even when it is not expressed) to behave morally has value (the thief and murderer still have that capacity, even when they act wrongly – and it is their *moral personhood* that intuitively deserves respect). Second, the idea that only humanity-insofar-as-it-is-capable-of-morality has unconditional value sits much better with the overall spirit of Kant's considered moral views, as Kant repeatedly emphasizes things like, "It is nothing other than [moral] *personality*…[the] capacity of being subject...[to] pure practical laws…by which alone [human beings] are ends in themselves," and,"[moral] *lawgiving* itself, which determines *all* worth, must for that very reason have a dignity, that is, an unconditional, incomparable worth; and the word *respect* alone provides a becoming expression for the estimate of it that a rational being must give."[x] These passages, and many others besides[xi], all indicate that it is not humanity *per se* but rather humanity *insofar as it gives moral laws* that warrants the respect ascribed to "humanity" by Kant's Humanity Formulation. Finally, if the present paper is correct, it is only through identifying humanity-insofar-as-it-is-capable-of-morality as the bearer of unconditional worth that we are capable of accomplishing a very important task – a task that Kant not only believed could be completed, but which we have

independent reasons to wish to complete: unification of the three formulations of the Categorical Imperative. Indeed, I submit that any interpretation that enables us to understand Kant's moral theory as unified (and intuitive) whole is clearly preferable, all things being equal, to an interpretation that leaves Kant's theory fragmented into three incommensurable formulas.

In light of this situation, I propose a unique interpretative approach. I propose that, whatever (conflicting) things Kant might have actually written about the value of humanity, this paper's argument for the unification of the Categorical Imperative is itself strong evidence that Kant ought to have taken it not to be humanity but only humanity-insofar-as-it-is-capable-of-autonomy that is of unconditional worth for Kant. Thus, for the sake of argument, I will assume my non-standard expression of the Humanity Formulation – proposition (1) – to be that formula's proper expression.

Now turn to,

(2) For Kant, humanity-insofar-as-it-is-capable-of-morality is identical to rational nature (i.e. the capacity of transcendental freedom).

The textual support for (2) is clear. All of Chapter 1 of the *Critique of Practical Reason* is devoted to showing it. For example, Kant writes, "The…question here…is whether pure reason of itself alone suffices to determine the will"[xii]; "it will not only be shown that pure reason is practical but that *it alone…is unconditionally practical*"[xiii]; "The law of causality from freedom, *that is, some pure practical rational principle*, constitutes the unavoidable beginning and determines the objects to which alone it can be referred"[xiv]; and finally, most definitively, "As a *rational being*…the human being can never think of the causality of his own will otherwise than under the idea of freedom; *for, independence from the determining causes of the world of sense (which reason must always ascribe to itself) is freedom*."[xv] Thus, (2) has clear textual support.

Now turn to,

(3) For Kant, rational nature is identical to the capacity that, when adopted, always *in fact* acts on practical principles that can function as universal laws of practical agency.

The textual support for (3) is clear. First, Kant writes, "if reason completely determined the will the action would *without fail* take place in accordance with [law]."[xvi] Then, in the most important passage of all (especially the final sentence), he writes:

> The practical use of common human reason confirms…[that] There is no one – not even the most hardened scoundrel…who, when one sets before him examples of honesty of purpose, of steadfastness in following good maxims…does not wish that he might also be so disposed. He cannot indeed bring this about in himself, though only because of his inclinations and impulses; yet at the same time he wishes to be free from such inclinations…Hence he proves, by this, *that with a will free from impulses of sensibility he transfers himself in thought into an order of things altogether different from that of his desires in the field of sensibility*….This better person…he believes himself to be when he transfers himself to the standpoint of a member of the world of understanding, as the idea of freedom, that is, of independence from determining causes of the world of sense, constrains him involuntarily to do…The moral "ought" is then his own necessary "will" as a member of an intelligible world, and is thought as "ought" only insofar as he regards himself at the same time as a member of the world of sense.[xvii]

In short, whenever we adopt the standpoint of pure practical reason – and hence (for Kant) really are transcendentally free[xviii] – we necessarily do act on laws of practical reason. Transcendental freedom is the capacity that always in fact acts on principles that could be laws of practical

action. Immorality is a failure to adopt the standpoint of pure practical reason. Thus, (3) has clear textual support.

This gives us,

(4) Thus (from 1-3), for Kant, our fundamental moral-practical obligation is to respect humanity-insofar-as-it-is-capable-of-morality, *the capacity that, when adopted, always in fact acts on principles that can function as universal laws of practical agency.*

The next premise in my argument is, as far as I can tell, never stated explicitly by Kant – but it seems obviously true:

(5) The one and only way to respect the capacity that, when adopted, always in fact acts on principles that can function as universal laws of practical agency, is to *express* that very capacity (i.e. always act on universal laws of practical agency).

Indeed, how else could one respect the capacity to always act on laws of practical agency except by in fact acting on laws of practical agency?

But in that case we have,

(6) Thus, (from 4&5 by identify), for Kant, our fundamental moral/practical obligation – to respect humanity-insofar-as-it-is-capable-of-morality as an end-in-itself [*the Humanity Formulation*] – is *identical to* acting on universal laws of practical agency [the *Universal Law Formulation*][xix]

Thus, we have,

(7) Thus (6, restated), for Kant, the Humanity Formulation of the Categorical Imperative ultimately *states nothing other* than that our fundamental moral/practical obligation is

to obey the Universal Law Formulation. [*Universal Law Formulation=Humanity Formulation*]

Now let us turn to,

(8) For Kant, always acting on universal laws of practical agency is identical to willing oneself to act on principles *independently of or abstracting away from any sensible wants or inclinations.*

Kant asserts (8) in many different places, including the following passage:

> Since the mere form of a law can be represented *only by reason*…the determining ground the will is distinct from *all determining grounds of events in nature*.[xx]

Because Kant is clear that all wants and inclinations (independent of a pure rational will, which acts on laws) are found in nature (i.e. in the sensible world)[xxi], Kant clearly affirms (8).

Premise (9) then follows by identity,

(9) Thus, (from 7&8, by identity), for Kant, our fundamental moral/practical obligation – to respect humanity-insofar-as-it-is-capable-of-morality as an end-in-itself [*the Humanity Formulation*] – is *identical to* acting on universal laws of practical agency [the *Universal Law Formulation*], which in turn is *identical to willing oneself to act on principles independently or abstracting away from any sensible wants or inclinations*.

Now turn to,

(10) For Kant, to will oneself to act on principles abstracting away from all sensible wants or inclinations is to act under the idea of a Kingdom of Ends.

The following passage demonstrates that Kant accepted (10):

> [S]ince laws determine ends in terms of their universal validity, *if we abstract away from the personal differences of rational beings as well as from all the content of their private ends* we shall be able to think of *a whole of all ends in systematic connection…that is, a kingdom of ends*…[xxii]

And so, finally, we have,

(11) Thus, (from 9&10, by identity), *The Unifying Interpretation*: for Kant, our fundamental moral-practical obligation is to

a. *Respect humanity-insofar-as-it-is-capable-of-morality* (the Humanity Formulation); which, by identity, just is to,

b. *Always act on principles that one could will to be universal laws of practical rationality* (the Universal Law Formulation); which, again by identity, just is to,

c. *Always act under the idea of a Kingdom of Ends*, i.e. on principles abstracting away from all sensible ends or inclinations (the Kingdom of Ends Formulation).

**Acknowledgements**

I would like to thank Geoff Sayre-McCord, Christopher Pines, Andrea Faggion, John Harris, and audiences at the Southwestern Philosophical Association, Mountain-Plains Philosophy Conference, Pittsburgh Area Philosophy Colloquium, and Midwest Ethics Society for their helpful comments.

**References**

Kant, Immanuel. (1797) *The Metaphysics of Morals*, in Mary J. Gregor (ed.), *Immanuel Kant: Practical Philosophy* (Cambridge: Cambridge University Press, 1996): pp. 353-604.

Kant, Immanuel. (1788) *Critique of Practical Reason*, in Mary J. Gregor (ed.), *Immanuel Kant: Practical Philosophy* (Cambridge: Cambridge University Press, 1996): pp. 133-272.

Kant, Immanuel. (1785) *Groundwork of the Metaphysics of Morals*, in Mary Gregor (ed.), *Cambridge Texts in the History of Philosophy* (Cambridge: Cambridge University Press, 1997).

Stocker, Michael. (1998) "The Schizophrenia of Modern Ethical Theories," in Roger Crisp and Michael Slote (eds.), *Virtue Ethics* (New York: Oxford University Press): pp. 66-78.

**Notes**

*I abbreviate Groundwork of the Metaphysics of Morals as G, The Metaphysics of Morals as M, and Critique of Practical Reason as C2.

[i] *C2* 5:3-4; italics added.

[ii] See *G* 4:436. Also see the Stanford Encyclopedia of Philosophy entry, "Kant's Moral Philosophy", §9, for an overview of the philosophical consensus regarding the distinctness of the three formulations.

[iii] Some readers might object that there are still other formulas I ought to discuss – for example, the so-called "Law of Nature" formula (*G* 4:421) and the "Autonomy Formula" (*G* 4:440). For reasons I cannot explain fully here, I do not believe that these are unique formulas (I believe they simply express what the other three primary formulas express). Indeed, I believe that there is relatively clear evidence at the end of section II of the Groundwork (*G* 4:436) that Kant held the three formulas I discuss here to be the (sole) three formulations of the Categorical Imperative: the formulations that express (a) the Categorical Imperative's "form" (the Universal Law Formulation), (b) its "matter" (the Humanity Formulation), and (c) its "complete determination" (the Kingdom of Ends Formulation).

[iv] Stocker (1998).

[v] *M* 6:390.

[vi] See *M* 6:399-403

[vii] *G* 4:433.

[viii] *G* 4:429.

[ix] *G* 4:435; italics added.

[x] *G* 4:436.

[xi] See *G* 4:435, 4:437. Also see *M* 5:25, where Kant writes that any "admixture" of sensible desires or inclinations to the will's lawgiving force "destroys its dignity and force."

[xii] *C2* 5:15.

[xiii] *C2* 5:15; italics added.

[xiv] *C2* 5:16; italics added.

[xv] *G* 4:452; italics added.

[xvi] *C2* 5:20.

[xvii] *G* 4:454; italics added.

[xviii] See *G* 4:448.

[xix] *C2* 5:30; *G* 4:401 and 4:421. Note: in the *Groundwork* (but not in the *Critique*) Kant sometimes says we are to act on principles that could be universal laws "of *nature*" (*G* 4:421); at other times simply "universal laws" (also *G* 4:421); and at other times "practical law[s]" (*G* 4:401). There is some question as to whether Kant's reference to universal laws *of nature* adds anything here. I will not address this issue here, as I think it is ultimately tangential to my discussion.

[xx] *C2* 5:28; italics added.

[xxi] See e.g., *C2* 5:30 and all of *G* III.

[xxii] *G* 4:433; my italics.