Unifying the Categorical Imperative, and Beyond*
Marcus Arvan

> [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.[1]

This paper demonstrates something that 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.[2] §1 of this paper explains Kant's theory of practical reason and morality at a purely intuitive level, showing how the three[3] formulations of the Categorical Imperative (the Universal Law Formulation, the Humanity Formulation, and the Kingdom of Ends Formulation) are intuitively unified. §2 then defends each premise in a formal argument for my Unifying Interpretation. §3 raises and resolves an objection to my strategy posed by Pallikathayil.[4] §4 then argues that my interpretation provides an intuitive analysis of how (and why) we should respect immoral *persons* while not respecting their immoral behavior. §5 argues that the Unifying Interpretation defended here is superior to rival interpretations of the Categorical

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

1 C2 5:3-4; italics added.

2 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.

3 Some readers might object that there are still other formulas– for example, the so-called "Law of Nature" formula (G 4:421) and the "Autonomy Formula" (G 4:440). For reasons I cannot explain here, I do not believe these to be unique formulas. I will, in any case, simply restrict my inquiry to the three formulations I discuss here, and leave questions about other formulations for elsewhere. For what it is worth, there is relatively clear evidence Kant himself held that three formulations I discuss to be, if not the only formulations of the Categorical Imperative, the most relevant to understanding its content. At the end of section II of the Groundwork (G 4:436), Kant clearly states that the Universal Law Formulation expresses the Categorical Imperative's "form," the Humanity Formulation its "matter," and the Kingdom of Ends Formulation its "complete determination." As readers will see, my Unifying Interpretation fits very well with these three descriptions.

4 Pallikathayil (2010).

Imperative. Finally, because Kant's kingdom of ends formulation is on my Unifying Interpretation the "master formulation" of the Categorical Imperative – the one formulation that *expresses* the true content of all other formulations of the Categorical Imperative – §6 briefly explores a new interpretation of the kingdom of ends formula.

An important caveat is in order before we begin. There has been a great deal of debate both about what Kant means by "humanity", and about what Kant takes to have unconditional value. For example, in his widely discussed recent book, *The Value of Humanity in Kant's Ethical Theory*, Richard Dean distinguishes between a "minimal" reading of the concept of humanity – a reading which understands humanity as the capacity to set and pursue ends – and a much more expansive which identifies humanity with a good will.[5] Then there is the closely related issue of to what Kant attaches unconditional value. In some places, Kant states that humanity is the only thing of unconditional value, and that we have a duty to respect humanity as an end in itself – as something that has "absolute worth."[6] Elsewhere, however, Kant says that it is only "humanity insofar as it is capable of morality" that has unconditional value and warrants respect.[7] These remarks seem explicitly inconsistent. Dean argues that it is possible rescue Kant from inconsistency if, and only if, we identify humanity with a good will. Dean's argument, however, has been the subject of forceful objections.[8]

This paper pursues a novel interpretative approach. Although I provide textual support for each premise in an argument for my Unifying Interpretation – showing how

---

[5] See Dean (2006).
[6] G 4:428.
[7] G 4:435.
[8] Again, see Dean (2006), Denis (2011), and Frierston (2007). Also see Wood (1999) and (2008).

Kant did in fact write things that support each of premises – this will *not* engage with other passages in Kant that do not appear to fit as well with those claims. In short, I will "pick and choose" passages from Kant's works that support my interpretation, while avoiding some passages that do not. Importantly, I hope to show that this strategy is justified in two ways: first, by showing the many things that Kant actually wrote in favor of my Unifying Interpretation are *coherent and philosophically sensible*; and second, by showing that whatever Kant might have written that does not fit with the Unifying Interpretation, the Unifying Interpretation is nevertheless *independently compelling*. My aim, in other words, is to show that whatever Kant might have actually written, there are philosophically compelling reasons to accept my Unifying Interpretation.

This approach might not appeal to some readers, particularly those most interested in Kant exegesis. Let me say a bit more, however, about why I believe it is worthwhile. Whatever Kant did say, there are compelling reasons to *seek* a unifying interpretation of the Categorical Imperative. Theoretical unity, after all, is an undeniable theoretical desideratum. Kant not only *claimed* that the formulas are unified; theoretical unity is also widely recognized to be a *virtue* in scientific and philosophical theories. It is something to aim for (all things being equal), and the reasons *why* it is something to aim for are clear. Without unity, a theory is fragmented and disconnected. For example, unless and until the formulas are unified, Kant's theory must be understood as comprising a disconnected set of three "fundamental" principles of morality. This is not a happy state of affairs, any more than physicists' inability to unify general relativity and quantum mechanics is desirable. There are, as such, strong reasons to take unification itself to have a high priority. Getting the physical world to hang together, as in physics, or getting the *moral* world to hang

together, as in ethics, are things to aim for. Thus, insofar as we care about theoretical unity at all (and few would deny that we should care about it), we should care not just about what Kant said but what he *could* have said to unify the formulas. This, at any rate, will be my approach. Although I will argue that there is textual evidence in Kant in support of my argument for the unity of the formulas, I will not try to settle once and for all that my interpretation is the only or best way to understand Kant's thought as a whole. I will instead simply hold up my interpretation as a possible way to approach Kant's ethics, draw attention to some highly attractive elements of it, and leave it to readers to decide whether it is the best way to go.

§1. Kant on Pure Practical Reason and Morality

Kant thought 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 seem not to have any choice whether to act upon their inclinations. We human beings act very differently. We are often capable of refusing to act upon our desires or inclinations, and indeed, acting independently of them. I may have a strong desire to tell a lie, but can will myself not to do so. Now, of course, one may doubt that we have the freedom to simply will things ex nihilo, absent any prior desires or inclinations. One can always attribute the phenomenon of "overcoming one's wants or inclinations" to another want or inclination. For example, we could explain my choice to

not tell the lie as me having the desire to do what is right, rather than as me acting independently of my wants and desires.

Now, it is not my aim here to evaluate or defend Kant's transcendental argument that we really do have the capacity of autonomy – the capacity to act independently of any desires, on *pure* will (which is found in *Groundwork III* and *Critique of Pure Reason*). We might also propose (as Kant seems to at some points) that it doesn't ultimately matter whether we really have autonomy – for insofar as, whenever we act, we at least seem to act "under the idea of freedom"[9], we might say that for this reason alone we must always "act under the idea" of pure practical reason (and hence, must act under the idea of the Categorical Imperative, even if we're not really transcendentally free). For the purposes of our inquiry, however, let us assume that we really do have the capacity of pure will as Kant claims.

Let us return to the capacity in question: our distinctly human capacity to overcome our inclinations and act from (what appears to be) pure will. 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 does not because they see that it would be wrong. We admire this person because they "overcame" their personal temptation and did the right thing on principle. 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 chose to develop better inclinations (and once again, we would admire their decision and strength of will to work at becoming a better person).

[9] G 4:447.

Indeed, what we admire about people seems to be their capacity to rise above their animal nature – their capacity to will themselves 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.[10] 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 pass the Categorical Imperative. That, however, seems like the wrong reason to visit. Yet this is a bad way to think about Kant, 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).[11] 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. Friends visit one another in hospitals at least in part because they *care* about the sick person. Thus, presumably Kant would say that we have an imperfect duty to develop caring attitudes towards others (which is exactly what he did say in the *Metaphysics of Morals*).[12] Second, friendship does not seem reducible only to caring. True friends, goes the common saying, are "there for each other no matter what." In a word, they are there for each other categorically. This is what separates true friends from "friends of convenience." True friends are not perfect. They do not always want to be there for you when you need them. But they are there for you regardless — they commit themselves to you as a matter of principle, no matter what, because that is what true

[10] Stocker (1998): 66 – 78.
[11] M 6:390.
[12] See M 6:399-403

friendship is. Indeed, people often heard people say things like, "I really don't want to hang out with so-and-so – but I really should," or "I have to do my friendly duties," etc. Rather than looking down on people who behave as a friend even when they do not want to, we admire them. We think that is one of the things makes them true friends.

Kant's basic point, then – that our capacity to act freely and overcome our desires is what gives life moral value – seems exactly right. It is the capacity that moves us when we see, for example, firefighters rush into burning buildings knowing that they may die. They do not want to die, but they risk their lives nonetheless because they know it is right. We consider them heroes because they overcome their fear based on principle. Contrast this example against a poor sod who signs up for military service not out of principled understanding but an impulsive nationalistic sentiment or desire for danger. This person does not impress us. We pity them and see them as mere "sheep" to the slaughter. In short, we admire people with the willpower to do the right thing out of understanding and principle. About this much, Kant seems right. His detractors (like Stocker) misunderstand the true depth of his theory. His theory explains *precisely* what we value about genuine acts of love, friendship, and courage. Nobody – not the truest friend, not the most faithful spouse, not the courageous person – is *always* inclined to be true, faithful, or courageous. Nobody is that virtuous. 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 *categorical choice* to do the right thing even when they don't want to. The true friend *chooses* to "be there" for the other person *categorically*, even when being there is hard. The faithful spouse *chooses*

to remain faithful *categorically*, temptations be damned. And the truly courageous person *chooses* to act in the face of danger, *categorically*, however great their fear might be.

What makes us distinctly "human," then, 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. 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.[13] 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 seem unified. Each formula can only be properly understood in terms of the others. All

[13] G 4:433.

three formulations really are, "at bottom only so many formulae of the very same law, and any one of them unites the other two in it."[14]

§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. 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."*[15] But does Kant really mean to say that humanity has unconditional value? Again, Kant is inconsistent about the value he ascribes to humanity. Indeed, 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:

> 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*.[16]

Again, 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,

---

[14] G 4:436.
[15] G 4:429.
[16] G 4:435; italics added.

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[17] – 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 *be a moral agent* that 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."[18] These passages, and many others besides[19], all indicate that it is not humanity *per se* but humanity *insofar as it gives moral laws* that warrants the respect ascribed to "humanity" by Kant's Humanity

[17] See *G* 4:437 and *M* 6:392
[18] G 4:436.
[19] 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."

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. 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"[20]; "it will not only be shown that pure reason is practical but that *it alone…is unconditionally practical*"[21]; "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"[22]; 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*."[23] Thus, (2) has clear textual support.

---

[20] C2 5:15.
[21] C2 5:15; italics added.
[22] C2 5:16; italics added.
[23] G 4:452; italics added.

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]."[24] 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.[25]

[24] C2 5:20.
[25] G 4:454; italics added.

In short, whenever we adopt the standpoint of pure practical reason – and hence (for Kant) really are transcendentally free[26] – 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 philosophically sensible:

(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*? Consider again my capacity to choose to "*be there for a friend"* categorically. What would it *be* for to respect that capacity – and indeed, *any* person's capacity – to on such principles categorically? Intuitively, the only way to *respect* the capacity to act on universal laws is to simply *do* it: act on, and only, universal laws.

But now if that is the case, then we have,

[26] See G 4:448.

(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*][27]

Accordingly, we also have,

(7) Thus (6, restated), for Kant, the Humanity Formulation of the Categorical Imperative ultimately *states nothing more or less* than that our fundamental moral/practical obligation is to obey the Universal Law Formulation. [*Universal Law Formulation=Humanity Formulation*]

Then turn to (8),

(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*.[28]

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)[29], Kant clearly affirms (8).

---

[27] 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.

[28] C2 5:28; italics added.

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*…[30]

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,

---

[29] See e.g., C2 5:30 and all of G III.
[30] G 4:433; my italics.

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).

This argument is deductively valid. It also has, as we have just seen, clear philosophical and textual support. It may not be the *only* to interpret Kant, but again, these matters are tertiary to my main concern. My concern is to see if the Categorical Imperative *can* be sensibly unified. Whatever else Kant might have said, all of the passages and ideas that I have appealed to in my argument are philosophically sensible and coherent. Perhaps other things Kant actually wrote stand in the way of interpreting him as (consistently) attached to this interpretation (since, as one reviewer put it, there are many places in which Kant treats the Universal Law and Humanity Formulations as "the" Categorical Imperative, whereas I take the Kingdom of Ends to specify what those two formulas amount to). Understanding Kant is of course of immense importance. But so too is understanding the Categorical Imperative, as an idea, on its own terms – and I believe we have seen that, whatever else Kant might have written, there is a coherent and philosophically sensible way to unify it.

§3. Response to Pallikathayil's Objection

Pallikathayil briefly considers, but rejects, a central aspect of my Unifying Interpretation: the idea that the Humanity Formulation should be understood in terms of the Universal Law Formulation. Pallikathayil writes,

> [I]n order to make the possible rational consent interpretation of what it is to treat someone merely as a means coherent, we need to identify some other rational norm that could be used to give the idea of rational consent content. This would be difficult to do within the Kantian framework. Kant identifies two kinds of norms of practical rationality: categorical imperatives, which apply to us unconditionally, and hypothetical imperatives, which apply to us in virtue of having adopting a certain end. As we will see, neither...can be used to give content…[to] the Formula of Humanity.[31]

> Kant emphasizes the claim that there is only one categorical imperative…For this reason, as an interpretative matter it would be difficult to treat another formulation of the Categorical Imperative, for example, the Formula of Universal Law, as a rational norm that is *independent* of the Formula of Humanity in the way needed to avoid…*circularity*…Substantively, this strategy would [also] face questions regarding how to justify privileging one of the formulations in this way.[32]

However, neither of these reasons for rejecting my proposal are sound. Consider first the circularity worry. On my interpretation, Kant avoids circularity because there is, at bottom, only one idea animating all three formulations of the Categorical Imperative: Kant's notion of transcendental freedom. For Kant, practical rationality – humanity-insofar-as-it-is-capable-of-morality – is identical to transcendental freedom. Since transcendental freedom in turn is identical to the capacity to act on universalizable principles, respecting humanity

[31] Pallikathayil (2010): 121.
[32] Ibid.

is then identical to acting on universalizable principles. There is, therefore, no circularity in understanding the Humanity Formulation in terms of the Universal Law Formulation. The Humanity Formulation and Universal Law Formulation are not distinct norms at all (as Pallikathayil suggests). They are identical norms.[33]

Now turn to Pallikathayil's substantive worry – the one about how we might justify "privileging" the Universal Law Formulation. Here again, Pallikathayil's error is in thinking the Humanity Formulation and Universal Law Formulation are distinct. Understanding the Humanity Formulation through the Universal Law Formulation does not privilege the Universal Law Formulation: it simply tells us (by identity) what the Humanity Formulation is. Thus, Pallikathayil's worries about my strategy are misconceived.

§4. The Unifying Interpretation and Respect for Criminals[34]

Richard Dean has argued that in order to rescue Kant's claims about humanity and unconditional value from inconsistency, we must identify humanity with a good will.[35] This move, however, has encountered powerful objections – in particular the objection that it denies unconditional worth to criminals and other immoral people who fail to realize a good will.[36] Allen Wood, in particular, has forcefully argued that no interpretation of Kant

---

[33] This, I submit, is what Kant meant when he wrote that, "[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." (C2 5:3-4; italics added) A keystone, in architecture, is the stone at the top of an arch. Without it, the arch will collapse. My interpretation (and only my interpretation) makes good on this claim. Transcendental freedom unifies all three formulations because respecting transcendental freedom (i.e. respecting our humanity-insofar-as-it-is-capable-of-morality) *just is* acting only on universal laws, which in turn *just are* principles that could be willed abstracting away from all contingent ends (in a kingdom of ends).

[34] This section responds to concerns raised by an anonymous referee.

[35] Dean (2006).

[36] See e.g. Wood (2008) and Frierson (2007).

can fly that denies that the criminal must be treated morally.[37] How does my interpretation handle this issue?

Notice that I have not identified humanity with a good will. I have distinguished humanity from humanity-insofar-as-it-is-capable-of-morality, where the latter is the capacity which, when adopted, realizes a good will. This is crucial because, as we will now see, my interpretation has wholly intuitive implications about criminals: it entails that they are due a certain kind of unconditional respect, but that their immoral behavior is unworthy of respect (both of which seem like common sense).

Now, at first glance, my Unifying Interpretation might appear to entail that criminals do not have unconditional value or warrant respect. After all, I have argued that it is only (a) humanity-insofar-as-it-is-capable-of-morality that has unconditional value, and that (b) humanity-insofar-as-it-is-capable-of-morality is the capacity that (when adopted) always in fact acts on universalizable maxims. These two claims appear to jointly entail that human beings have unconditional value only when they in fact act on universalizable maxims. In fact, they really do entail this – yet, on the Unifying Interpretation, this does not mean that criminals are unworthy of respect or lack unconditional value. Let me explain why.

The Unifying Interpretation states that to respect humanity(-insofar-as-it-is-capable-of-morality) just is to act on univeralizable maxims, but that universalizable maxims in turn just are maxims that could be willed abstracting away from all contingent ends (as members and subjects of a kingdom of ends). This, however, is just to say that all human beings, even criminals, have the same unconditional value and are worthy of the very same respect: all are worthy of, and are to be treated according to, maxims that could

[37] See Wood (1999).

be willed abstracting away from all contingent ends. Thus, even when the criminal fails to act in this way (i.e. morally), he/she is still worthy of exactly the same respect as a person with a good will. Both are to be treated according to principles that all could will abstracting away from all contingent ends (as members and subjects of a kingdom of ends). Now, of course, this might not seem very helpful, because now we have to ask: what is it to treat all human beings in that way? Indeed, this brings us back to an utterly unique feature of the Unifying Interpretation, its claim that the Kingdom of Ends formulation is the "master" formulation that specifies (by identity) what the other two formulations amount to. On the Unifying Interpretation, we cannot state more clearly what unconditional value all human beings have, and what sort of respect all are due as human beings, until we arrive at a clear interpretation of the Kingdom of Ends Formulation. The Unifying Interpretation entails that the other two formulations of the Categorical Imperative simply do not have any clear sense except through the Kingdom of Ends Formulation. This is why the present paper cannot simply end with the Unifying Interpretation. Until we say clearly what the Kingdom of Ends Formulation comes to, we cannot really say what *any* formulation of the Categorical Imperative comes to. We cannot say what sort of unconditional value *any* human being has (criminal or otherwise), or what sort of respect any human being is due, until we properly interpret the Kingdom of Ends Formulation.

§5. Comparing the Unifying Interpretation to Alternative Interpretations

The most common interpretation of the Humanity Formulation, due to Korsgaard and O'Neill, holds that we treat humanity as an end-in-itself when and only when we treat

people according to principles to which they could possibly consent.[38] Call this the Possible Consent Interpretation (PCI). Japa Pallikkathayil has argued persuasively that PCI fails both as an interpretation of Kant, but also as a moral view (having counterintuitive implications so severe as to be "fundamentally flawed").[39] Pallikkathayil then defends a very different interpretation of the Humanity Formulation, according to which we must understand respect for humanity through Kant's concept of equal external freedom, which ultimately involves Kant's political philosophy (and so, in order to understand the Humanity Formulation, we must move to political theory). Call this the Equal External Freedom Interpretation (EEFI).

Pallikathayil has already shown, in my estimation, that Korsgaard and O'Neill's Possible Consent Interpretation (PCI) of the Humanity Formulation cannot be correct. Here, though, is a deeper reason why PCI cannot be correct. Kant says explicitly that morality is a matter of acting according to principles that all could will abstracting away from all personal differences of rational beings and the content of their private ends.[40] The PCI, as such, is simply inconsistent with what Kant explicitly wrote. Morality, for Kant, is not about treating people according to principles they could consent to; it is about treating them according to principles they could consent to abstracting away from all differences between rational beings. That is a very different claim, and Kant explicitly makes it. Thus, PCI is false, at least as an interpretation of Kant.

---

[38] See Korsgaard (1996): 106–32, 137–40, 295–96; and O'Neill (1989): 105–25.
[39] See Pallikathayil (2010), esp. pp. 116-125.
[40] G 4:433.

Next, here is why Pallikathayil's Equal External Freedom analysis of the Humanity Formulation – the claim that we respect humanity by *respecting equal external freedom* – is incorrect. Pallikathayil writes,

> In his moral philosophy, one of the things that Kant is most concerned to argue is that internal freedom—that is, autonomy—is achieved through following the Categorical Imperative. *In what follows, I am not concerned with that part of his project*. We are simply going to take the Formula of Humanity as our starting point. The intuitive thought is that, if the value of humanity can be used to generate requirements on how one affects the external freedom of others, it will make sense to regard the violation of these requirements as treating someone merely as a means: violations will limit the ability of others to engage in self-directed action and, in that sense, will involve directing others rather than allowing them to direct themselves. And this seems to be a way of treating someone as on a par with a mere tool.[41]

Here is the problem with this. We cannot specify what equal external freedom is until we understand *internal freedom*. Kant thinks that we must always act as legislators and subjects to a Kingdom of Ends – but the idea of a kingdom of ends is intrinsically tied to internal freedom. In order to know what a Kingdom of Ends is, we must, "abstract away from the personal differences of rational beings as well as from the content of their private ends." (G 4:434) The very idea of "equal external freedom" presupposes an analysis of the Kingdom of Ends. We cannot know what equal external freedom is until we perform the task of abstracting away from private ends and imagine all ends in systematic connection.

[41] Ibid: 133 (my italics).

Thus, Pallikathayil's Equal External Freedom Interpretation will not do. One cannot know what equal external freedom amounts to without in turn interpreting the Kingdom of Ends Formulation, which is in turn a difficult question in its own right. It is unclear, for example, whether Kant interpreted the kingdom of ends correctly in his political theory. Rawls, for one, gives a very different interpretation of the idea of a kingdom of ends in his theory of justice as fairness.[42] For my part, I think Rawls' interpretation is more correct. In any case, one cannot simply appeal to Kant's political theory, as Pallikathayil does. One must ask whether Kant's theory interprets the Categorical Imperative correctly – which is just to say that Pallikathayil has the cart before the horse.

§6. A Tentative Proposal for a New Interpretation of the Kingdom of Ends Formulation: Fixing Rawls' Error, and Toward a Morality of Fair Compromise

Let us now investigate the idea of a kingdom of ends ourselves (since, on the Unifying Interpretation, it is the "master" principle that expresses the true meaning of Universal Law and Humanity formulations). I cannot, for reasons of space, enter into here a detailed summary and discussion of existing interpretations of the Kingdom of Ends formulation. Instead, I would like to say some new things about one (widely criticized) interpretation of the Kingdom of Ends formula: John Rawls' claim that his "original position" embodies it.[43]

Rawls' claim that the original position embodies the kingdom of ends has been widely criticized. Flikschuh, for example, argues that Kant's idea of a kingdom of ends is fundamentally metaphysical, not political (as Rawls' original position interprets it).[44] In response, I want to introduce a new criticism of Rawls, and argue that once we address this

[42] Rawls (1971): §40.
[43] See Rawls (1999): §40.
[44] See e.g., Flikschuh (2009).

new criticism properly, the result is that something relatively close to Rawls' original position – a distinctly non-political version of it (that should address Flikschuh's worries) – really does embody the idea of a kingdom of ends.

We have seen that Kant identifies a kingdom of ends as a systematic union of ends arrived at through abstraction away from the content of all contingent ends. Rawls argues that his position embodies this idea insofar as its "veil of ignorance" embodies an abstraction away from contingent ends.[45] Rawls believes that insofar as the veil of ignorance withholds from citizens all potentially "self-individuating" information – knowledge of their race, gender, social class, talents, particular ends, etc. – the original position models a class of purely rational agents who cannot deliberate based on their contingent ends.[46] At least offhand, this does sound a lot like a model of Kant's kingdom of ends.

Flikschuh maintains that in understanding the kingdom of ends in this kind of political matter – as pertaining to citizens – Rawls runs afoul of the fact that Kant's idea of a kingdom of ends is fundamentally metaphysical, not political. I agree. Rawls' original position contains a fundamental mistake, at least as a model of a kingdom of ends: it does not actually abstract away from the content of all contingent ends. Here is why: Rawls assumed for the sake of constructing a theory of justice – that is, for the sake of political theory – that all citizens in society are committed to a higher-order, contingent common end: the mutual cooperation on fair, reciprocal grounds.[47] Notice too that this is precisely the place in Rawls' theory to which anti-Rawlsians object the most vehemently.

[45] Again, see Rawls (1999): §40.
[46] Ibid: §4.
[47] Ibid:

Libertarians like Nozick, for example, ask: who has a right to make me cooperate with others on "fair" grounds?[48] According to libertarians, we cannot simply assume in political theory that people will or even ought to interact reciprocally on fair grounds (whether a person wants to cooperate with someone else should be their right to decide!). And so, it seems, Rawls not only made a suspicious move in political philosophy; he made an assumption that seems directly opposed to the Kantian idea of a kingdom of ends. A proper model of the kingdom of ends would not simply abstract away from most contingent ends, assuming all the while that people share a contingent end (i.e. fair cooperation). No, a proper model of the kingdom of ends should abstract away from all ends. Flikschuh, then, is perfectly right. Rawls' original position fails as a model of the kingdom of ends precisely insofar as it assumes the end of political cooperation.

But now what if we were to correct for this error, and re-imagine the original position not as a political model but as a metaphysical moral model that truly did abstract away from all contingent ends (just as the Kingdom of Ends formulation requires)? What would such a revised original position look like? Would it plausibly model Kant's metaphysical idea of a kingdom of ends? We now turn briefly to this question.

Although Kant is clear that the kingdom of ends is supposed to be a systematic union of all ends (arrived at through abstracting away from all contingent ends), I believe it may be more fruitful, for the time being, to apply our "revised original position" – one in which no contingent ends are assumed from behind the veil of ignorance – to two individual persons considered in isolation. My reason for beginning in this way has to do with matters of simplicity, clarity, and "intuitiveness." If we can determine what a proper

[48] See Nozick (1974).

model of a "kingdom of ends" would look like for two people, we could then abstract away from there and apply the same model to all possible people (thus modeling a true, systematic union of all possible ends). This, anyway, is the hope. Let us proceed.

Consider two ordinary people behind an "absolute" veil of ignorance, one that withholds from each of them all information related to contingent ends. These two people, then, do not even know if they want to cooperate with one another. Second, in order to make the example intuitive, let us describe what each of these two people are actually like. Let us suppose, on the one hand, that we have a Racist-Restaurant-Owner here on the one side and, on the other hand, a Person-of-Discriminated-Against-Race here on the other. We will suppose, of course, that Racist-Restaurant-Owner has the contingent end of not selling food to the Person-of-Discriminated-Against-Race, and that the Person-of-Discriminated-Against-Race has as a contingent end to eat at Racist-Restaurant-Owner's restaurant. Does the Racist-Restaurant-Owner have a moral obligation to let Person-of-Discriminated-Race dine at his restaurant? I assume that everyone reading this paper will agree that the Racist-Restaurant-Owner does have such an obligation, and indeed, that it is wrong of the Racist-Restaurant-Owner not to serve Person-of-Discriminated-Against-Race. Finally, let us return to back behind the "absolute" veil of ignorance. Suppose you were behind a veil of ignorance and you did not know whose contingent ends were yours (i.e. you do not know whether you are the Racist-Restaurant-Owner or Person-of-Discriminated-Against-Race). In that case you would not know any of your own contingent ends. The absolute veil of ignorance requires you to deliberate in complete abstraction away from any knowledge of which ends are (contingently) yours. Well, then, behind this veil of ignorance – one that, finally, appears to embody Kant's idea of a kingdom of ends (as abstracting away

from all contingent ends), applied at least to two individuals – what would you choose? How would you deliberate? Here is an answer: you would have to ask yourself, "What would be worse, (A) turning out to be a racist who has to serve someone of a race he doesn't like (side with Racist-Restaurant-Owner), or (B) turning out to be someone who cannot eat at a restaurant due to racial discrimination?" Offhand, it seems obvious that the latter is much, much worse for anyone. Although having to serve someone of a particular race may offend a racist, not being able to eat at a restaurant due to discrimination is a serious and lasting assault on the self-respect of the person discriminated against. It thus seems that if we apply our revised original position to this case (as a model of Kant's kingdom of ends), we get the right moral answer, which is that (from behind the veil of ignorance), all persons should agree, whatever their contingent ends, that it is wrong for the Racist-Restaurant-Owner not to serve the Person-of-Discriminated-Against-Race.

Now, of course, I only teased out this example at a highly intuitive level. Making the analysis perfectly rigorous will probably turn out to be very difficult. Here, though, is the important thing: the revised original position not only (i) appeared to properly model the idea of a kingdom of ends (as applied to two isolated individuals); it also (ii) appeared to give the right moral answer. If these are not strong marks in favor of the revised original position I have proposed, I do not know what is.

The revised original position also has tantalizingly illuminating offhand implications for what is perhaps the most difficult and longstanding problem in all of moral and political philosophy: the problem of fundamental disagreement. Consider, for example, debates between utilitarians and Kantians in moral philosophy, or debates between Rawlsians and libertarians within political philosophy. It is no secret that these "debates" have ended in

something very close to a stalemate. Almost nothing that utilitarians say in arguing for utilitarianism seems to convince Kantians that they are wrong, and conversely, almost nothing that Kantians say seems to convince utilitarians that they are wrong. Similarly, almost nothing that Rawlsians say seems capable of convincing libertarians that they are wrong, and nothing that libertarians say seems capable of convincing Rawlsians that they are wrong. These and other such stalemates seem to be the result of different people having fundamentally different moral intuitions. Indeed, when it comes to Rawlsians and libertarians in particular, both sides typically begin from very different premises.[49]

I believe that our revised original position embodies a plausibly Kantian idea of what it is to respect the humanity in ourselves and others when it comes to these sorts of disagreements. For let us reflect, first offhand, about what humanity is for Kant. If my Unifying Interpretation is correct, humanity (insofar as it is capable of morality) is simply the capacity to act abstracting away from one's contingent ends. Now consider Rawlsians and libertarians. Is it not the case that they both think their favored views do something like this? The answer is clearly yes. Indeed, Rawls and Nozick both argue that their respective approaches to political theory embody Kant's idea of respecting humanity. But now if two people fundamentally disagree on what respecting humanity amounts to, my Unifying Interpretation along with the revised original position gives us a model of what it is for two people who fundamentally disagree to respect the other's humanity(insofar-as-it-is-capable-of-morality). In order to model respect for humanity(insofar-as-it-is-capable-of-morality) as applied to fundamental moral disagreement, my model says that we must

[49] Compare e.g. Nozick's (1974) foundational premises versus Rawls' (1999) acceptance of the method of "reflective equilibrium."

imagine (e.g.) a committed Rawlsian and a committed libertarian behind an "absolute" veil of ignorance. We will now see that this model provides a fascinating and plausible new analysis of what it is for people who fundamentally disagree to respect one another.

Suppose you were behind an absolute veil of ignorance and did not know whether you would turn out to be a liberal-egalitarian, a libertarian, a Marxist, etc. In this case, you would know that regardless of who you turn out to be, you will think your favored political theory is more justified than the rival theories you reject. But now of course, since you are behind an absolute veil of ignorance, how can you possibly choose? Because the different theories are mutually incompatible, you cannot simply say, "To each their own." You must decide between them. But deciding in favor of any one of them would potentially leave you the one whose favored theory is rejected. Clearly you don't want that outcome — so what are you to do? You must require all parties to compromise equally. You must ask the hard-core libertarian to give up his/her hard-core libertarian political theory in favor of a more moderate position because, for all you know, you will turn out to be a liberal-egalitarian or Marxist. Similarly, you must ask the hard-core liberal-egalitarian and Marxist to give up their fundamental beliefs in favor of a more moderate position, because for all you know, you will turn out to have libertarian sentiments. The end-result, then, is an ethic of fair compromise. Our new model of the Kingdom of Ends Formulation seems to entail that whenever there is true, fundamental moral disagreement (between people who are all equally committed to moral equality, i.e., unlike the racist restaurateur, who has an obligation on our model to not be racist), we respect each other through compromise. This is, I submit, a plausible and enlightening view. Indeed, I for one have always wondered: how do we respect one another if we insist upon the correctness of our own

Rawlsian/libertarian/etc. views when other reasonable human beings fundamentally have a different point of view? Our new model of the Kingdom of Ends answers this question in the negative. It says that we respect one another on matters of fundamental (and genuine) moral disagreement (i.e. disagreement not clouded by, e.g., racist tendencies, but rather a common commitment to moral quality) if and only if all disagreeing parties compromise.

Our new model of the Kingdom of ends – the "absolute" original position – thus has some very strong attractions. It not only seems to (A) get the Kingdom of Ends formulation right where Rawls got it wrong; it also (B) seems to give the right kinds of moral answers to cases like that of the Racist-Restaurant-Owner (i.e. racism is wrong), and finally (C) it gives a tantalizing, distinctive new moral analysis of how we ought to respond to fundamental, genuine moral disagreement. Although the absolute original position obviously needs a great deal of further philosophical defense and elaboration – which, unfortunately, due to space constraints, is impossible here – I submit that these three prima facie results show that the "absolute" original position deserves to be taken seriously, not just as a model of the Kingdom of Ends formulation, but as a model of morality more generally.

§7. Conclusion

If my Unifying Interpretation of the Categorical Imperative is correct (as I believe it has been shown to be), the Kingdom of Ends formulation is the "master formulation" of the Categorical Imperative. It tells us what the other two formulations come to. We cannot understand the Universal Law Formulation or Humanity Formulation at all in isolation (which has been the dominant approach); we can only understand them properly in terms of the Kingdom of Ends Formulation. We can, therefore, understand the true normative import of Kant's ethical system when, and only when, we properly understand the Kingdom

of Ends Formulation. This, however, raises an obvious problem. It is commonly admitted that the Kingdom of Ends Formulation is the least well-understood formulation of the Categorical Imperative.[50] Although there are some well-known interpretations of the Kingdom of Ends formulation[51], its proper interpretation is still very much open to debate. Finally, I have tried to add something new and distinctive to this debate, showing how, in my view, Rawls' fundamental error in taking his original position to model Kant's kingdom of ends – namely, Rawls' affirmation of a common contingent end (that of fair social cooperation) – can be corrected. I argued that an "absolute" original position – one that truly abstracts away from all contingent ends – not only appears to properly model the idea of a kingdom of ends, but also, that this new original position has tantalizingly plausible implications as a model of morality (particularly in its moral analysis of fundamental moral disagreement).

[50] See e.g. the Stanford Encyclopedia of Philosophy entry, "Kant's Moral Philosophy", §8.
[51] See Rawls (1971) and Hill (1992).

References

Dean, Richard (2006). *The Value of Humanity in Kant's Moral Theory* (Oxford University Press).

Denis, Lara (2011). "Humanity, Obligation, and the Good Will: An Argument Against Dean's Interpretation of Humanity." *Kantian Review*. 15, 1: 118-141.

Flikschuh, Katrin (2009). "Kant's kingdom of ends: metaphysical, not political**,"** in: Timmermann, Jens, (ed*.) Kant's 'Groundwork of the metaphysics of morals': a critical guide*. Cambridge University Press, Cambridge, UK.

Frierson, Patrick R. (2007). "Review of *The Value of Humanity in Kant's Moral Theory* by Richard Dean", *Notre Dame Philosophical Reviews*, http://ndpr.nd.edu/review.cfm?id=9364.

Hill, Thomas (1992). *Dignity and Practical Reason in Kant's Moral Theory*. (Ithaca: Cornell University Press).

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).

Korsgaard, Christine (1996). *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press).

Nozick, Robert (1974). *Anarchy, State, and Utopia*. (Basic books).

O'Neill, Onora (1989). "Between Consenting Adults," in *Constructions of Reason* (Cambridge: Cambridge University Press).

Pallikkathayil, Japa (2010). "Deriving Morality from Politics: Rethinking the Formula of Humanity", *Ethics*.

Rawls, John (1971). *A Theory of Justice*. (Cambridge, MA: The Belknap Press of Harvard University Press).

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

Wood, Allen W. (1999). *Kant's Ethical Thought* (Cambridge: Cambridge University Press).

Wood, Allen W. (2008). *Kantian Ethics* (Cambridge: Cambridge University Press).