ON THE UNIVERSAL LAW AND HUMANITY FORMULAS

by

Sven R. Nyholm

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Philosophy)
in the University of Michigan
2012

Doctoral Committee:

Professor Elizabeth S. Anderson, Co-Chair
Associate Professor Sarah Buss, Co-Chair
Professor Peter A. Railton
Professor Donald H. Regan

To my grandmother

ACKNOWLEDGEMENTS

I would first like to thank my committee – Liz Anderson, Sarah Buss, Peter Railton, and Don Regan – for their support. During my time at Michigan I have also benefited greatly from the course work I did – especially under Allan Gibbard, Stephen Darwall, and Victor Caston – so I would like to thank my other teachers at Michigan, too. My approach to the study of the history of philosophy is very strongly influenced by what I learned in some of these courses, but also greatly inspired by the Ancient Philosophy reading group whose concern with the proper and precise translation of the great historical works has served as a model for me in my engagement with Kant's texts.

In the preparation of these chapters I have been particularly helped by comments on an earlier draft by Liz Anderson, and also by the discussions about Kant's ethics I had with Sarah Buss in Zürich in the summer of 2011. I was also particularly helped by the feedback I received by Peter Railton, Rich Thomason, and Gordon Belot when I presented parts of chapter two as a practice job talk in September of last year.

Since I wrote these chapters away from Ann Arbor – in Berlin and Stupferich in Germany, Borgholm in Sweden, and Durham, NC – I haven't had a chance to discuss their content with my fellow Michigan graduate students. But during my time at Michigan I have certainly learned a lot from Steve Campbell, Jason Konek, Nathaniel Coleman, Dave Wiens, Lina Jansson, and all the others. Finally I would also like to thank Katharina Uhde, her family, and my own family.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

The translations of passages from the works whose abbreviations in the footnotes are listed below are all mine. When I have translated the given passages I have consulted Mary J. Gregor and Werner S Pluhar's English translations of Kant's works as well as Jeanette Emt, Fredrik Linde, and Joachim Retzlaff's Swedish translations. The German versions I have used are all from the widely available *Suhrkamp Taschenbuch Wissenschaft Immanuel Kant Werkausgabe in 12 Bänden* edition whose editor is Willhelm Weischedel. In what follows I cite Kant's works by giving the abbreviations below (if and only if the works in question are among those for which I use my own translations) and the volume numbers and page numbers in the *Immanuel Kant: Gesammelte Schriften (Akademie-Ausgabe)*, as is customary, except for in the case of the first *Critique*, where I give the page numbers in the original first and second editions (which is also fairly customary). Here, in the order in which the works were published, are my abbreviations along with the original German titles and, within brackets, the English titles most often used:

KrV: *Kritik der reinen Vernuft* (*Critique of Pure Reason*)

P: *Prolegomena zu jeden künftigen Metaphysik* (*Prolegomena to any Future Metaphysics*)

G: *Grundlegung zur Metaphysik der Sitten* (*Groundwork for the Metaphysics*

	of Morals)
MN:	*Metaphysische Anfangsgründe der Naturwissenschaft* (*Metaphysical Foundations of Natural Science*)
KpV:	*Kritik der Praktischen Vernuft* (*Critique of Practical Reason*)
KU:	*Kritik der Urteilskraft* (*Critique of the Power of Judgment*)
R:	*Die Religion Innerhalb der Grenzen der bloßen Vernuft* (*Religion within the Limits of Reason Alone*)
TP:	*Über den Gemeinspruch: Das mag in der Theorie richtig sein, taugt aber nicht für die Praxis* (*Concerning the Common Saying: That May Work in Theory, though Not in Practice*)
MS:	*Die Metaphysik der Sitten* (*The Metaphysics of Morals*)
ÜVR:	*Über ein Vermeintes Recht aus Menschenliebe zu Lügen* (*On a Supposed Right to Lie out of Benevolence*)
AP:	*Anthropologie in pragmatischer Hinsicht* (*Anthropology from a Pragmatic Point of View*)

CHAPTER I

Introduction: The Human Nature Formula

1. Why Yet Another Dissertation on How to Interpret Kant's Ethical Theory?

Our topic is how to understand Immanuel Kant's "universal law" and "humanity" formulations of the categorical imperative, and the exact relation between the two. The former reads (in the formulation used in the *Metaphysics of Morals*) "act on the basis of a maxim that could hold as a universal law"[1]; the second "so act that you treat the humanity in your own person, as well as in every other person, always at the same time as an end, and never as a means only."[2]

Since my choice of topic is in itself everything but groundbreaking, the first thing I need to do is to motivate this choice. I shall limit myself to offering two main motivating reasons. Before doing so I will, however, first note another limitation I have placed myself under: namely, to engage exclusively with the discussion of Kant's ethics within Anglophone moral philosophy, and in particular fairly recent contributions to this discussion. This leads directly to the first motivation behind the discussion that follows.

Firstly, although Kant's ethical theory is very widely discussed and often severely criticized both by those who sympathize with Kant and those who don't, Kant's ethical

[1] (MS: 6:225)
[2] (G: 4:429)

theory is *not*, I believe, well understood within contemporary moral philosophy. This means that many of the criticisms that are directed at Kant's ethical system misfire and that, insofar as Kant has anything to teach us, many of Kant's intended contributions to ethical theory go unappreciated.

That Kant's ethical theory isn't well understood within contemporary moral philosophy isn't surprising. It does of course most of all have to do with the fact that Kant's texts are simply hard to understand.[3] It is, to use our own main topic as our example, all but immediately obvious how subjecting ourselves to guiding principles (or "maxims") that could hold as universal laws is equivalent, as Kant argues that is, to always treating the humanity in each person as a purpose in itself, and never as a means only. As we will see in the chapters to follow, most commentators believe that these formulas have different practical implications, and therefore find this claim to be, as one commenter puts it, "puzzling".[4] But Kant's ethics' not being well-understood within contemporary Anglophone moral philosophy also has to do with how many of Kant's readers read the ethical works by Kant that they do read (which is often limited to a sub-set of these) in light of contemporary discussions of ethics within Anglophone analytical philosophy (which tend to focus exclusively on the mere *permissibility* of candidate

[3] Kant's works are also hard to translate into English. Indeed, the major translations that tend to be used are, I will argue, questionable at key points in the texts, which is part of what has led to some of the major misconceptions about how to understand Kant's theory.

[4] Though she's initially "puzzled" by this claim – most of all since the two formulas contain very different concepts – Onora O'Neill eventually argues her way to an interpretation on which the formulas can be reconciled (O'Neill 1989). Thomas Hill and Christine Korsgaard are two prominent writers who argue that these formulas differ in their practical applications (Hill 1980); (Korsgaard, 1996). Derek Parfit argues that while the humanity formula implies that the permissibility of actions is dependent of what the agent's attitudes are, the universal law formula can be interpreted so as to not have this implication. (Parfit, 2011a) Allen Wood, in turn, argues that the universal law formula is not even meant to have any practical implications, and that it is a mere first step in an argument leading to the humanity formula, which, Wood argues, is the one from which all more specific imperatives derive. (Wood, 1999; 2008).

courses of action[5]) along with all the preconceptions about Kant's ethics therein. This is not conducive to achieving a fair understanding of Kant's views.[6]

When Kant, for example, says that the *morality* of an action is not determined by the consequences of this action, most contemporary readers immediately start thinking about current debates about consequentialism[7] and understand Kant as claiming that the consequences of our actions are morally irrelevant. But by the "morality" of an action Kant means the relation the agent's decision-making has to the moral law: i.e. whether the agent acted as she did out of respect for the moral law or whether she acted for some other reason, not taking into account anything morally relevant.[8] And that what the agent's prior deliberation itself is like isn't determined by the consequences that actually result when she goes ahead and takes action does *not* necessarily mean that in deliberating, she is morally permitted to disregard all considerations having to do with what consequences her actions might have. So on the basis of Kant's claim that the

---

[5] See especially section 2 of chapter 4 below.

[6] That Kant's approach to ethics is heavily influenced by the ethical theories of various Ancient schools (such as the Epicureans and the Stoics), that Kant himself puts forward his theory as a kind of virtue-ethical theory in his main work on substantive ethics, *The Metaphysics of Morals*, is also often not taken into account, which partly accounts for the tendency many readers seem to have to find Kant's claims about duties to ourselves strange. In terms of virtue-ethics there is, however, nothing strange about this idea. A person who is flourishing as a human being not only relates to others in distinctive ways, but also to herself in ways distinctive of complete human flourishing. For Kant's view of what it is for a human being to flourish completely in accordance with the basic laws of morality (as he conceives of them), see especially the two sections on human flourishing and the highest good in chapter 4.

[7] *Consequentialism* is, roughly, a view that has two main components, namely: (1) there are evaluative facts about what would be the best ways for things to go, or be, that are wholly independent of all moral considerations; and (2) what we ought morally to do is whatever would make the world conform to these facts about what would be non-morally best. See, for example, (Pettit 2003) and other contributions to (Darwall 2003).

[8] Most readers instead read the various Kant's claims about the "morality" of our actions as if, in making these claims, he were talking about what he himself calls the "legality" of our actions. Here's how he puts this distinction in the introduction to *The Metaphysics of Morals*: "The conformity of an action with the law of duty is its legality (legalitas); the conformity of the maxim of an action with a law is the morality (moralitas) of the action." (MS: 6:225)

morality of an action is not determined by its consequences we cannot, as so many commentators do, conclude that Kant thinks that the consequences of people's actions are morally irrelevant.

The first major motivation behind the following discussion of Kant's ethics, then, is simply that Kant's ethical theory is widely misunderstood, which I find unfortunate (especially since there at the same time is universal agreement that Kant is among the greatest philosophers within our tradition). Here I am of course merely asserting that Kant's ethical theory – in particular his universal law formula – isn't widely understood. In the chapters to come I will explain and argue for that claim at great length.

Secondly: we need, I believe, to investigate and clarify all major alternatives to the methodological intuitionism that dominates so much of contemporary moral theory. Much under the influence of writers such as John Rawls (in particular his "reflective equilibrium" methodology[9]) and Derek Parfit (with his extensive use of fanciful thought experiments[10]), a great deal of contemporary ethics is heavily intuition-driven. It is widely assumed that there is no alternative to this approach. It is assumed that by thinking about different general ethical principles, particular examples and situations, and our own dispositions to make certain moral judgments under certain circumstances, the chief aim of ethical theory is to work towards a state in which there is a match between the implications of general principles "we" – and it is not always clear who the "we" in question are – are inclined to accept and the intuitive responses we have with regard to

[9] (Rawls 1971) This is not to say that Rawls's own primary method of first-order moral reasoning is a kind of intuitionism. It is not. But Rawls does claim that once we have used the method he suggests (which we do not need to describe here), we must decide whether or not to accept the results (and these first-order methods themselves) by checking whether they match up with our "considered judgments," which are intuitively held moral beliefs such as the conviction – to use one of Rawls's own examples – that slavery is wrong. (Rawls 1971: 20); (Rawls 2001: 29)

[10] (Parfit 1984; 2011a-b)

how one ought to conduct oneself in particular situations. T.M. Scanlon, for example, goes so far as to write that:

> this method, properly understood, is … the best way of making up one's mind about moral matters. […] Indeed, it is the only defensible method: apparent alternatives to it are illusory. [11]

This common conception leads many philosophers to either become intuitionists about ethics in the sense of taking their own intuitions about general principles and particular cases to serve as a reliable guide to how people should live their lives, or to become skeptics about ethical thought.

The grounds for skepticism tend to depend either on the observation that people differ in their intuitive responses to particular examples as well as in their considered judgments[12], or on theories (some of which are empirically based) of the nature and natural history of our intuitive responses and judgments.[13] I share much of this skepticism about intuition-driven moral reasoning, partly for these reasons. But I disagree with the conclusion that this should lead us to general skepticism about moral reasoning. And the reason for that is that I don't share the view that it is illusory to think that there are alternatives to intuition-driven searches for reflective equilibriums. The general type of theory of which I believe Kant's theory to be an instance – namely, what I will follow others in calling "constitutivism" in ethical theory[14] – is precisely an example of a type of

[11] (Scanlon 2003: 149)
[12] (Mackie 1977)
[13] (Street 2006); (Greene 2008)
[14] (Smith 2011; forthcoming); See also (Velleman 2000) and (Kastafanas 2011). Constitutive theories understand basic moral and practical principles in terms of the conditions that our agency must live up to in order for us to robustly be able to exercise autonomous agency whereby we can be fully responsible for our actions and, some add, thereby be competent members of what is sometimes called the "moral community" (Darwall 2006). Another similar line of thought is that which is associated with Jürgen Habermas and some of the other contemporary members of the

theory that doesn't take intuitions about cases as the basic input and ultimate arbitrator of ethical reasoning. So to decide whether all supposed alternatives to reflective equilibrium-seeking intuition-driven moral reasoning really are illusory, one of the things we must do is, I believe, to thoroughly investigate how exactly to understand the theory that Kant puts forward, since it might be one of the non-illusory alternatives that Scanlon and others overlook.

Our topic, then, is how to understand the basics of Kant's ethics – particularly the relation between the universal law and humanity formulations of the categorical imperative – and the two main motivations prompting our discussion are: (1) that Kant's ethical theory, for various reasons, unfortunately isn't widely understood within Anglophone moral philosophy; and (2) that Kant's theory seems to offer an alternative to the intuitionist – and therefore sensibility-relative – methodology at the heart of most contemporary philosophical discussions of both the foundations and applications of ethics. In this introductory chapter I will roughly outline the reading of Kant's ethical theory I will argue for in subsequent chapters and then also outline the main chapters of this dissertation.

2. Korsgaard on Self-Constitution in The Ethics of Plato and Kant

The reading of Kant's theory I offer in this dissertation draws heavily on recent work by Christine Korsgaard in which she presents a way of thinking about ethics that

---

so-called Frankfurt school, namely the *discourse ethical theory* according to which basic moral principles spell out the conditions we must live up to in order to be able to participate competently and equally in social practices of justification and shared deliberations about what specific norms are to govern our social interaction. Such theories are also not driven by intuitions about cases. See, for example, (Habermas 1991) and (Forst 2011).

synthesizes ideas from Kant, Plato, and Aristotle.[15] My reading of Kant does, however, also greatly differ from Korsgaard's in important respects – both regarding how to understand the universal law formula and with regard to how to understand the humanity formula – and much of what follows in the chapters below circles around critical engagements with Korsgaard's treatment of various aspects of Kant's ethics. I will, however, take the aspects of Korsgaard's recent work that I draw on as the starting point of my discussion, namely her work on what she calls the "Constitutional Model" of human agency, and how it relates to the principles of morality.

The constitutional model of human agency is contrasted with the "Combat Model" under which human actions are outcomes of struggles between different forces within the agent, such as competing desires, impulses, and passions. The clearest example of this combat model is probably Hobbes's theory according to which a person's *will* is identified with the last desire that a person has before taking action, and on which deliberation is viewed as a struggle between different desires that all, so to speak, try to become that last desire that constitute the person's will.[16] In contrast to this type of view,

[15] The parallels Korsgaard draws between Kant's theory and Aristotle's ethics have to do with what kinds of action theories, on Korsgaard's readings, these respective philosophers are operating with. In Kant's case, Korsgaard's suggestion is that the maxims Kant thinks all our actions are performed on are principles specifying what *means* to take for the sake of what *ends*. Since I disagree, for reasons to be given in chapter two, with this reading of what a maxim is for Kant – at the very least when it comes to the maxims that are candidate universal laws – I will not be discussing the parallels Korsgaard thinks there are between Kant and Aristotle's respective action theories. I do agree, though, that there are many more similarities between the ethics of Kant and Aristotle than is normally recognized, having to do, for example, with the central role friendship (or what Kant calls "the moral friendship") has in both of these philosophers' accounts. Kant also offers an argument in the first section of the *Groundwork* that has a lot in common with Aristotle's famous function argument (Aristotle 1999). And the reading of Kant's theory of a fully flourishing human being that we will offer in chapter 4 furthermore has, I think, similarities with Aristotle's theory of human flourishing (which is why I there put things in terms of what, on Kant's view, it is for an agent to *flourish* from the point of view of the highest good, as Kant conceives of it).

[16] (Hobbes 1651)

Korsgaard presents the "Constitutional Model", which is her own preferred model, in the following way:

> What distinguishes action from mere behavior and other physical movements is that it is *authored* – it is in a quite special way attributable to the *person* who does it, by which I mean, the *whole* person. The Constitutional Model tells us that what makes an action yours in this way is that it springs from and is in accordance with your constitution. But it also provides a standard for good action, a standard which tells us which actions are most truly a person's own, and therefore which actions are most truly *actions*.[17]

In short, on this model, "…the function of action is *self-constitution*."[18] But, what exactly does this have to do with morality? Korsgaard writes:

> [As] Plato… taught us, in the *Republic*, …the kind of unity required for agency is the kind of unity that a city has in virtue of having a just constitution… Following [Plato and Kant's] lead, …I argue that the kind of unity that is necessary for action cannot be achieved without a commitment to morality. The task of self-constitution, which is simply the task of living a human life, places us in a relationship with ourselves… We make laws for ourselves, and those laws determine whether we constitute ourselves well or badly. And … the only way in which you can constitute yourself well is by governing yourself in accordance with universal principles which you could will as laws for every rational being.[19]

---

[17] (Korsgaard 1999: 3)

[18] (Korsgaard 2009: xii)

[19] (Korsgaard 2009: xii)

I believe that it is indeed correct to see Kant as roughly following Plato's lead in the *Republic*: that is, as trying to analyze virtue via an idea of a just republic. In Kant's case, though, virtue turns out being analyzed via the idea of a positively free will and the republic is not exactly Plato's ideal city, but rather the kind of republic Rousseau describes in the *Social Contract*: in which we, according to Rousseau's argument, can be free at the same time as being subject to restraining laws for the reason that these are laws that, as members of the body politic, we impose on ourselves.

Rousseau even directly makes what sounds precisely like the claim that Kant's ethical theory is based around, namely that "over and above all this, [we can also] add, to what man acquires in the civil state, *moral liberty*, which alone makes him truly master of himself; for the mere impulse of appetite is slavery, while obedience to a law which we prescribe to ourselves is liberty." (Rousseau 1762: 10, Emphasis added)

Rousseau is of course here talking about external laws voted upon by members of the body politic, which arise out of the "general will" that members form and partake in through their

But the requirement to act on guiding principles (or "maxims") that could qualify as laws for all rational beings is, on the Kantian conception of morality, the most fundamental imperative of morality; this is the universal law formulation of the categorical imperative. It therefore follows, if Korsgaard is right, that:

> …a commitment to the moral law is built right into the activity that, by virtue of being human, we are necessarily engaged in: the activity of making something of ourselves. The moral law is the law of self-constitution, and as such, it is a constitutive principle of human life itself.[20]

So in order to at all exercise *agency*, as opposed to merely behaving or reacting to things that happen to us, we must, Korsgaard argues, follow the moral law, which is thus the constitutive principle of all human agency: the principle that must govern our choices if we are to count as acting at all.

Is this how to understand Kant's view? My view is that Korsgaard is right that Kant's ethical theory is a kind of "constitutivism" on which the moral law is the constitutive principle of a certain kind of agency, but that Korsgaard's own constitutivism – which she also attributes to Kant and Plato – is different from Kant's. Korsgaard's constitutivism, I believe, is much more rigid than Kant's (and therefore lacks some of the most important strengths of Kant's theory). The next section explains what I mean by this.

---

collective deliberations. Kant applies this idea to the inner principles governing a person's own will, which becomes autonomous when the person chooses principles that are fit to serve as universal laws. That is how Kant is using Plato's *Republic* strategy with materials from Rousseau's *Social Contract*.

[20] (Korsgaard 2009: xiii)

3. How Kant's Constitutivism (as I understand it) Differs from Korsgaard's

Kant himself, I believe, does *not* hold the view that agency is possible only if our wills are governed by the moral law (i.e. only choosing substantive guiding principles that could serve as universal laws for all reason-endowed beings). Kant instead holds that agency is indeed possible outside of the complete rule of the moral law, and this is why the moral law functions as an *imperative* for human beings.[21]

Kant doesn't put these things in terms of "agency", however, but instead in terms of "freedom". To get a feel for what Kant's view is, consider this passage from the introduction to the *Metaphysics of Morals* (which sums up much of Kant's theory):

> …choice, which can be determined through *pure reason*, is called free choice. That [choice], which is only determinable through inclinations (sensual drives, stimuli), would be animal choice (arbitrium brutum). Human choice, in comparison, is of such a sort that although it is affected by drives, it is not determined thereby [and yet] is in itself (without the acquired skillfulness of reason) not pure, but nevertheless capable of being determined by a pure will. The *freedom* of the power of choice is this independence from *determination* through sensual drives; this is the negative concept thereof. The positive one is: the capacity of pure reason to be practical for itself. This is however not possible other than through the subjection of the maxims of all actions under the condition of their fitness [to serve as] universal laws … [and] these laws of freedom … are called *moral*.[22]

---

[21] I discuss Kant's understanding of the categorical imperatives of practical reason briefly in section 5 below. The idea is, roughly, that we get categorical *imperatives* of practical reason by using the wills of imagined versions of ourselves whose wills are autonomous (but who otherwise are like us in their desires, inclinations, needs, etc.) as *standards* to live up to: what we *ought* to do is what we *would* do in these nearby possible worlds in which we govern ourselves on the basis of maxims that are fit to serve as universal laws. (G: 4:449; 4: 454)

This way of describing Kant's view will, in many readers, bring to mind the views Michael Smith puts forward in works as (Smith 1994) and I will, precisely for that reason, relate Kant's view to one of the distinctions Smith draws in his work towards the end of section 5, namely the distinction between "exemplar" and "advice" models of normative reasons for action (Smith 1995). For a discussion of norms of reason as "standards of correctness" (much like the squares or rulers that builders use), which we can fail or succeed to meet, see (Railton 2000:1-2).

[22] (MS: 6:213-4)

I understand Kant to be saying that even the person who is not wholly determined by the moral law of pure reason can nevertheless exercise agency (i.e. enjoy negative freedom of the will), since such a person can be so constituted that she is *not*, to use Hume's phrase, a complete "slave of her passions"[23]: her actions are not, that is, necessarily wholly determined by her impulses, her instincts, and her direct sensual responses to cues in her immediate surroundings in the way in which we imagine the behaviors of most non-human animals to be.[24]

Thus in the following passage from the *Critique of Pure Reason,* Kant writes in a way that clearly suggests that he thinks that if a creature possesses the capacity to exercise some degree of *impulse-control*, then that is already sufficient for it to qualify as an agent, as somebody who – as Kant puts it – enjoys "practical freedom":

> A capacity for choice... is merely animal-like (arbitum brutum), [if it] cannot be determined otherwise than through sensual drives/predispositions, i.e. pathologically. That capacity for choice, however, that can be determined independently of sensual drives/predispositions, [and] which [can be determined through] motives that can only be represented through reason, is called a free

[23] Hume famously wrote that, "Reason is and ought only to be a slave of the passions" (Hume 1741). From where Hume gets the idea that reason *ought* to be the slave of passions, especially since he argues that reason *is* merely a slave of passions, has always seemed like a mystery to me. This must be an addition made for the sake of stylistic effect, and it does indeed make for a striking phrase. For two of the most thoroughly developed recent Hume-inspired accounts of practical reason on which, just as Hume claims, practical reason can ultimately only be a "slave" and never a ruler of the passions, see (Persson 2005) and (Schroeder 2007).

[24] Note that the standard way to think of *freedom* generally during this period is what Philip Pettit calls the "republican" way, whereby freedom means non-domination by an external master. A citizen, for example, is free when she is not subject to the arbitrary power of another person, such as an absolute monarch. The will, Kant similarly takes it, is "negatively" free when it is not subject to the arbitrary power of impulses and passions. Freedom, thus, is *not* defined as the capacity to do whatever one wants, but as the absence of external domination. For Pettit's discussion of the republican concept of freedom, which Pettit tries to reestablish in the realm of political thought, see (Pettit 1997). Part of what confuses contemporary readers of Kant's texts is, I believe, that this *republican* way of thinking about freedom is much further from their minds than the standard *liberal* way of thinking about freedom, whereby freedom consists in having options and/or being able to do what one desires. For more on the relevant genealogy of the concept of *liberty*, see (Skinner 1998).

> capacity for choice (arbitum liberum), and everything that is a basis or an effect of it is called practical. Practical freedom can be proven on the basis of experience. Because not only that which excites, i.e. that directly agitates the sensibilities, determines human choices, but we have also a capacity to through representations of what is itself only harmful or useful in the future overcome the impressions of our sensual desire-faculty; these considerations, however, regarding that which is desirable for our whole condition/state, i.e. good and useful, derive from reason.[25]

Kant's view, then, is not that agency is only possible in accordance with the moral law.[26] Kant's point is rather, I believe, that the agency of a person who is not governed by the moral law is unstable or insecure (or that it, in a sense, is a less free type of agency than that which a person who *is* governed by the moral law exercises). The agent who is *not* governed by the moral law is somewhere on a *continuum* between the following two extremes: full autonomy through the rule of her own practical reason, at the one extreme, and a complete lack of agency and freedom through the rule of irresistible impulses, at the other extreme. The more she is governed by her immediate impulses, the less responsible she is for what she does: the less of an autonomous agent she is. The more she is governed by her capacity to determine herself to action through her representations

[25] (KrV: B831, A 803) See also the footnote at (R: 6:58) in which Kant writes of the Stoics that these "philosophers derived their universal moral principle from the dignity of human nature, from its freedom (as an independence from the power of the inclinations), and they could not have laid down a better or nobler principle for its foundation," which again suggests that, as Kant here puts it, "an independence from the power of the inclinations" is sufficient for agency or freedom, which means that if we can enjoy some amount of such independence even if we are not completely governed by the moral law, then agency is possible outside of the complete rule of the moral law within us. What governing oneself on the basis of the moral law does is, instead, to robustly secure this kind of independence, which gives us what Kant calls positive freedom in addition to mere negative freedom. Or so I argue that we ought to understand Kant's view.

[26] His view rather seems, again, to be that as soon as our reason has any kind of influence over our behavior we are exercising some form of agency. Note also that in the *Religion*, Kant argues that a person can choose *evil* over *good* by making the promotion of her own happiness, rather than a moral principle, into her most basic guiding principle or maxim. (R: 6:29, 36) Such a choice can, Kant claims, be *imputed* to the agent. (R: 6:37) This again clearly suggests that Kant thinks that there at the very least is some form of agency – which falls short of fully autonomous agency – that we can exercise outside of the kind of inner rule of law we achieve by choosing our maxims on the basis of their fitness to serve as universal laws.

of principles of action, the more fully responsible she is for what she does: the more does she approach the possibility of full autonomy.

What the person who is not governed by the moral law – i.e. that of choosing her guiding principles on the basis of their fitness to serve as universal laws – lacks is self-mastery, autonomy, or what Kant also calls "positive freedom". She does not necessarily lack the capacity for agency as such, and when she acts she is not necessarily acting on an irresistible impulse. But she resides, as Kant puts it, in an "ethical state of nature" out of which she can only emerge by subjecting herself to maxims of virtue that could hold as universal laws. As Kant writes in *Religion within the Limits of Reason Alone* (again using an analogy between inner and outer freedom):

> A *juridico-civil* (political) *state* is the relation of human beings to each other inasmuch as they stand jointly under public juridical laws (which are all coercive laws). An *ethico-civil* state is one in which they are united under laws without being coerced, i.e. under *laws of virtue* alone.
>
> Now, just as the … *juridical state of nature* is opposed to the first, so is the *ethical state of nature* opposed to the second. … [And] just as the state of lawless external (brutish) freedom and independence from coercive laws is a state of injustice and of war, each against each, which a human being ought to leave behind in order to enter into a politico-civil state, so is the ethical state of nature a … feuding between the principles of virtue and a state of inner immorality which the human being ought to endeavor to leave behind as soon as possible.[27]

What the moral law is a constitutive principle of, then, is not agency (or negative freedom) as such (which consists of independence from the rule of impulses and passions), but instead autonomous agency or positive freedom: the kind of fully

[27] (R: 6:95-7) We here again see, of course, that Korsgaard is indeed right to say that Kant, following Plato, thinks of virtue on analogy with the just lawfulness within the civil constitution of a republic under the rule of law.

responsible agency that consists in self-governance through our capacity to think about action in terms of general and universal principles, i.e. our practical reason.[28]

Return now to Korsgaard's claim that action in accordance with the moral law amounts to self-constitution. Let's relate this idea to the humanity formula. Since our *humanity*, as Kant thinks of it, consists of our peculiar dual nature as beings capable of autonomy who nevertheless also are animals subject to various needs and desires[29], which together help to form our personal conceptions of happiness, we most fully realize our own humanity when our pursuit of happiness does *not* lead us to lose our autonomy (and when we instead bring these different aspects of our nature into harmony with each other). We most fully realize our own humanity, in other words, when we subordinate our pursuit of happiness (without abandoning it) to the condition that our basic guiding principles could serve as universal laws of autonomous human agency. But doing so amounts to making the preservation and full realization of the distinctively human kind of autonomy into the most general purpose around which our actions are organized. This, as I will argue at length in the chapters that follow, is why the moral law, as Kant thinks of it, can also be said to instruct us human beings to so act that we always treat the humanity in each person as an end, and never as a means only.

The constitutive *aim* of our human brand of autonomous agency is, thus, simply our own humanity itself.[30] In that respect Korsgaard is right, I think, to say that the Kantian view is that we constitute ourselves (as the particular kinds of beings that we are)

---

[28] (G: 4: 412)

[29] See, for instance, (MS: 6: 420).

[30] And this is also the sense in which, according to Kant, our humanity, as he puts it, "exists as a purpose in itself": the preservation and full realization of the humanity within each person is the most general aim our actions would be organized around if we were completely governed by the most basic principles constitutive of autonomous human agency.

through the principles we subject ourselves to, and moreover that we constitute or fully realize our own humanity only if and insofar as we subject ourselves to the moral law. (I expand on this line of reasoning a little more in the section following this one.)

To sum up this section: I agree with Korsgaard that it is Kant's view that we constitute ourselves as autonomous agents by subjecting ourselves to maxims that can serve as universal laws; but I disagree with the idea that the only kind of agency that there is, as Kant sees things, is this fully autonomous kind. The Kantian view, as I understand it, is rather that agency occurs wherever we enjoy more or less independence from the rule of our impulses, but that fully autonomous agency can only be achieved through the subjection to self-adopted principles that are fit to serve as universal laws of reason-governed agency.[31] As noted above, I also have further disagreements with Korsgaard, particularly with regard to how to understand the universal law and humanity formulas. But we will save those disagreements for later chapters.

4. The Role of Kant's Constitutivism in the *Groundwork* (and the Human Nature Formula)

I started out by saying that the topic of this dissertation is how to understand the universal law and humanity formulations of the categorical imperative and the relation between the two. I then went on to offer a qualified endorsement of Korsgaard's claim

[31] By subjecting ourselves to such principles, we leave the ethical state of nature behind and enter into a state of virtue in which we are fully responsible for our conduct – a state in which we can also offer sensibility-neutral justifications of our actions both to ourselves and to others since we then govern ourselves by principles whose universal following would allow all to most fully realize their own humanity independently of what their particular conceptions of happiness and the good life are. Part of what the Kantian theory offers us, then, is a sensibility-neutral way of reasoning about ethics that doesn't require appealing to sensibility-relative intuitions.

that Kant's ethical theory is a form of constitutivism, i.e. a theory that sees the fundamental principles of morality as principles constitutive of autonomous agency. Since it might be wondered what one thing has to do with another – i.e. what the relation between the two main formulations of the categorical imperative has to do with Kant's theory's being a kind of constitutivism – I will now offer a rough sketch of the overall argument regarding how to understand Kant's theory that I will be laying out across the chapters that follow. In so doing I will try to explain why understanding Kant as a constitutivist is the main key to understanding the relation between the universal law and humanity formulations of the categorical imperative.

Return first briefly to Korsgaard's comparison between Kant's approach to virtuousness and Plato's approach to the virtue of justice (as a virtue of a person) in the *Republic*. Recall that Plato has Socrates suggest that we understand justice of the soul by first forming a view of the constitution of a just city and then applying that view to the inner state of the person.[32] Following Korsgaard I think it is correct to think of Kant as using a similar strategy with regard to freedom of the will: Kant applies Rousseau's idea that we achieve civil liberty by subjecting ourselves to laws that we give to ourselves as members of the body politic[33] to the issue of freedom of the will, claiming that the will secures freedom from external rule by subjecting itself to self-legislated laws (i.e. self-adopted maxims whose form and further properties make them suitable to serve as universal laws for all reason-governed beings.)

When it comes to understanding the nature of these laws of free agency themselves in the Kantian way, however, the analogy with the constitution of a just

[32] (Plato 2004)
[33] (Rousseau 1762)

republic no longer applies. Because when it comes to what kind of laws we are to subject ourselves to through our choices of guiding principles, Kant instead uses the analogy of natural laws (as he understands them within his critical philosophy). As he puts it in the *Critique of Practical Reason*, laws of nature serve as the "type" on which the laws of freedom (the laws of morality) are modeled.[34] This is why Kant claims, in the *Groundwork*, that the universal law formula is equivalent to what is sometimes called *the law of nature formula*: so act that the maxim of your will could also hold as a universal law of nature.[35]

Laws of nature on Kant's view are the most general principles in accordance with which things exist and operate in accordance with their particular natures (or their particular constitution).[36] And when it comes to the laws of autonomous agency, these laws are therefore the most general principles in accordance with which potentially autonomous agents can exist and fully realize their own nature as beings capable of operating according to self-adopted laws. The subject-matter, so to speak, of all maxims that can hold as universal laws, then, are *reason-governed beings themselves*: they are the things whose existence and flourishing is thought to be made possible only if they operate in accordance with the particular substantive principles in question.[37]

Consider next Kant's different uses of the term "nature". In what he calls its *material* sense, "nature" refers to the sum-total of everything that is a potential object of the senses.[38] But in what Kant calls the *formal* sense, a particular "nature" is a particular constitution: a type of existence that is possible in accordance with particular general or

[34] (KpV: 5:67-72)
[35] (G: 4:429)
[36] (P: 4: 318-9); (MN: 4:467); (MN: 4:469) See also chapter two, section six.
[37] (G: 4:429)
[38] (MN: 4:467)

universal laws.[39] So when Kant uses the phrase "the reason-endowed nature", he is referring to a particular type of entity[40] (namely, beings of the sort that can be governed by their own practical reason) whose existence is possible only in accordance with certain general or universal laws (in this case, maxims we give to ourselves that are fit to serve as laws of our nature, i.e. universal principles in accordance with which we can exist and flourish as beings capable of being governed by self-adopted guiding principles[41]).

Now if maxims that could qualify as universal laws of the reason-endowed nature are understood as principles of action that each reason-endowed being could adopt and in accordance with which they could both persist and fully realize their particular nature as beings that can be governed by their own practical reason, then it immediately follows that in operating in accordance with such principles, we are thereby making the existence and realization of this particular kind of nature into the most general end or purpose around which all our actions are organized. This, as I will now try to explain, is why the humanity formula turns out being equivalent to the universal law formula (as the universal law formula applies to human beings[42]).

In our case as human beings, as already noted above, we are not only reason-endowed beings capable of autonomous agency, but also animals with desires, needs, and

[39] As Kant puts it in the *Groundwork*: "nature in the most general sense (as regards *form*)" refers to "the existence of things insofar as it is determined in accordance with universal laws." (G: 4:421) See also: (MN: 4:467).

[40] Cf. (Timmermann 2006)

[41] Some of Kant's examples are: "live in accordance with your nature;" "build your mental and bodily powers for fitness for all ends that may spring out from you;" "the duty of respect" (i.e. "not to make others throw themselves away to pander to your ends"); and "love your neighbor as yourself". (MS: 6: 419; 449; 451) Additionally the most basic requirement of virtue for Kant is perhaps the "*duty of apathy*," which is to become *master* over one's "affects" and *ruler* of one's "passions." (MS: 6: 407-8) (Note that Kant is here adopting Stoic terminology, whereby "apathy," for example, (as in ἀπάθεια or *apatheia*) does *not* mean indifference. More on this in chapter four.)

[42] I say "as it applies to human beings" because the universal law formula is meant to apply to all possible reason-endowed beings, whereas the humanity formula is limited to all human beings.

a resulting general wish for happiness. This is our humanity, as Kant thinks of it.[43] So if we make our own nature as this particular variety of reason-endowed beings into the thing whose existence and flourishing is to be achieved through our subjection to self-adopted maxims modeled on laws of nature, then in following these principles we would in effect be making the existence and full realization of humanity (i.e. our own particular kind of reason-endowed nature) into the most general end or purpose[44] around which our actions were organized. If we were governed by such principles (whatever precise principles they might be[45]), then we would thus never be treating beings of our human kind only as means for other purposes, but would always at the same time make sure that we treated their existence as a purpose in and of itself. And this, as I will argue in the chapters that follow, is exactly why Kant claims that the categorical imperative (as applied to human beings) can also be expressed as the humanity formula: so act that you

[43] See, for example, (MS: 6: 420 & 435)

[44] I say "end or purpose" here (and will often do so in what follows) since it is not altogether clear in my judgment that translating Kant's "Zweck" into "end" is a great idea (given the connotations "end" has for many of us now). I also prefer translating Kant's "vernünftiges Wesen" into "reason-endowed being" for two reasons: I want to avoid the connotations related to "rational" that are brought about by the normal translation of Kant's term into "rational beings". Second, Kant doesn't think that reason-endowed beings are always at the same time reason-*governed* beings. Reason-endowed beings are, rather, beings capable of being reason-governed, but sometimes or often need to make an effort in order for it to be so. (In the *Anthropology from a Pragmatic Point of View*, for example, Kant writes, "It is thus the case for us that when it comes to put the human being into a class within the system of living nature and characterize him, nothing remains but: that he has a character, which he himself brings about in how he is capable of perfecting himself on the basic of his own chosen ends; through which he can from being an animal capable of reasoning (*animal rationabile*) turns himself into a reason-governed animal (*animal rationale*)." (AP: 7: 321))

[45] We already saw some examples in footnote 41 above, but we will get to some more examples in the next chapter. It is important to do so because one of the major problems with the ways in which Kant's universal law formula is usually approached is that writers tend to focus on Kant's examples of situation-specific practical principles that *cannot*, in his judgment, be universal laws rather than Kant's examples of basic principles that do qualify as universal laws.

always treat the humanity in each person, never as a means only, but always at the same time as an end.[46]

Thus to act in accordance with self-adopted maxims that could serve as universal laws of our own nature (i.e. as the principles in accordance with which beings of our kind can exist and flourish) is to so act that we never treat our own nature as a means only, but always at the same time as an end or purpose in itself. Let us now relate this idea to Kant's constitutivism. Since (1) part of what is distinctive about our nature (as Kant thinks of it) is the very capacity to subject ourselves to principles that, in this way, take the form of laws – since it, in other words, is part of our nature that we are capable of autonomy in this sense – *and* (2) the requirement to act in accordance with maxims that could be universal laws is the moral law in its basic form, it follows that the moral law is the constitutive principle of our nature insofar as we are members of the more general class of reason-endowed beings. Since, however, (3) we are not only beings capable of autonomy, but (4) we are also beings who are animals with various inclinations, needs, etc. – since we are, as Kant uses the term, *human* beings – it more specifically follows that the principle of humanity is the constitutive principle of our particular kind of reason-endowed nature.

[46] Now what complicates things is that *part* (though *not* the whole) of what is distinctive of our nature is our capacity for positive freedom: our capacity to act in accordance with maxims that could serve as universal laws. So there can appear to be a threat of a complex kind of circularity here. We are to act in accordance with self-given principles in accordance with which we can all exist as the particular kind of animals with the particular kind of needs we have etc., but we are *also* to act in accordance with self-given principles in accordance with which we can exist *as beings that can act in accordance with self-given principles that could serve as laws for beings of our own kind*. Does that not lead, one might reasonably wonder, to some kind of regress or vortex or something along those lines? No, Kant would reply, because principles of the latter kind are mainly prohibitions against courses of action that would undermine our own or other people's ability to govern themselves using their own practical reason, etc. This is part of why Kant says that the reason-endowed nature exists as a negative end, i.e. one that is *not* to be acted against. (G: 4: 437)

Or, as we could also put it, it follows that the universal law formula as it applies to us human beings can be formulated as follows: act only on maxims that all human beings could follow and whose following would allow us all to preserve and fully realize our particular human nature. If we wanted to give this a name, we could call it *the human nature formula*.

But we should keep in mind that this human nature formula is simply the universal law/law of nature formula applied to our particular brand of the more general "reason-endowed nature". The idea is also not that the universal law formula is to be derived from facts about human nature, but rather that, as the universal law formula applies to our particular type of reason-endowed nature, it becomes equivalent to the human nature formula (whose following thus is equivalent to following the humanity formula[47]).

With these ideas in mind we begin to get a sense for how to understand one of the sets of claims in the *Groundwork* that otherwise appear most obscure, namely the following two. That (1) each reason-endowed being with a will necessarily – and for a reason that is valid for all – conceives of its own existence as a purpose in itself (and thereby as an "objective purpose" given by pure reason itself); and that (2) the grounds

[47] One can argue as follows: in choosing her maxims on the basis of their fitness to serves as laws of the kind of nature she exemplifies (and thus following the universal law/law of nature formula), the human being would need to choose maxims on the basis of their fitness to serve as universal laws for the preservation and full realization of the particular human nature (the human nature formula). But if we followed the human nature formula, we would be making the preservation and full realization of the human nature into the most general purpose around which we'd organize our actions, and we would never subordinate this purpose to any other purpose, such as the pursuit of our own happiness. It thus follows that the human nature formula is equivalent to the humanity formula: so act that you always treat the humanity in each person as a purpose in itself, and never as a means only (the humanity formula): i.e. as something whose preservation and full realization is a purpose that should never be subordinated to any other purpose. Since the human nature formula just is the universal law formula as it applies to us, and the human nature formula is equivalent to the humanity formula, the universal law formula, as it applies to us, is equivalent to the humanity formula.

for why this is so are contained in the arguments given in the third part of the *Groundwork*, which is about freedom of the will.[48]

What Kant does in the third part of the *Groundwork* is, firstly, to argue that all agents with a will of their own "act under the idea of freedom" (i.e. the idea of not being completely determined by their impulses, instincts, inclinations, etc.), and that a will can only securely enjoy this independence from the rule of external forces if it operates in accordance with laws it gives to itself. But laws that we give to ourselves cannot, he continues, be anything other than self-adopted basic guiding principles (or "maxims") that have properties that make them fit to serve as laws for all beings possessing a will of their own (i.e. universal principles in accordance with which these beings can exist and flourish as beings of their particular kind). And the requirement to choose one's maxims on the basis of their fitness to serve as universal laws is the most fundamental law of morality. This means that to fully realize our partial nature as beings capable of fully autonomous agency, we must subject ourselves to the moral law; it means, in other words, that the moral law is the constitutive principle of fully autonomous human agency. This, in turn, means that we can only securely enjoy the kind of independence from the

[48] In full, the relevant passage, which will be the main subject of chapter four, runs like this:

"If, then, there is to be a supreme practical principle and, with respect to the human will, a categorical imperative, it must be one such that, from the representation of what is necessarily an end for everyone because it is an *end in itself,* it constitutes an *objective* principle of the will and thus can serve as a universal practical law.

The ground of this principle is: *the reason-endowed nature exists as an end in itself.* The human being necessarily represents his own existence in this way; so far it is thus a *subjective* principle of human actions. But every other reason-endowed being also represents his existence in this way consequent on just the same reasonable ground that also holds for me;* thus it is at the same time an *objective* principle from which, as a supreme practical ground, it must be possible to derive all laws of the will. The practical imperative will therefore be the following: *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.*

* [Kant's footnote:] Here I put forward this proposition as a postulate. The grounds for it will be found in the last Section." (G: 4:428-9)

rule of our impulses etc. that we assume in all of our decision-making – and thereby ourselves become the fully responsible authors of our own actions – if we submit ourselves to the moral law.

But in accordance with the moral law, the humanity within each person is a purpose in itself (since choosing one's maxims on the basis of their fitness to serve as laws for the preservation and realization of our own human nature amounts to making the preservation and full realization of our shared humanity into the most general purpose around which our actions are organized). So if, as Kant claims, each agent with a will of its own cannot think of herself as securely enjoying the kind of practical freedom she necessarily self-attributes (as she does in acting under the idea of freedom) unless she subjects herself to the moral law, it follows that the preservation and full realization of her own humanity is a purpose in itself in accordance with the law she must subject herself to in order to make herself into a fully autonomous being (and thereby fully secure and fully realize the practical freedom she attributes to herself). If we don't subject ourselves to the moral law, whereby we make the humanity within each person into the most general purpose around which all our actions are organized, then we necessarily, in other words, remain in an ethical state of nature in which the practical freedom we self-attribute is all but secure, and there instead is an unstable type of war of all against all among the different motivational forces that exist within us. And this, I believe, is precisely why Kant claims that (1) each reason-endowed being with a will necessarily represents her own existence as a purpose in itself and (2) that this is established by the arguments given in section three of the *Groundwork*.

To simplify even more: the argument Kant has in mind is, roughly, that (1) we act under the idea of a practical freedom we can only fully secure and fully realize by subjecting ourselves to the moral law (doing so makes us fully autonomous); that (2) our own humanity is a purpose in itself from the point of view of the moral law; and that (3) our own humanity, therefore, is a purpose in itself from the point of view of the basic law we must subject ourselves to in order to fully secure and fully realize the kind of practical freedom under the idea of which we always act (insofar, that is, as we think of ourselves as having a will of our own).

All the interpretative claims offered in this section will be argued for at length in the chapters below. But I hope that what has been said in this section already gives an idea of why, as I believe, understanding Kant's theory as a kind of constitutivism is the key to understanding his universal law and humanity formulas, and the relation between the two.

5. The Kantian View of Categorical Imperatives of Practical Reason

We can next, to return to another topic mentioned above, briefly consider how to understand the Kantian view of categorical *imperatives* of practical reason in relation to the human being. I've already explained how, on my understanding of Kant's view, the preservation and full realization of our own humanity is the constitutive aim of fully autonomous human agency.[49] But, as I've also argued, Kant's view allows for a

[49] To repeat, Kant's view is that maxims that could serve as universal laws of our nature are basic guiding principles that all human beings could follow and at the same time preserve and fully realize their nature as beings who despite having needs and desires nevertheless can be governed by their own practical reason (as opposed to their own or other people's whims and impulses). If we follow such guiding principles, then we in effect make the preservation and full realization of our humanity into the most general purpose around which our actions are organized, and this

distinction between fully autonomous human agency and less than fully autonomous – and therefore "heteronomous"[50] – human agency.[51] In fact, Kant's view allows, as we've seen above, for a continuum between fully autonomous agency, and, on the other extreme, complete governance by irresistible impulses (e.g. extreme drug-addiction) and/or uncontrollable responses to cues in one's immediate surroundings (e.g. uncontrollable phobias). Kant's view is that most human beings are somewhere on the middle of this spectrum, or perhaps rather somewhere on the side approaching autonomy, and this is why, on his view, the moral law takes the form of a categorical *imperative* in relation to the human will: it serves as a *standard* that we must live up to if we are to fully realize the humanity within us and ourselves become fully responsible for our own actions.

If, without the need for any effort, we naturally always acted on the basis of self-adopted maxims that could serve as laws for the realization of the humanity within us, then we would have what Kant calls *holy* wills (an impossibility for human beings).[52] The moral law would then not take the form of an imperative in the relation to our wills. If, however, acting in accordance with self-adopted maxims that could serve as universal laws of our humanity requires continual effort on our part, as Kant thinks it does for human beings, then we are being *virtuous* insofar as we are conducting ourselves in

---

means that the most general principle guiding the will of an autonomous human being can also be said to be to always treat the humanity in each person as an end, and never as a means only. That is the sense in which the realization of our own humanity is the constitutive aim of fully autonomous human agency.

[50] (G: 4:433)

[51] The problem with Korsgaard's reading of Kant, as I've also argued, is precisely that it seems unable to account for this feature of Kant's view since Korsgaard thinks that only autonomous agency is agency at all.

[52] (G: 4:439)

accordance with such principles.[53] And while the moral law describes rather than prescribes the basic principle in accordance with which fully autonomous wills *would* operate, in relation to us it takes the form of an imperative of virtue telling us how we *ought* to conduct ourselves.[54] It is a standard we must live up in order to assume full control over ourselves or, as Korsgaard puts it, in order to constitute ourselves as fully responsible authors of our own actions with autonomous wills of their own.

How, it might now be asked, does this view relate to the distinction Michael Smith draws between *exemplar* and *advice* models of reasons for action?[55] On the former kind of view, as Smith defines it, what a person *ought*, or has *most reason*, to do is what an ideally rational version of her would do in the situation she is in. On the advice model, however, what a person ought or has most reason to do is what an ideally rational version of her would advise her non-ideal self to do in the situation that her non-ideal, actual self finds herself in (with her various non-ideal attitudes etc.). It might seem as if Kant's view is of the former kind, i.e. that it tells us that we ought to do what a version of us with a *holy* will would do if she were in the situation we're facing. But that, I think, is not right.[56] (I will also return to this issue in chapter 4 below.)

---

[53] (KpV: 5: 128); (MS: 6: 394-5)
[54] (G: 4: 454)
[55] (Smith 1995)
[56] Very strictly speaking, Kant's theory constitutes a third alternative here, but we don't need to be quite that refined to make the point I wish to make here. (Briefly: Smith's advice model is about how an ideally rational version of us would want or advise the actual version of us to act, but Kant's model, I believe, is about how the nearest possible non-ideal, but nevertheless autonomous version of us would act. It is not an exemplar model in Smith's sense since it is not about how an altogether ideal – or holy – version of us would act. Indeed, Kant thinks that it is close to a contradiction to imagine a holy human being: such a being would not be human since she would not be subject to the kind of limitations, needs, and desires distinctive of our humanity.)

Since we don't have holy wills, and couldn't possibly acquire such wills either, but we instead have to exert effort or perhaps even struggle in order to bring ourselves to act on maxims that could serve as universal laws, what we ought to do is, as I understand Kant's theory, what *virtuous* versions of us would do. This means that what we ought to do is what versions of us that have our desires, drives, hopes, fears, needs, etc., but who nevertheless manage to be autonomous would do.[57]

So to use an example of Smith's[58], supposing that we are overcome by frustration (perhaps due to losing a game of squash), what we ought to then do is what allows us to remain autonomous despite the fact that we are sometimes, as in this case, subject to these kinds of episodes. Supposing that we realize that we would hit our opponent with our racket if we went over to shake his or her hand due to our frustration, but that instead, say, counting to ten a few times would allow us to regain our composure, then what the Kantian theory implies, as I understand it, is that – although a holy version of us would easily be able to walk over and shake our opponent's hand – what we ought to do (all else being equal) is here to count to ten a few times, and thereby resume control over ourselves, before we do anything else. It is only in that way that, in this particular situation, we would be able to make our being subject to these kinds of emotional outbursts (and this is of course a somewhat extreme example) compatible with our ultimately being governed by our practical reason, rather than by our impulses.

[57] Does Kant think that our various inclinations etc. are bad and to be regretted? No. Commenting on our various human-specific natural inclinations in the *Religion*, Kant writes: "Natural inclinations are, *considered in themselves, good*, i.e. unobjectionable, and to have a will to exterminate them would not only be futile, but also harmful and reprehensible; one must rather only tame them, so that they will *not* wear each other out, but instead can be brought to harmonize into a whole, called happiness." (R: 6:58)

[58] Also in (Smith 1995)

I will return to the relation between what the fully virtuous person would do in a particular situation, given the options that she would regard as open to her, and what actions are *permissible* in particular circumstances in the last chapter. As we will see, one useful way of thinking about whether an action is, as Kant puts it, in accordance with duty is to think of it as one of the actions that the fully virtuous person, operating in accordance with maxims of virtue, would regard as open to her in the situation she is in.[59]

6. On the Coming Chapters

I have now offered a glimpse of what is to come in the remaining chapters by briefly explaining how I understand the universal law formula, the humanity formula, and the relation between these two (which I did partly by introducing what I called the human nature formula). As I will argue in the chapters that follow, once we come to read Kant in this way – and I will argue at length that we should do so – many of the standard objections to Kant's ethical theory will be found to be invalid. The last thing I will do in this introductory chapter is to offer a map of the chapters that follow.

In addition to this introductory chapter, this dissertation has three further chapters. In the second chapter our subject will be that of how exactly to understand the universal law formula: both how to go about forming a better understanding of this formula and exactly how to understand the formula itself. We will start by outlining as many as seven major objections that have been raised against this formula and in the process also uncover the interpretative assumptions that underlie these objections. Having done that I

[59] And, again, this does *not* mean that the only permissible actions are those that a holy being would consider and/or take an interest in, but rather that the actions that are permissible are all those that, as Kant puts it, can "co-exist" with the autonomy of the will. (G: 4: 439) For more on Kant's views on permissibility, and how they relate to his views about virtuousness, see the discussion in the first half of chapter four.

will argue that these interpretations fail and then suggest 8 crucial steps we need to go through in order to form a better understanding of the universal law formula. Once we've done that we will arrive at the understanding sketched above – according to which maxims that could be universal laws are basic guiding principles that could guide all agents across situations and whose following would allow all agents to persist and flourish in their nature as autonomous human agents – and having done that, we will then explain how the standard objections fail once we understand the formula in this way.

We will already see in chapter two how (as I've briefly explained above) acting in accordance with self-adopted maxims that could be universal laws of the reason-endowed nature amounts to always treating the reason-endowed nature as a purpose in itself. Chapter three considers how to understand Kant's reasoning leading up to his initial statement of the humanity formula in the *Groundwork*. It starts by considering and criticizing Christine Korsgaard's much-discussed reconstruction of Kant's argument for the humanity formula. It also reviews other reconstructions that are similar to Korsgaard's (such as Allen Wood's), and argues that these reconstructions all fail, partly by failing to properly distinguish between the "absolute value" that Kant attributes to virtuousness and the status of being a "purpose in itself" that Kant attributes to the humanity within each person.

As is already clear from what has been said above, rather than being about the implications of attributing absolute value to humanity, Kant's argument for the principle of humanity is instead intimately tied to his constitutivism according to which, as I have said, the constitutive aim of autonomous human agency is the preservation and full realization of the humanity within all of us. Kant's main idea is that since we can only

robustly secure the kind of practical freedom that we assume in all practical deliberation if we make ourselves autonomous, and since doing so involves making the humanity in each person into our most basic purpose guiding all our actions (in accordance with the universal law formula, as it applies to us, i.e. in accordance with the human nature formula), reflection on the preconditions that must obtain in order for us to stably enjoy full practical freedom whereby we are fully responsible for our own actions ought directly to lead us to representing our own humanity as an end in itself, independently of what our own particular desires and personal needs are. In addition to further spelling out this argument of Kant's, chapter three also investigates Kant's ideas about the absolute (or non-relative) value of virtue, as these ideas relate to his concept of morality as involving a striving to be worthy of happiness.

Chapter four considers and responds (on Kant's behalf) to objections to Kant's theory (and in particular to the humanity formula) that are based on what I believe to be misunderstandings about Kant's views on permissibility, virtuousness, and what a morally ideal life for a human being would be like. Many writers (arguments recently given by T.M. Scanlon and Derek Parfit will serve as my main examples here) understand Kant as taking the *permissibility* of candidate courses of action to depend on what attitudes, motives, or principles we would be acting on in performing these actions: this, they in particular believe, is what determines whether or not these actions conform to the humanity formula. And objections to the humanity formula are then raised on the basis of how permissibility, according to other premises of these arguments, does *not* depend on the attitudes etc. of the agent doing the acting. But as I will explain, many of these objections rest on a simple failure to take note of Kant's distinction between what

he calls the "legality" and "morality" of our actions: whether our actions are merely in accordance with the moral law or whether they are also performed out of respect for it.

In criticizing Kant's philosophy, many writers also fail, as we will see, to take note of the view of the highest good for human beings that Kant sets out and returns to in several of his main works (and here our example will be an objection to Kant's theory raised by Rae Langton, which involves the real life example of Kant's own correspondence with one of his contemporary admirers, the young Austrian woman Maria von Herbert). Chapter four explains the relation between the humanity formula and Kant's theory of the highest good (both for individual people and for a possible world as a whole) and responds (on Kant's behalf) to the objection (which Langton raises) that Kant's moral theory sets out an unattractive ideal of human flourishing.

With the reading of Kant's principles that I will be arguing for roughly sketched, and the chapters to come briefly outlined, we're now ready to really start our discussion. We will start with the universal law formula. This formula, I will argue, is not well understood within contemporary debates, and most criticisms directed against it are, therefore, directed at a straw man.

CHAPTER II

Rediscovering the Universal Law Formula

1. Introduction

The three chapters on Kant's universal law formula in Derek Parfit's recent *On What Matters*[60] offer a helpful summary of many of the main objections that have been raised against this formula, and also add some new objections to it. But these chapters also, I believe, serve as an illustration of how this formula of Kant's has been almost universally misinterpreted.

Though there is no received view on how to interpret Kant's famous formula, there are various distinctive features that we can associate with the now most widely accepted readings. These features are what give rise to the various objections that Kant's formula allegedly faces. But the features of the main standard readings that give rise to these objections are, I will argue in this chapter, all based on interpretive mistakes and/or overlook key features of Kant's overall view.

The standard readings also have trouble making sense of Kant's claim that by subjecting ourselves to guiding principles (or "maxims") that could hold as universal laws, we would in effect be treating the humanity in each person as a purpose in itself, which means, as Kant puts it, that the universal law and humanity formulas are really just "different statements of the very same law."[61] This is sometimes called Kant's

[60] (Parfit 2011a)
[61] (G: 4:436)

"equivalence claim"[62] and most readers are puzzled by it, argue that it is false since the formulas have different substantive implications, or ignore it in their interpretations of these two different formulas.[63]

Since (1) the standard readings of the universal law formula leave it open to various seemingly decisive objections, and (2) these readings cannot make sense of Kant's equivalence claim, a charitable reader has good reason to suppose that there must be something wrong – perhaps fundamentally wrong – with the standard ways of understanding the universal law formula. And on closer inspection of Kant's texts it becomes, I shall argue, clear that the standard readings involve some obvious interpretative mistakes.[64]

In this chapter I will suggest eight steps that we need to take in order to form a better understand of Kant's universal law formula. In going through these steps and forming a new view of how to understand this formula (as well as of the so-called law of nature formula), we will as a result also directly come to see why it is that Kant claims that it is equivalent to the humanity formula. (Or, better yet, we will come to see why it is that the universal law formula, which is supposed to apply to all possible reason-endowed beings, is equivalent to the humanity formula when and insofar as it is applied to the reason-endowed beings of the human variety[65].)

Since our reading not only shows Kant's formula not to be open to the various seemingly devastating objections that have been raised against it, but also explains Kant's

[62] See, for example, (Allison 2011) and (O'Neill 1989).
[63] See section 4 below.
[64] …some of which, as we will see, have to do with problems with the English translations of key passages that many of Kant's critics and followers use.
[65] …and thus becomes what I called *the human nature formula* above: act only on that maxim that, through your will, could also become a universal law for the preservation and full realization of the particular human nature. (See section 4, chapter 1.)

equivalence claim, there is, I shall conclude, strong reason to replace the standard readings with this new reading. As for Kant's arguments for the universal law formula, I will not discuss these in this chapter. Nor will I suggest any new arguments for this formula. Those will be topics for another day.

The last thing I will mention in this introduction is this. Although this introduction frames our discussion as a reaction to Parfit's three chapters on the universal law formula in his recently published book, our discussion could equally well be thought of as a reaction to the various authors – such as Barbara Herman, Onora O'Neill, Thomas Hill, Christine Korsgaard, and Allen Wood – whose work Parfit is drawing on and adding to.

2. On The Standard Readings

In the *Groundwork for the Metaphysics of Morals*, Kant's main aim is nothing less than to "seek out and establish the highest-order principle of morality."[66] This would be the most general principle governing the decision-making of a "pure and good will", the possession of which Kant believes to be "the indispensible condition of the worthiness to be happy."[67] Kant ends up formulating his suggestion of what that principle is in several different ways, but the universal law formula is his chief formulation. In its

[66] (G: 4: 392)

[67] (G: 4: 393) It is an essential part of Kant's conception of "the doctrine of morals" that it is "not a teaching of how to be happy, but of how to be worthy of happiness." (KpV: 5:130); (TP: 8:278) At the same time Kant also thinks that ethics more generally concerns itself with the "the concept of freedom", and "the laws of freedom." (KU: 5:171); (G: 4:387) Kant unites these two aspects of his view of what ethics is in his theory that the moral law (i.e. the basic principle governing a good will) is itself the constitutive principle of a fully autonomous and positively free will. (G: 4: 453-4)

main *Groundwork* formulation, it reads: "act only on that maxim that you could at the same time will that it were a universal law".[68]

In the *Critique of Practical Reason*, Kant asserts that the formulation he uses there – "so act that the maxim of your will always at the same time could hold as a universal law" – is indeed the "foundational law of pure practical reason."[69] And in *The Metaphysics of Morals*, which is Kant's main work in substantive normative theory, Kant claims that we can derive both all externally enforceable rights that human beings have within a just juridical state and all internal ends of human virtue from the formulation he uses there[70]: "act on the basis of a maxim that can at the same time hold as a universal law." There can be no doubt, then, that Kant thinks that he has made a major discovery.[71]

There is no consensus about how exactly to understand the universal law formula, or the relation between its various formulations. But, as mentioned above, there are some fairly widely shared views about how to understand some of the key aspects of Kant's formula. The main aim of this section is to spell out some of these common assumptions. The idea here is not to sketch what might be called *the* standard reading, since there isn't one, but rather to spell out some key features associated with the main readings that give rise to the various seemingly decisive objections that have been raised against Kant's formula.

---

[68] (G: 4: 421)
[69] (KpV: 5: 30)
[70] (MS: 6:219)
[71] Marveling at all the implications he thinks we can derive from this simple principle, Kant remarks, as if to reassure his readers that he understands that they may be skeptical of his claims: "The simplicity of this law in comparison with the great and various consequences that can be drawn from it must seem astonishing at first, as must also its authority to command without appearing to carry any incentive with it." (6:225)

The first interpretative suggestion is what I will call *the actual maxim feature* of many of the main readings. This is the suggestion that Kant takes whether or not a particular course of action is permissible to depend on the actual maxim on which the agent performing the action is acting. As Parfit puts it (not thinking, as I understand him, that he is saying anything controversial): "Whether our acts are right or wrong, Kant claims, depends on our *maxims*, by which Kant usually means our policies and their underlying aims."[72]

The idea here is that, in stating his universal law principle, part of what Kant is claiming is that whether what, say, a person telling a lie is acting contrary to duty or not depends on what actual guiding principle this person is acting on in the situation she is in. If, on this reading, that actual maxim that the person is acting on (whatever it is) could not hold, or be willed, as a universal law, then that means that she acted in an impermissible way (whereas had she acted on some other maxim instead, then her lie could have been permissible in the circumstances).

The second interpretative suggestion we will consider I will call *the highly personalized maxims feature*: that the maxims that are possible candidates for universal laws can be very specific plans, policies, or norms about how to behave in very specific situations if you have certain particular goals, and if certain idiosyncratic things are true of you and those with whom you are interacting. Paul Dietrichson, for example, discusses the maxim "if I give birth to a baby weighing less than six pounds, I shall do everything in my power to kill it" and worries that Kant's "criteria of universalizability" would allow

[72] (Parfit 2011a: 275) In making this claim, as we will see, Parfit (and others who also make it) overlook Kant's distinction between the "morality" and "legality" of our actions, the latter of which corresponds to what Parfit means when he talks about whether or not an action is *wrong*. But it is, as we shall see, only the *former* that depends on what maxims we are acting on.

this to qualify as a universal law.[73] Barbara Herman similarly discusses and worries about the maxim "In order to avoid crowded tennis courts, I will play on Sunday mornings (when my neighbors are in church and the courts are free)"[74], and Allen Wood has a discussion of "the maxim of making a false promise on Tuesday, August 21, to a person named Hildreth Milton Flitcraft"[75].

This means, to return to the Dietrichson example, that this feature of the standard readings takes it that some rule that could only possibly be followed by women at the time of their lives when they could become pregnant (and who might give birth to babies weighing less than six pounds), could even so possibly be universal laws applying to all possible beings with practical reason (or at least to all human beings). Or, to use the Wood example, that a "maxim" of making a false promise on a specific date to a person with a particular unusual English name could serve as a guiding law for all reason-endowed beings.[76] This is so even though Kant's ethical theory, as he emphasizes, is supposed to apply to all possible beings with practical reason, not simply particular human beings with their personal desires and aims in the specific cultural circumstances in which they find themselves.

The third common interpretative suggestion is that whether some maxim could be a universal law is to be a matter *either* of whether it is logically conceivable that everyone

[73] (Dietrichson 1969) (Quoted in (Korsgaard 1996b: 82).)
[74] Herman doesn't think that this could be a universal law; the worry is instead that it would be okay to act on this maxim, even though it could *not* be a universal law. This implies that Herman thinks that this sort of rule about how to organize one's tennis schedule is of the general kind that Kant is talking about when he is talking about how we are to act on maxims that could be universal laws. See (Herman 1993: 138) and also (Wood 1999: 105-6)
[75] (Wood 1999: 102)
[76] I put "maxim" within quotes here since in giving this example, Wood is so obviously using the term in way that has very little to do with the way that Kant and others for whom "maxim" is a natural part of their language use it.

acts on it *or* of whether everyone would be able to achieve the aim contained in the maxim if it were universally acted upon. Commenting on Kant's formula in the introduction of *Utilitarianism*, John Stuart Mill, for example, writes of Kant that:

> when he begins to deduce from this precept any of the actual duties of morality, he fails, almost grotesquely, to show that there would be any contradiction, any logical (not to say physical) impossibility, in the adoption by all rational beings of the most outrageously immoral rules of conduct.[77]

On Mill's reading, then, whether something can be a universal law of action depends only on whether it is a maxim we can coherently imagine as universally acted upon without there being any logical (or physical) impossibility involved.

Korsgaard, in contrast, takes the "practical" view. She thinks that all "properly formulated" maxims specify some *aim* and some action meant as a *means* of achieving it. Some maxim could be a universal law, Korsgaard suggests, if and when its universal following would be consistent with everybody's being able to achieve the end contained within the maxim by taking the course of action suggested by the maxim.[78] Though Korsgaard favors this practical reading over the logical one (and though she also discusses a third "teleological" reading[79]), it is clear that both Korsgaard and most other commentators think that whether some maxim can be a universal law is a matter of whether its being universally followed is logically or "practically" possible, and that

[77] (Mill 1861/1963: 207)
[78] (Korsgaard 1996b: 92; 2009: 10)
[79] This is a reading on which whether something could be a universal law depends on whether it fits with nature's teleological purposes vis-à-vis the human being. (Korsgaard 1996b: 87-92)

that's the only "test" a maxim must pass. Call this *the either logical or practical feature*.[80]

The next feature of the standard readings has to do with the formulation of the universal law principle in the *Groundwork* that says that we are to act only on that maxim that we could at the same time *will* that it be a universal law. Although Kant only ever talks about whether we "could will" that some maxim were to become a universal law, it has become standard to add "rationally" between the "could" and the "will". Parfit, for example, writes that in various passages, "Kant seems to ask what we could *rationally* will, or choose" and that when we "apply these formulas, we must appeal to some beliefs about rationality and reasons" [81]. And, Parfit continues, for it to be interesting to discuss Kant's different formulations of the universal law formula, we should rely on what we ourselves take to be "true beliefs about rationality and reasons."[82]

Korsgaard similarly writes that the "Formula of Universal Law is a test of the sufficiency of the reasons for action and choice which are embodied in our maxims" and that each "interpretation must presuppose some notion of rationality in determining

[80] This interpretative suggestion is not altogether a mistake since Kant does indeed think, as we will see below, that it is a requirement for something to possibly be a universal law that it be possible for all to act on this principle and to have it as their maxim (i.e. basic guiding principle). Thus Kant sometimes uses examples in which he argues that something couldn't possibly be a universal law simply because it could not be a principle everyone could follow. In his second *Groundwork* example at (G: 4:422) illustrating reasoning using the law of nature formula, for example, Kant notes that a principle to make lying promises when we are in need of money couldn't be a universal law since nobody would trust anybody enough to lend them their money if it is was universally believed that people in financial need would make such lying promises. So Kant does indeed think that maxims fit to be universal laws must be able to function as guiding principles for all. This is, however, and as we shall see below, not a *sufficient* condition for a maxim to be fit to serve as a universal law. And the interpretative mistake involved in suggestions such as those of Mill and Korsgaard is precisely, therefore, to think it a sufficient condition.

[81] (Parfit 2011a: 285, 287)

[82] (Ibid.)

whether a rational being can will the universalization of a maxim."[83] Different writers who have agreed on this have, of course, suggested very different "notions of rationality". Rawls, for example, suggests that Kant is talking about what would maximize expected utility in a hypothetical choice situation in which you don't know what your own particular values are and in which you don't know the probabilities of being any particular person[84]. And Parfit suggests that we should use a notion of rationality according to which there are some "objectively" good ends – such as avoiding future agony – that any fully rational being must care about in order to qualify as rational or responsive to reasons.[85] We can call the common ground between these different interpretative suggestions *the rational willing feature* of the standard readings.

The fifth and last feature of the standard readings that I will highlight is *the interpretation in isolation feature*. By this I mean that it is, or at least seems, quite common that efforts to interpret the universal law formula pay limited attention to the wider context of Kant's philosophy as a whole, Kant's ethical theory as a whole, or even Kant's claims about the relations among the different moral principles he discusses. Here's just one example of what I mean.

In the *Groundwork* Kant claims that the universal law formula is equivalent to *the law of nature formula*: "so act that the maxim of your will could hold as a law of nature".[86] And in the *Critique of Practical Reason* Kant claims when we choose maxims on the basis of their fitness to serve as universal laws, the laws of nature can serve as the

[83] (Korsgaard 1996b: 79, 101)
[84] (Rawls 2000: 175)
[85] (Parfit 2011a: 285)
[86] (G: 4:422)

"type" on which to model these self-legislated moral laws.[87] Now Kant's theoretical philosophy also contains explanations of what Kant means both by "nature" in the different uses he thinks this word has as well as what Kant means by "laws of nature"[88]. But even so it is nevertheless the case that even those who take quite scholarly approaches to their interpretation of Kant's universal law formula – such as Korsgaard and Wood – don't use Kant's claims about how he understand *laws of nature* or about what he means by "nature" when they work their ways towards their readings of the universal law formula.

We have now pointed out five features associated with some of the most widely discussed interpretations of Kant's universal law formula: the actual maxim feature; the highly personalized maxim feature; the either logical or practical feature; the rational willing feature; and the interpretation in isolation feature. We can now turn to some of the many objections to Kant's formula that are based on these interpretative suggestions.

3. Seven Objections to the Universal Law Formula

We will consider as many as seven objections against the universal law formula. These are all objections intended to illustrate that when put to the test, Kant's formula turns out to have a whole host of completely outrageous implications.

As we are reviewing the objections and the allegedly crazy implications of Kant's principle, we will be relating these objections to the various different features of the standard readings discussed above and also to the different formulations of the universal

[87] (KpV: 5: 69-70)

[88] Not only that – the *Groundwork* and the *Critique of Practical Reason* themselves also, albeit more briefly, offer explanations of what Kant means by these things! See, for example, (G: 4: 422) (KvP: 5: 69-70) For Kant's explanations of his use of "nature" and his ideas about the laws of nature (and of different natures) within his theoretical philosophy, see sections 5 and 6 below.

law formula mentioned above. In the next section we will briefly review some of the pessimistic conclusions that even those favorable to Kant's moral theory draw on the basis of these types of objections. But we will also, more importantly, turn the discussion on its head by suggesting that in contrast to those other negative conclusions, a better conclusion to draw is that the problem must lie with the various features of the standard interpretations that give rise to these objections rather than with Kant's formula itself.

So if the following at times will feel like an unnecessarily drawn out refutation of an obviously doomed philosophical theory that couldn't possibly be made to work, please keep the following in mind. Our ultimate aim in this section is to raise doubts about the soundness of various features of the standard readings of the universal law formula that invite the objections we'll look at, *not* to illustrate how Parfit and others have decisively refuted Kant's formula.

Objection 1. We will start with the basic formulation of the universal law formula that Kant puts forward in the introduction to the *Metaphysics of Morals*: act on a maxim that can at the same time hold as a universal law. Parfit calls this the *impossibility formula* and the main objection he uses against it is a set of variations on an objection often associated with the work of Barbara Herman.[89] We can therefore call this *Herman's objection*.[90] The objection starts with the following observation.

In the *Groundwork*, Kant claims that strict duties (i.e. ones that admit of no exceptions) are established by showing that maxims permitting or prescribing such acts

[89] (Herman 1989)

[90] This might also be called *Mill's Objection* since this seems to be precisely the sort of objection Mill is hinting at in the quoted sentence above (which he, however, doesn't back up with any particular examples of what he has in mind).

couldn't be conceived of as universal laws.[91] We can furthermore safely assume that there is a moral duty not to kill, harm, or deceive others when this is beneficial to us and we can get away with it. But we can, this objection continues, imagine a world in which everybody successfully follows maxims such as "kill others whenever this promotes our self-interest and we can do so without being found out and punished".[92] This would surely not be a nice world to live in, and we could perhaps not sensibly desire for such a maxim to be a universal law, but it could, according to this objection, conceivably *be* a universal law. So, this objection concludes that Kant's impossibility formula fails to, as Parfit put it, "condemn" what Herman calls "convenience-killing".[93]

Herman's objection involves *the either logical or practical feature* of the standard readings: that whether some maxim could be a universal law depends *only* on whether it is logically or practically conceivable that everyone could act on it. Let us next turn to an objection that assumes what I have called *the actual maxim feature*: that the moral status of an agent's action depends on the actual maxim on which she is acting.

Objection 2. Parfit calls this objection *the mixed maxims objection*. Parfit assumes that the universal law formula takes the moral *permissibility* of an agent's action (in

---

[91] (G: 4: 424)

[92] Indeed, Hobbes's famous state of nature-based argument for why we ought all to obey a common sovereign crucially depends on our ability to conceive of such a world. (Hobbes 1651)

[93] Herman herself thinks that the availability of this objection shows that we must take the most important version of Kant's formula to be that according to which there is a requirement that we be able to *will* some maxim as a universal law without a "contradiction in the will". This idea, in turn, Herman understands as amounting to using maxims that spell out what we regard as valuable about particular courses of action, and the test, Herman claims, ends up being whether these evaluations of possible courses of action under certain descriptions are compatible with valuing rational beings as ends in themselves. The maxim of convenience-killing is not compatible with valuing rational beings in this way, and therefore this must be how to understand Kant, Herman concludes. But Herman's original objection, with which she starts her paper on this topic, would still stand: that there is no "contradiction in conception" in imagining a Hobbesian state of nature in which everyone follows the principle of killing others when it is convenient to do so and they can get away with it. (Herman 1989)

Kant's terms whether the action is "in accordance with" or "contrary to duty") to depend on what actual maxim the agent is acting on, and whether that maxim could be, or be willed, as a universal law. This can't be right, Parfit argues, since there are numerous possible maxims that it would sometimes be *right* to act on, but at other times obviously *wrong* to act on. One example, he claims, is the egoistic maxim "always do what is best for me".[94]

If a committed egoist realizes that saving a drowning child would give him some great benefit, then his egoistic maxim would lead him to save the child. This act would clearly not be contrary to duty. But, says Parfit, even egoists could *not* rationally will or desire that all people always act on the egoistic maxim. So, this means, Parfit takes it, that Kant's formula implies that it is wrong for the egoist to save the drowning child: if he did this, he'd be acting on the maxim "always do what is best for me" and that maxim cannot rationally be willed as a universal law.[95] That is a bad implication.

Objection 3. Consider next an objection that in addition to those other features also depends on *the highly personalized feature*: the interpretative assumption that maxims that could potentially be universal laws could be highly specialized, involving customized references to particular situations that agents with idiosyncratic properties are in. This is the objection that Parfit calls *the rarity objection*, which runs as follows.

Start by returning to the example from Wood of a "maxim of making a false promise on Tuesday, August 21, to a person named Hildreth Milton Flitcraft". The rarity of opportunities to act on this maxim helps to make it true, Wood suggests, that everybody could act on this maxim, which suggests, he takes it, that Kant's impossibility

[94] (Parfit 2011a: 289-93)
[95] (Ibid.)

formula implausibly implies that it would be permissible to make a false promise to somebody named Hildreth Milton Flitcraft if we were to encounter such a person on a Tuesday that happens to be August 21. Following Wood's lead, Parfit similarly suggests that if we come across a woman in a red dress, who is eating strawberries while reading the final pages of Spinoza's ethics, then we could make it permissible to steal her wallet by formulating a maxim that says to do so if we encounter such a woman and then acting on this eccentric maxim. It is, Wood and Parfit's arguments continue, not right to act in these ways, however, and so Kant's formula seems to generate the wrong results in these cases. If these particular objections succeed, then we could more generally or perhaps always make immoral actions morally permissible by adopting highly eccentric principles on the basis of which we'd perform these immoral actions. And that is surely a very bad consequence for a moral theory to have.

Objection 4. Turn next to *the threshold objection*, which makes use of the highly personalized feature and the actual maxim feature. This objection goes like this. There are some maxims, Parfit first claims, that have the following features: (1) everyone could conceivably act on them; (2) if the number of people who act on them is below some threshold, there is clearly nothing wrong about their acting in these particular ways; but (3) if everybody or people above the threshold act on this maxim, then what they'd be doing would clearly be wrong. Two examples Parfit uses are "have no children so as to be able to devote your life fully to philosophy" and "move to Iceland to absorb the spirit of the Nordic sagas".[96] It's perfectly fine if some people below some rough threshold act on these maxims. But it would be wrong for all people to do this, Parfit claims, since this

[96] (Parfit 2011a: 308-9)

would mean the end of the human race or at least result in chaos on Iceland. So, Kant's formula supposedly gets it wrong again.

Objection 5. Consider yet another objection that depends on the actual maxim feature of the standard readings, and also on Parfit's "value-based" assumptions about what agents could *rationally* will.[97] This objection returns to Herman's objection, but adds a new twist to it. A defender of Kant's theory might point out, we can first note, that though we can imagine a world in which people kill others when that benefits them and they can get away with it, we could not *rationally* will this as a universal law[98]: it would be bad for us to live in such a world, and as Hobbes argues, we have self-interested reasons to want there to be a central authority that harshly punishes violence and murder.[99] This is often, but *not always* true, Parfit claims. We might find ourselves in unusual circumstances, he suggests, in which we could rationally will, or choose, that everybody be permitted to cause great harm to others since in those particular rare circumstances this would be very good from the point of view of our personal self-interest.

It could happen, Parfit imagines, that you and I both have been bitten by a poisonous snake out in the desert, that you had prudently brought some antidote along with you, and that the only way for me to save myself would be to steal this antidote from

[97] What it is rational for a person to do always depends, Parfit takes it, on what is good for this person as an individual *or* on what is impartially good. Following (Sidgwick 1907), Parfit takes this to mean that while we could rationally become complete altruists who are always only interested in promoting the common good, we can equally rationally be egoists who are mainly concerned to do what is good for us on self-interested grounds. Such "value-based" theories of rationality differ from subjective theories, Parfit claims, since they assume the existence of objective values, whereas the latter understand rationally in terms of efficiency and consistency in the pursuit of desire-fulfillment. See chapters 5-6 in (Parfit 2011a), particularly pp. 130-40.

[98] This, as we saw above, is what Herman herself does in response to what I called Herman's objection above.

[99] (Hobbes 1651)

you.[100] Doing so would mean that you would die from your snake bite. It would clearly be wrong to steal your antidote, Parfit claims, but I could, in the rare circumstances I am in, rationally will that everybody be permitted to do great harm to others if the benefits to themselves are great enough. In these rare circumstances, it is self-interestedly rational for me to want everyone who has been bitten by a snake to be permitted to steal the only available antidote even if its rightful owner has also been bitten by the same snake.

We could normally not rationally will to live in worlds in which everybody's permitted to act in these ways, but this is, Parfit claims, a rare type of circumstance in which we could will this (namely, the type of situation in which the stakes are this high). And the moral status of an action we might perform depends, on Parfit's reading of the universal law formula, on what we could rationally will, in the particular situation we are currently in, that everybody be permitted to do. Again Kant's formula allegedly gets the wrong result. Parfit calls this *the high stakes objection*.[101]

Objection 6. Turn next to another objection that Parfit discusses that also depends on assumptions about what sorts of things we could rationally will that everybody always does. This objection, *the ideal world objection*, comes in two versions. Thomas Hill discusses the first version in connection to Kant's *Realm of Ends* principle, which says to act in ways that make a realm of ends possible (this being the possible state of affairs in which everyone treats the humanity in each person as a purpose in itself, and never as means only). The problem, Hill writes, is that

> acting in this world by rules designed for another can prove disastrous. In a world of perfectly rational rule-followers perhaps the imperative 'Never Lie' would

[100] (Parfit 2011a: 331)
[101] (Parfit 2011a: 331)

> make sense; but not in our world… To adopt our principles *as* ideal legislators seems a good idea; but to make them *for* ideal law-makers does not.[102]

Parfit's main version of Hill's objection notes that we could rationally will that everyone acts on the maxim "never use violence". So, we ought never, Parfit's objection takes Kant's formula to imply, to use violence. But if we were attacked and the only way of defending ourselves would be to use violence it would not, he asserts, be wrong to do so. So Kant's formula appears to generate the wrong result.[103]

We could furthermore, Parfit dramatically adds, rationally will that everybody acts on the following maxim: "Never use violence, unless some other people have used aggressive violence, in which case kill as many people as I can."[104] This gruesome maxim does allow us to defend ourselves when we are attacked. And we can rationally will that *everybody* follow it, Parfit takes it, since the world would then be a safe place. But, since everyone doesn't follow this maxim, this maxim, which we could supposedly will as a universal law, instructs us to kill as many people as we can. Kant's universal law principle, Parfit in effect concludes, offers homicidal maniacs a moral argument for the crazy conclusion that we are all permitted to go out and kill as many people as we possibly can.

Objection 7. As if this was not bad enough, consider next the last major objection Parfit launches against Kant's principle, what he calls *the non-reversibility objection*. This objection is based on *the rational willing feature* of the standard interpretations. It notes that on certain very common assumptions about rationality, people in privileged

[102] (Hill 1972: 314-5) Korsgaard similarly writes that "Kant's standard of conduct is designed for an ideal state of affairs: we are always to act as if we were living in the Kingdom of Ends, regardless of possible disastrous results." (Korsgaard 1996b: 149)
[103] (Parfit 2011a: 312)
[104] (Parfit 2011a: 315)

social positions can sometimes rationally choose a situation in which everybody follows norms that permit them to oppress other social groups.[105] Kant's universal law formula is open to this objection, Parfit claims, because what it tells us to ask ourselves is only whether we *ourselves* could rationally will that everybody be permitted and disposed to follow candidate maxims. And there are many people that, on most views of rationality, could rationally will that they be permitted to keep treating others in oppressive ways.[106]

To summarize this section: if Parfit, Dietrichson, Hill, Wood, Mill and company are right, Kant's universal law formula in its different formulations outrageously implies that we are morally permitted to engage in convenience-killing, mass-murder, deception when it benefits us etc., while at the same time forbidding us to save drowning children if we are egoists and requiring us not to devote our lives to philosophy or to move to Iceland, etc, etc. These alleged implications of Kant's principle, as I have shown, all depend on a number of standard assumptions about how to understand the principle. What should we make of all this?

4. Two Radically Different Conclusions We Can Draw

We have already seen that in Mill's assessment of the universal law formula, when Kant "begins to deduce from this precept any of the actual duties of morality, he fails, almost grotesquely, to show that there would be any contradiction, any logical (not to say physical) impossibility, in the adoption by all rational beings of the most outrageously immoral rules of conduct." What conclusions do Parfit, Wood, and the

[105] These are cases in which the roles cannot be reversed since one group possesses a characteristic the other group cannot acquire and vice versa, hence the name of the objection (if, that is, I am understanding Parfit correctly).
[106] (Parfit 2011a: 334-8)

others discussed above draw about the universal law formula based on the various different objections above?

Kant's impossibility formula, Parfit claims, "spectacularly fails".[107] The *Groundwork*'s willing formula, Parfit thinks, is better, but even so also fails. If we drop Kant's appeal to the agent's maxim, and the idea of what the agent herself could rationally will – and we instead ask what everybody could rationally will from the point of view of certain "objective" and "impartial" values – then we would, Parfit suggests, revise Kant's formula in "some wholly Kantian ways".[108] Such a view, Parfit claims, would be "remarkably successful" and all these objections would then "disappear".[109]

As some of Parfit's commentators point out, however, Parfit's "wholly Kantian" revisions of Kant's principle – especially their appeal to supposedly objective non-moral values that we know through direct rational intuition[110] – are deeply un-Kantian.[111] So it is better to conclude that if Parfit is right in his criticisms, then we should conclude with him that Kant's formula "spectacularly fails."

Allen Wood concludes that, as a principle intended to tell us which actions are right and wrong, the universal law formula is "radically defective"[112] and "pretty useless"[113] since it "systematically yields both false positives and false negatives when we try to employ it generally".[114] But Kant's formula is not even meant to serve as a basis from which to derive our moral duties, Wood also claims. It is, for Kant, "merely the first

[107] (Parfit 2011a: 327)
[108] (Parfit 2011a: 294)
[109] (Parfit 2011a: 342)
[110] (Parfit 2011b)
[111] See, for example, (Wolf 2011)
[112] (Wood 2006: 345)
[113] (Wood 2002: 172)
[114] (Wood 2006: 345)

stage in a philosophical search for the supreme principle of morality"[115]. Kant's different formulations of the universal law formula are just "the earliest and most abstract formulas Kant derives in the course of a progressive argument" and "also the least adequate expressions of the supreme principle of morality, and the poorest in practical consequences."[116]

Herman concludes that despite a "sad history of attempts […] no one has been able to make [the universal law formula] work"[117]. Hill in turn claims that, on its own, the universal law formula cannot even give us "even a loose and partial guide to action". [118] And Onora O'Neill suggests at one point that the universal law formula is not meant to serve as a principle on the basis of which we can decide what our duties are, but that it instead is intended by Kant as a criterion for determining whether our actions have or lack moral worth: whether in acting as we do, we are being virtuous.[119] At another point she suggests that as a principle for deciding what our duties are, the universal law formula gives either unacceptable guidance or no guidance at all.[120]

What about the relation between the universal law formula and the humanity formula? Wood, as we just saw, thinks that the universal law formula is just a first stumbling step on the way to a better principle, namely the principle of humanity, which Wood claims is "Kant's formula of choice for applying the moral law". Korsgaard thinks

---

[115] (Wood 1999: 110)
[116] (Wood 1999: 110) Critics using the kinds of objections discussed above fail to understand, Wood claims, that after discussing the universal law formula, Kant "proceeds immediately to specify the substantive value on the principle rests" (Wood 1999: 107) and that "Kant's formula of choice for applying the moral law is not [the universal law formula] but [the principle of humanity]" (Ibid.: 110).
[117] (Herman 1993: 104, 132)
[118] (Hill 2002: 122)
[119] (O'Neill 1989: 130)
[120] (O'Neill 1975: 125, 129)

that the universal law formula and the principle of humanity yield different conclusions when we use them in moral reasoning about important cases.[121] And Hill writes of Kant's so-called equivalence claim that it is not correct since the formula of humanity appears to "go beyond the famous first formula" since it, unlike the universal law formula, appears to declare "a rather substantive value judgment with significant practical implications."[122] Hill in effect agrees with Wood, then, that the formula of humanity is, as he puts it, "independent" of the universal law formula and that it is only or mostly from the former, and not the latter, that Kant is able to derive any substantive implications about what our duties are.[123]

Are these the right conclusions to draw? Or has something gone wrong, not with Kant's universal law formula, but instead with our understanding of it? On what we might call *the first option*, we accept one of the standard readings of the universal law formula and conclude (a) that it has crazy implications and (b) that it is disconnected from Kant's other moral principles, even though Kant himself claims that they are "at bottom only so many different formulations of the same basic law."[124] The writers above all support versions of the first option. But there is also a second option.

---

[121] Korsgaard thinks, for example, that the universal law formula does not have the implication that we must always, in all contexts, be completely truthful, whereas the humanity formula, as she interprets it, does have that implication. See (Korsgaard 1996b: 135-40, 152-3)

[122] (Hill 1980: 98)

[123] Hill thinks that we get specific duties out of the humanity formula on account of how it, as he interprets it, asserts a basic value to the rationality of human beings that trumps all other values. On the basis of this reading, Hill then worries that this assigns undue priority to rationality, and not enough value to human welfare. (Hill 1980) Hill seems to overlook that Kant, as we have already seen above, claims that treating the humanity in each person as an end itself importantly involves making their happiness into an end of ours quite independently of whether we think they are worthy of happiness or not.

[124] (G: 4:436)

On *the second option*, we reject the standard readings of the universal law formula, with the features that invite the objections above, and assume that the universal law formula (a) does not have all these crazy implications and (b) is closely related to Kant's other moral principles, just as Kant repeatedly says that it is. We then work our way toward a new reading of the universal law formula based on these two assumptions.

The *principle of charity* in interpretation favors the second option. Indeed, the principle of charity requires that we argue as follows: Step 1: charity requires us to assume that Kant understands the relation between his different principles and also that his main principle, which he calls "the highest-order principle of morality", does not generate outrageous implications. Step 2: the standard readings of this principle, with the features surveyed in section 2 above, imply that Kant is confused about the relations among his different principles and that the universal law formula generates outrageous implications. Conclusion: we must reject the standard readings, with their various troublesome features, and find a different reading instead.

How can all these other writers have gone so wrong in their attempts to understand Kant's universal law formula? In the sections to follow, in which I suggest various steps we need to take in order to work our way towards a better understand of Kant's formula, I will point out some key interpretative mistakes I believe most commentators make, some having to do with the use of poor translations of key passages, others having to do with missing important general points Kant himself makes about the details of his theory. But the main problem, I believe, is what I above have called *the interpretation in isolation feature*: that many readers assume that Kant is a writer whose individual works can easily be understood in isolation from his other works, and whose

individual claims can easily be understood independent of the general contexts in which they occur. I will expand on what I mean as we go along.

5. Eight Steps Towards a New Understanding of the Universal Law Formula

Having reviewed some of the objections against the universal law formula that seem most devastating, and having argued that they should lead us to question the standard readings of this formula, which also can't make sense of the relation between it and the humanity formula anyway, what I will do now is to suggest eight interpretative steps that, in my view, we need to take in order to form a better understanding of the universal law formula. Once we have gone through these eight steps and derived a new understanding of the universal law formula on the basis of these steps we will also, as we will see, be in a position to see why the humanity formula, just as Kant claims it is, turns out being equivalent to the universal law formula (at least as this formula applies to the human variety of what Kant calls the "reason-endowed nature"[125]). What follows will at times perhaps be a little provocative, but the previous sections have shown that we need to question the standard readings, so we shall need to not be very reverential towards the main Kant-interpreters in the contemporary debate.[126]

First Step: *Forget everything you've learned about how to understand Kant's universal law formula, so as to be able to start from scratch in your thinking about this formula.* This is, of course, the main part of what I meant when I said that what would follow would be a little provocative, but we've seen that there is a pressing need to

[125] To quickly see what I have in mind here, recall my discussion of what I called *the human nature formula* above.

[126] After all, it is their interpretations of Kant's formula that have given rise to all these seemingly decisive objections against it.

rethink our common ways of thinking of Kant's formula, and the best way to do so is to ignore the most common ideas about how to understand it dominating the debate. We shall also soon see that the most influential readings contain significant interpretative mistakes.

Step Two: *Find out what Kant regards as the ultimate basis for the existence of a moral law.* According to some highly influential Kant-scholars, most notably Allen Wood and Paul Guyer, the basis on which Kant's whole ethical theory is founded is the assertion of a *basic value*, namely the basic value of rational beings (according to Wood[127]) or of freedom (according to Guyer[128]). Thus Wood writes that "[p]erhaps the most fundamental proposition in Kant's ethical theory is that rational nature is the supreme value and the ground of whatever value anything else might possess."[129] Guyer writes that "the various formulations of the fundamental principle of morality that [Kant] offers can be understood as formulations of the rules necessary in order to realize the value of freedom."[130]

Now it is indeed true that the discussion in the *Groundwork* takes the unconditional goodness of a good will as its starting point. And Kant does also think, as we saw in footnote 67 above, that it is an essential part of the concept of *morality* that it has to do with the worthiness to be happy, a worthiness that he thinks is conditional upon having a good will. But Kant does, nevertheless, *not* think that the unconditional goodness of a good will is the ultimate basis of morality. His view is rather – as he very explicitly puts in the second section of the *Groundwork* – that:

---

[127] (Wood 1999; 2008)
[128] (Guyer 1998; 2006)
[129] (Wood 1999: 121)
[130] (Guyer 2006: 179)

> …it is of the greatest importance … just because moral laws are to hold for every rational being as such, to derive them from the general concept of a reason-endowed being as such…[131]

or, as he also puts it, that:

> …[the] question is therefore this: is it a necessary law for *all rational beings* always to appraise their actions in accordance with such maxims that they themselves could will to serve as universal laws? If there is such a law, then it must already be connected (completely a priori) with the concept of the will of a reason-endowed being as such.[132]

What Kant goes on to do in the third section of the *Groundwork* is precisely to argue that the basic moral law of choosing one's maxims on the basis of their fitness to serve as universal laws is the constitutive principle of a fully autonomous – and therefore positively free – will.[133] It is, Kant argues, only by subjecting ourselves to maxims that are fit to serve as universal laws that we can assume full control over ourselves and thereby robustly secure the kind of practical freedom we think of ourselves as having when we think of ourselves as beings with a will of our own. As Kant puts it in the introduction to *The Metaphysics of Morals*:

> The *freedom* of the power of choice is [an] independence from *determination* through sensual drives; this is the negative concept thereof. The positive one is: the capacity of pure reason to be practical for itself. This is however not possible except for through the subjection of the maxims of all actions under the condition of their fitness [to serve as] universal laws … [and] these laws of freedom … are called *moral*.[134]

In sum, Kant's view of the ultimate basis of morality – as he himself summarizes it in the very first sentence of the preface to the *Religion within the Limits of Reason Alone* – is that:

---

[131] (G: 4:412)
[132] (G: 4:426)
[133] (G: 4:452)
[134] (MS: 6:213-4)

> Morality … is based on the concept of the human being as a free being, but who, just for that reason, binds himself through his reason to unconditional laws…[135]

Our second main interpretative step, then, is to take note of how Kant, in this way, regards the universal law (which is the most basic formulation of the moral law) as the constitutive principle of an autonomous – and therefore positively free – will. And the reason I am suggesting that we take this step on the way to a new reading by reviewing various passages where Kant comments on what he regards as the ultimate basis of morality is partly a *negative* one: namely, that it allows us to see that we should *not* follow writers like Wood and Guyer, who claim that Kant's most basic moral principle is based on or asserts a basic value. Kant instead has *something else* in mind, something having to do with the autonomy of which we are capable as reason-endowed beings with a will of their own.

Step Three: *In constructing your theory of what it is for maxims to qualify as universal laws of autonomous (human) agency, focus on maxims that Kant puts forward as examples of maxims that <u>can</u> be universal laws, <u>not</u> on his examples of ones that are*

---

[135] (R: 6:3) I am here translating Kant's "die Moral" into *morality*, so in the above-quoted sentence "morality" means what we would ordinarily mean by it. When Kant uses "morality" in his special sense of the property an action might have of being performed out of the respect for the moral law, the word he uses is instead "Moralität," which I also translate into *morality*. The sentence above could, in other words, equally well have been translated into "ethics is based on the concept of the human being as a free being, etc."

I can here also note that in the preface to the Critique of Practical Reason, Kant similarly writes that, "freedom is the condition of the moral law, …[it] is the *ratio essendi* of the moral law." (KpV: 5: 4) And in his lectures on philosophical theology, Kant furthermore relates what he regards as the ultimate basis of morality to the idea of morality as the teaching of how to be worthy of happiness, when he states that, "morality contains the conditions, as regards the conduct of rational beings, under which alone they can be worthy of happiness. These conditions, these duties, are apodictically certain; for they are grounded in the nature of a rational and free being." (Kant 1817/1996: 28: 1072) In these same lectures he also states that, "*morals*, the whole system of duties, … is cognized *a priori* with apodictic certainty through pure reason. This absolute necessary morality of actions flows from the idea of a freely acting rational being and from the nature of actions themselves." (Kant 1817/1996: 28:2011)

not fit to be universal laws. I suggest this because most writers seem to construct their theory of what it is for maxims to qualify as universal laws on the basis of Kant's examples of maxims that *cannot* be universal laws. The usual strategy, however, is a bad one. The maxims that Kant rejects as candidate universal laws are all of very different kind than those put forward as examples of maxims that can be universal laws. The former are particular maxims about how to achieve particular ends or sets of ends; the latter are very general principles instructing us to adopt certain basic ends and basic principles about end-setting and end-pursuit as such.[136]

Here first are some maxims Kant uses as examples of ones that *cannot* be universal laws: "from self-love I make it my principle to shorten my life when its longer duration threatens more troubles than it promises agreeableness"; "when I believe myself to be in need of money I shall borrow money and promise to repay it, even though I know that this will never happen"; and "increase my [financial] assets by every safe means".[137]

Here, in contrast, are maxims Kant puts forward as examples of possible universal laws: "that I make it my maxim to act rightly is a demand ethics makes on me"[138]; "[i]n accordance with the ethical law … 'love your neighbor as yourself', the maxim of benevolence (practical love of human beings), is a duty of all human beings towards one another"[139]; "the maxim not to treat humanity as a means only"; "live in accordance with

[136] See next step for more details.
[137] (G: 4:422); (KpV: 5: 27)
[138] What does "acting rightly" mean here? It means following (what Kant presents as) the highest-order principle of externally enforceable law-giving within a just juridical state, namely that "Any action is right if it can coexist with everyone's freedom in accordance with a universal law, or if on its maxim the freedom of choice of each can coexist with everyone's freedom in accordance with a universal law" (MS: 6:230)
[139] A stronger variation of this *maxim of benevolence*, which Kant cites with approval in the *Religion*, is this: "Love everyone as yourself, i.e. promote his welfare from an unmediated good-will, on not derived from selfish incentives". This maxim, and also the more general maxim to

your nature"[140]; "make yourself more perfect than your nature has already made you"; and (similarly) "build your mental and bodily powers for fitness for all ends that may spring out from you".[141]

The members of the former set are all principles about how to pursue specific ends; the members of the latter are all very general principles instructing us to adopt certain ends or principles about end-setting and end-pursuit as such. This is almost universally overlooked.

Korsgaard, for example, claims that all "properly formulated" maxims specify some particular *end* and the *means* intended to achieve it.[142] But this suggestion clearly seems modeled on Kant's examples of maxims that cannot be universal laws, *not* on his examples of maxims that can be.[143] If we look at the latter set, then we immediately see that they do *not* have the features Korsgaard (and many others) claim that all "properly formulated" maxims have. All these writers, I suggest, model their understanding on the wrong set of examples. This (to be a little provocative again) is further reason to start from scratch and ignore what we have previously learned about Kant's formula.

Step Four: *Take a <u>close</u> look at what Kant says when he explains what he means by "maxims" in his formula.* This step crucially involves using better translations of some of Kant's key claims makes about maxims than the ones usually used. What I especially

---

"Do your duty from no other incentive except the unmediated appreciation for duty itself", are two maxims about which Kant writes that, "these commands are not merely laws of virtue but precepts of *holiness,* which we ought to strive after … in view of which the striving itself is called *virtue*." (R: 6: 161)

[140] We'll get to the crucial question of what Kant means by "nature" below.

[141] (MS: 6: 231; 451; 449; 419)

[142] (Korsgaard 1996b: 82; 2009: 10)

[143] Nor does, to give another example, Herman's claim in (Herman 1989) that maxims that could be universal laws pick out the features of possible courses of action that we see as valuable about these courses of action seem to at all fit with Kant's own actual examples of maxims that can hold as universal laws.

have in mind is the term "Grundsätze", which is usually simply translated into "principles", but which ought really to be translated into "basic principles" or "foundational principles".

"Grund" literally means "ground", but "basis" or "foundation" seems like a better English translation, given the context. And "sätze" is the plural of "satz", which means "proposition," but can also acceptably be translated into "principle". Reading the passages I have in mind with the right translation, i.e. with "basic principle" rather than merely "principle" for "Grundsatz", is eye-opening.

Consider for example this key passage, which is the very first set of claims Kant makes in the introductory chapter of the *Critique of Practical Reason* called "On the Foundational Principles of Pure Practical Reason":

> Practical foundational principles are propositions that contain a universal determination of the will, which have several practical rules under them. They are subjective, or *maxims*, when the condition is regarded by the subject as holding only for her will; [they are] objective, however, or practical *laws*, when cognized as objective, i.e. as valid for each reason-endowed being.[144]

Consider also this footnote from the *Groundwork* that explains the idea of a maxim:

> A *maxim* is the subjective principle of action, and must be distinguished from the *objective* principle, namely the practical law. The former … [is] the *basic principle* in accordance with which the subject *acts*; the law, however, is the objective principle valid for every reason-endowed being, and the basic principle according to which she *ought to* act, i.e. an imperative.[145]

---

[144] (KpV: 5:19)

[145] (G: 4:422, emphasis added) We can further note that in *The Metaphysics of Morals*, Kant explains the relation between basic subjective principles of action and objectively valid principles in the following way:

> The categorical imperative, which at all only asserts what obligation is, is: act on a maxim that can at the same time hold as a universal law. Thus you must first consider

The self-adopted or self-adoptable maxims that can be universal laws are *not*, then, specific rules about how to conduct oneself in specific situations in view of particular ends. They are, instead, "universal determinations of the will" that may contain "several rules under them". They are "basic principles" of the will. (This, we can also note, fits perfectly with the examples we looked at above of maxims that *can* be universal laws.)

What would a principle need to be like in order to be both a "universal determination of the will" and also "valid for each reason-endowed being"? I suggest that what Kant means is not simply a very general principle, but also a principle that (a) could offer *cross-situational guidance*[146] for (b) *all* reason-endowed beings[147]. That is to say, in order to potentially be a universal law, some basic principle of end-setting and end-pursuit would need to be able to guide us not just in particular situations and in view of particular ends. It would need to be able to robustly guide our decision-making across different situations, independently of what, as Kant puts it, "our private ends" are. It needs, in other words, to be what Kant calls a "formal principle" of the will: a guiding principle that "abstracts from subjective ends".[148]

Why think that this is Kant's view? Because Kant writes that "material

---

> your actions according to their subjective basic principles; however, whether this basic principle is also objectively valid, you can only know in the following way: that when your reason subjects it to the test of conceiving yourself as giving universal law through it, it qualifies for such a giving of universal law. (MS: 6:225)

We furthermore get the same thing already in the *Critique of Pure Reason*, where Kant writes, "Practical laws, insofar as they are the subjective reasons for action, i.e. the subjective basic principles, are called *maxims*." (KrV: B840; A 812) Kant in other words defines *maxims* as basic principles in all of the first and second *Critiques*, the *Groundwork*, and *The Metaphysics of Morals*. But with the poor English translations of these passages commonly used, which simply say "principles", this unfortunately goes completely unnoticed.

[146] … and thereby contain a "universal determination of the will".

[147] … and thereby be "valid for each reason-endowed being".

[148] (G: 4:427)

principles", which are principles related to particular ends or sets of ends, "are all, as such, of the same kind, and belong under the general principle of self-love or of our own happiness".[149] And material principles of self-love and happiness, Kant further writes, "are all empirical and cannot give rise to any practical laws."[150]

To understand what Kant means by calling these "principles of our own happiness" and saying that they are "merely empirical" in the *Critique of Practical Reason*, we can, for example, consult the following crucial passage in the *Critique of Pure Reason*:

> The practical is anything that is possible through freedom. If, however, the conditions of the exercise of free choice are empirical, then reason cannot have anything but regulative use, and only serve as a unifier of empirical laws, as for example in the teaching of prudence with the unification of all ends given to us by desires into one end, happiness, and the harmonization of the means of achieving it, meaning that reason could only have as its business the achievement of sensibly recommended ends, and give rise to no purely a priori laws. In contrast to this, pure practical laws would give us ends a priori, not empirically conditioned, but absolutely commanded ends, which would be products of pure reason. These are the moral laws, hence they belong only to the practical use of reason, and admit of a canon.[151]

As we can see here, what Kant means by "empirical principles" of reason are more or less complex means-end principles about how to achieve specific ends or sets of ends. Our own happiness, we can further see, refers to our various personal ends unified into a coherent whole.

Notice that these passages from the first and second *Critiques* clash directly with Korsgaard's widely accepted claim that all properly formulated maxims are principles saying what to do – what *means* to take – in order to achieve specific ends. That is

[149] (KpV: 5:22)
[150] (KvP: 5:21)
[151] (KrV: A800, B828)

exactly the kind of "empirical" principle that Kant claims couldn't possibly qualify as a universal law. What can qualify as universal laws are instead general principles that prescribe general ends ("make yourself more perfect than nature has already made you" etc.) or principles about end-setting and end-pursuit as such ("the maxim not to treat humanity as a means only" etc.).

The crucial fourth step, then, is to have a close look at what Kant says about the nature of maxims and the basic guiding practical principles that, in his judgment, can be universal laws. And what we see – once we use a more precise translation of some of the key passages – is that maxims that are candidate universal laws are basic principles containing "universal determinations of the will" that are "valid for each reason-endowed being". This means, I have suggested, principles prescribing general ends and basic principles of end-setting and end-pursuit as such that can robustly guide us across situations and that could also guide all reason-endowed beings across situations.

The first four steps to take, then, are the following: (1) to forget what we've learned about how to understand the universal law formula and instead start from scratch; (2) to take note of how Kant's strategy is to try to derive the moral law from the concept of the reason-endowed being as a self-determining being with a will; (3) to build our interpretation on Kant's examples of maxims that, according to him, can be universal laws, *not* on ones Kant claims can't be universal laws; and (4) to take a close look, using precise translations, at what Kant says in passages that introduce the idea of a *maxim*, as it is used in the universal law formula and to notice that in these passages, Kant presents maxims as "basic principles" containing a "universal determination of the will" under which "there may be many practical rules".

5. Eight Steps Towards a New Reading, Continued

Having focused on the "maxim" part of the universal law formula, we will now move on to the part of the formula according to which we should choose our maxims or basic guiding principles *on the basis of their fitness to serve as universal laws*. The first thing we must do is the following.

Step Five: *Pay close attention to Kant's claim that the universal law formula is equivalent to the law of nature formula.* Just after stating the universal law formula in the second section of the *Groundwork*, Kant immediately claims it is equivalent to the law of nature formula: "so act that the maxim of your action through your will were to be a universal law of nature."[152] Kant then illustrates his formula with the four famous examples, not by using the universal law, but instead by using the law of nature formula. And in the *Critique of Practical Reason* he claims that though moral laws are not laws of nature, the laws of nature are nevertheless to serve as the "type" on which we are to model the maxims that we choose as laws of our own nature.[153] This means that in trying to come to understand the universal law formula, we cannot ignore, but ought instead to focus much of our attention on, the law of nature formula. This leads directly to the next interpretative suggested step.

Step Six: *Investigate what Kant means by "nature" within his philosophy in general, and in particular what he means by the phrase "the reason-endowed nature" within his practical philosophy (and even more specifically what he means when he talks about "the human nature").* We have just seen in the foregoing step that Kant thinks that

[152] (G: 4:429)
[153] (KpV: 5: 67-71)

there is an important analogy to be made between agents – or what he calls "the reason-endowed nature"[154] – operating according to moral laws and the rest of nature operating in accordance with natural laws.[155] This means that coming to understand what Kant means when he talks about *nature* and *laws of nature* in general will help us understand what he means by maxims that could be universal laws of the "reason-endowed nature" (and, for that reason, of the human variety of the reason-endowed nature) in particular.

The word "nature", Kant claims, has both *formal* and *material* senses. He explains these senses in the *Metaphysical Foundations of Natural Science*, which appeared the year right after the *Groundwork* was published, in the following way:

> If the word nature is taken simply in its *formal* meaning, where it means the first inner principle of all that belongs to the existence of a thing, then there can be as many different natural sciences as there are specifically different things, each of which must contain its own peculiar inner principle of the determinations belonging to its existence. But nature is also taken otherwise in its *material* meaning, not as a constitution, but as the sum total of all things, insofar as they can be the *objects of our senses*, and thus also of experience. Nature, in this meaning, is therefore understood as the whole of all appearances, that is, the sensible world, excluding all nonsensible objects.[156]

When Kant puts the law of nature formula he uses "nature" in what he above calls the *formal* sense. Indeed, just before translating the universal law formula into the law of nature variation, Kant writes that it is "the universality of laws, in accordance with which effects occur, [that] constitutes what is actually understood by *nature* (in accordance with form), i.e. the existence of things, insofar as it is determined by universal laws."[157] What Kant means specifically by the "reason-endowed nature", then, is the particular kind of

[154] (G: 4:437)
[155] (G: 4: 412); (KpV: 5: 67-72)
[156] (MN: 4:467)
[157] (G: 4:429)

constitution that reason-endowed beings have in their existence as such beings (the "human nature" being the particular variety of that more general kind that we constitute[158]).

Step Seven: *Investigate what Kant means by "laws" within his philosophy in general, and what he means by "laws of nature" in particular.* Knowing what Kant means by "nature" in the law of nature formula, the next thing we need to know is, of course, what he means by a *law* of nature (and/or a law of a particular type of nature). Consider first the following passage from the Kantian logic manual that Kant commissioned his former student Jächse to write on his behalf, on the basis of his lecture notes, in the year 1800:

> All rules according to which the understanding operates are either *necessary* or *contingent*. The former are those without which no use of the understanding would be possible at all … If now we put aside all cognition that we have to borrow from *objects* and merely reflect on the use just of the understanding, we discover those of its rules which are necessary without qualification, for every purpose and without regard to any particular objects of thought, because without them we would not think at all … [T]his science of the necessary laws of the understanding and of reason in general … we call *logic*.[159]

If the laws of logic are the pure principles of the understanding we must subject ourselves to, or think in terms of, in order to at all be able to think coherently, then what are laws of a nature, and in particular the laws of the reason-endowed nature? It is hard to find elaborate explanations of what Kant means by laws of nature (or the laws of a particular kind of nature), but the following claim from the *Metaphysical Foundations of Natural Science* is helpful:

---

[158] Thus Kant often uses phrases such as "the human and every other reason-endowed nature," as he does, for example, at (G: 4: 428) and (G: 4: 431).
[159](Kant 1800/1992: 528)

> Laws [are] principles of the necessity of that which belongs to the existence of a thing. (4:469)

These would be laws of what Kant calls a particular nature in the *formal* sense of nature. Consider also the following remark that Kant makes in the *Prolegomena* about the laws of *material* nature (i.e. the totality of all that can be the objects of our senses):

> lawfulness in the connection of appearances, i.e., nature in general, [is something that] we cannot come to know through any experience, because experience itself has need of such laws, which lie *a priori at the basis of its possibility*.[160]

On Kant's view, then, laws of particular natures are basic principles stating what is necessary for the possibility of the existence of things with particular kinds of constitution (i.e. with particular natures). (Laws of material nature similarly state what is necessary in order for us to be able to have coherent and unified experiences on the basic of our various perceptions.[161]) Or so I suggest that we understand Kant's view.[162]

---

[160] (P: 4: 318-9, emphasis added)

[161] Thus Kant writes, in the *Prolegomena*, that, "…the possibility of experience in general is […] at the same time the universal law of nature, and the principles of the former are themselves the laws of the latter. For we are not acquainted with nature except as the sum total of appearances, i.e., of the representations in us, and so *we cannot get the laws of their connection from anywhere else except the principles of their connection in us, i.e., from the conditions of necessary unification in one consciousness, which unification constitutes the possibility of experience*." (P: 4:319, emphasis added)

[162] For (what appears to be) an alternative view, see, for example, Kitcher (1986), according to which (if I understand this paper correctly) laws, for Kant, are but true general principles that don't always contain anything about what is *necessary* in relation to some thing or phenomenon. Mary Gregor's discussion of "the laws of freedom" at the basis of morality in (Gregor 1963) seems to understand laws in the sense I suggest above, since Gregor (I believe rightly, and in contrast to most contemporary Anglophone commentators) understand the laws of morality as being the most general principles it is necessary for us to follow in order to for us to enjoy negative and positive freedom of the will. Gregor's book is, I think, not receiving the full attention it deserves in the contemporary discussion of how to understand Kant's moral theory (and *The Metaphysics of Morals* in particular).

Step Eight: *Find out what, according to Kant, is distinctive of the reason-endowed nature in general, and the human variety of the reason-endowed nature in particular.* If laws of specific kinds of nature are the basic principles containing the necessary preconditions for the existence of things of the given kinds of nature, and a nature (in the formal sense) is a particular constitution, then what, we next need to know, is distinctive of the reason-endowed nature? What is the distinctive feature, in other words, of the reason-endowed nature such that this feature can only exist in accordance with particular basic principles? We have already seen the answer in our second step: it is practical freedom or agency.

What is distinctive about the reason-endowed nature in general, which sets it apart from the rest of nature, is – Kant thus claims – that whereas "everything in nature operates in accordance with laws … only the reason-endowed being has the capacity to operate in accordance with a *representation* of a law, i.e. to act on principles, or a *will*."[163] And to think of oneself as having a will, Kant also argues, is to think of oneself as not being wholly determined in one's decision-making by things outside of one's own practical reason, such as one's own or other people's impulses and whims. What the laws of the reason-endowed nature in general are laws of, then, is the existence of a kind of being, or nature, that is governed by its own capacity to both represent and act on principles.[164]

Now it is furthermore, Kant also claims, distinctive of us as *human beings* that we have a *dual* nature whereby at the same as we are reason-endowed beings with a will, we

[163] (G: 4:412)
[164] Kant somewhat dramatically calls such a nature a "supersensible nature". Thus in the *Critique of Practical Reason*, Kant writes, "The moral law is really a law of causality through freedom, and thus of the possibility of a supersensible nature." (KpV: 5:47)

are also *animals* with drives, needs, desires, fears, and other motivational status in relation to which we are, as some writers put it, *passive*[165] (which is to say that we don't ourselves decide to have these various inclinations, desires, needs, etc.). This dual nature that we have as animals with needs, drives, desires etc. who are nevertheless at the same time capable of autonomous agency is what, as Kant thinks of it, constitutes our distinctive human nature or, as he also puts it, our *humanity*.[166]

What the laws of the *human* variety of the reason-endowed nature are laws of, then, is the existence and realization of the nature of a being that while being subject to various drives, inclinations, needs, desires, etc., (and therefore having a resulting general desire for happiness, which is the state where everything goes in accordance with our desires and wishes), also is capable of self-determining, autonomous agency. Again, as Kant puts it in the *Metaphysics of Morals*:

> Human choice … is of such a sort that although it is affected by drives, it is not determined thereby [and yet] is in itself (without the acquired skillfulness of reason) not pure, but nevertheless capable of being determined by a pure will. The *freedom* of the capacity for choice is this independence from *determination* through sensual drives; this is the negative concept thereof. The positive one is: the capacity of pure reason to be practical for itself. This is however not possible except for through the subjection of the maxims of all actions under the condition of their fitness [to serve as] universal laws … [and] these laws of freedom … are called *moral*.[167]

We can now summarize these last four steps. In addition to the initial steps (1)-(4) that I suggested in the foregoing section, these last steps towards a better understanding of Kant's universal law formula that I have suggested are: (5) take note of how the universal law formula is said to be equivalent to the law of nature formula; (6) note that "nature",

[165] See for example the introduction to (Korsgaard 1996b).
[166] See, for example, (MS: 6: 420)
[167] (MS: 6:213-4)

in the *formal* sense relevant to the law of nature formula, is supposed to pick out a particular kind of entity with a particular constitution; (7) note that laws of particular natures within Kant's philosophy are basic principles in accordance with things with certain kind of nature can exist and operate; and, finally, (8) note that what is distinctive about the reason-governed nature in general is the capacity to exist and operate in accordance with representations of laws, (i.e. guiding principles that have properties making them fit to serve as laws), *and* that what is distinctive about the human variety of the reason-endowed being in particular is that we are capable of autonomous self-governance while at the same time it is the case that we, as human *animals*, are subject to various needs, desires, and other motivational influences in relation to which we are passive[168], and that we have a general desire for happiness (i.e. the possible state of affairs in which things, as far as possible, go in accordance with our desires and wishes).

6. How to Understand the Universal Law Formula (and Its Relation to the Humanity Formula)

Having taken these eight steps we are now in a position to put forward our new reading of Kant's universal law formula. And once we state this reading, it will be directly clear why the universal law formula, as it applies to us human beings, is equivalent to, or directly implies, the humanity formula. So let's get right to it.

We can begin by quickly stating some of the distinctive features that we have seen that Kant attributes to the maxims that, in his view, are fit to serve as universal laws.

[168] As noted earlier on, Kant doesn't regard our natural inclinations as evils, but writes instead that "Natural inclinations are, *considered in themselves, good*, i.e. unobjectionable, and to have a will to exterminate them would not only be futile, but also harmful and reprehensible; one must rather only tame them, so that they will *not* wear each other out, but instead can be brought to harmonize into a whole, called happiness." (R: 6:58)

What is distinctive of such maxims is, *firstly*, that they are *self-adopted basic guiding principles* that all others could also adopt and act on, and that we could be guided by across situations (see in particular step four). This we can call the basic guidance condition.[169] But what is also distinctive of these maxims is, *secondly*, that they are basic guiding principles that everyone could act on and, *as a result of this*, both *preserve* and *fully realize* their particular nature as beings capable of autonomous self-determination through the rule of their own practical reason (see in particular steps five through seven). And in the case of our particular *human* variety of this reason-endowed nature, what is distinctive of basic guiding principles that could at the same time serve as laws of our *human nature* is furthermore, *thirdly*, that they are guiding principles whose following would allow us to also make our "animal" side with its associated needs, desires, and general wish for happiness to fully harmonize with the capacity for autonomy that we at the same time possess (see in particular step eight).

This means that the maxims that are fit to serve as universal laws of our human nature are principles in accordance with which we can flourish in accordance with *both* the "moral" side *and* the "animal" side of our nature: maxims in accordance with which we can both achieve *autonomy* and *happiness*. Maxims that could serve as universal laws of our particular variety of the reason-endowed nature are, then, basic guiding principles all human beings could adopt and follow, and whose following would allow for, and help to bring about, the preservation and full realization of our distinctive humanity. As I already suggested in chapter one, the universal law formula as it applies to human beings

[169] This shows that what I above called the either logical or practical feature of the standard readings (according to which maxims that can be universal laws are ones we can all conceivably or practically act on) is indeed a necessary condition on a maxim to be fit to serve as a universal law, but we have also seen that it is not sufficient.

in particular can, therefore, also be restated as what I above have called *the human nature formula*: act on the basis of self-adopted guiding principles that all could act on, and whose following would help to preserve and fully realize the distinctive human nature within all of us.[170]

Note that in telling us to *act* on the basis of *maxims* that are fit to serve as universal laws, the universal law formula in effect instructs us to do two things: (1) to *adopt* basic guiding principles that have certain properties, and (2) to also *act* on the basis of, and in accordance with, these guiding principles. A *maxim*, as Kant uses the term, is a basic guiding principle or "rule" that " the subject makes into a principle for herself,"[171] and this means that if we are not already operating in accordance with self-adopted basic principles in accordance with which we and all others could flourish in our distinctive human nature, then the universal law formula (as it applies to us in the form of the human nature formula) instructs us to adopt such principles.

The universal law formula also, of course, instructs us to *act* in accordance with these principles.[172] But that, as I am pointing out, is not *all* that it does. That the universal law formula not only instructs us to act in accordance with maxims that could be universal laws, but that it also, in effect, instructs us to adopt such maxims and to make

---

[170] Now if we were to come into contact with other reason-endowed beings, then the universal law would, it seems, forbid us to limit ourselves to acting on the human nature formula, and we would instead be under a requirement to make our acting on this particular formula compatible with the preservation and full realization of those other reason-endowed natures as well (insofar as they, too, were so constituted that they could autonomously conduct themselves in accordance with maxims that could serve as laws we would all be subject to).

[171] (MS: 6:225)

[172] It is not enough, Kant thinks, that we have the right mindset: the moral law also requires of us that we actually act in accordance with it. Thus Kant cites with strong approval what he interprets as a part of Jesus' moral teaching according to which (as Kant puts it), although "the pure moral disposition of the heart" is "above all the goal for which the human being should strive", it nevertheless is *also* the case that "these pure dispositions … should also be demonstrated in *deeds*." (R: 6:159-61, emphasis in original)

them into our operative guiding principles, does of course have to do with how what is distinctive of our nature – as Kant thinks of it – is that we have the capacity to operate in accordance with laws that we give to ourselves (our capacity for autonomy). This aspect of our nature cannot be fully realized if we merely act in ways that conform to principles that are fit to serve as laws of our nature: the realization of this aspect of our nature also require that we actually do adopt such principles as our guiding principles.

Now if we choose our basic guiding principles on the basis of their being guiding principles all could adopt and whose following would help to secure the preservation and full realization of the humanity within all of us, we thereby make the humanity within each person into the most general purpose around which our actions are organized. We also make sure that our own pursuit of happiness is always compatible with the preservation and full realization of the humanity within each person. In so doing we abstain from ever treating the humanity in any person as a mere means that exists only for the sake of our own pursuit of happiness.

This means that in subjecting ourselves to maxims in accordance with which our human variety of the autonomous agency can be preserved and fully realized, what we are thereby doing is just what the humanity formula instructs us to do. We would so act that we never treat the humanity in any person as a means only, but instead at the same time always make sure that we treat the preservation and full realization of the humanity within each person as a purpose in itself.[173] This is why, just as Kant claims, the

[173] One way of putting this point is to say that the humanity formula is itself, out of logical necessity, the most general maxim that is fit to serve as a universal law with regard to the human variety of the reason-endowed nature. This is, I suggest, why Kant at one point in the *Groundwork* says that all morally acceptable maxims have a *form,* given by the universal law formula, but also at the same time a *matter,* which is specified by the humanity formula. (G: 4: 436) This way of putting things – i.e. taking the humanity formula to be the most general maxim

humanity formula is equivalent to the universal law formula (as the latter applies to us human beings in particular).

7. Why the Seven Objections Discussed Above All Misfire

With these conclusions that we have drawn on the basis of the eight interpretative steps suggested above in mind, we can now return to the five features associated with the standard readings reviewed above, and, even more importantly, the seven objections against the universal law formula also reviewed above. Since these seven objections depend on those five interpretative suggestions associated with the standard readings, whether those objections succeed partly depends whether those five suggestions are correct. Whether the seven objections succeed also depend on whether they respect the various interpretive constraints we have put forward above in our eight steps towards a new reading. We will quickly find that this is not the case, that the five features

---

that could be willed as a universal law of the human variety of the reason-endowed nature – also fits with how Kant introduces the idea of the humanity within each person as a purpose in itself in the second book of *The Metaphysics of Morals*. Thus immediately after stating the version of the universal law formula that he uses there – "act on a maxim of ends the having of which could hold as a universal law" – Kant directly states that:

> In accordance with this principle the human being is an end to herself as well as to others, and it is not enough that she not be permitted to treat any human being as an end only (since this is compatible with indifference), but to make the human being as such into an end is in itself a duty for human beings. (MS: 6:395)

But rather than saying that the humanity formula is the most general maxim that is fit to serve as a universal law, we can equally well say, I believe, that as it applies to human beings, the universal law formula turns into the human nature formula (act on the basis of maxims all human beings could adopt, and in accordance with which the human nature in all of us could be preserved and fully realized), which is equivalent to the humanity formula (since acting in the accordance with the former would amount to making the preservation and full realization of the humanity within each person the most general purpose around which all our actions were organized).

associated with the various standard readings are interpretive mistakes, and that the seven objections, therefore, all misfire.

Let's start with the *actual maxim* feature: the assumption, that is, that the universal law formula takes the moral *permissibility* of possible courses of action to always depend on the actual maxim an agent is acting on. The interpretative assumption, in other words, that whether some course of action would be a right or wrong (permissible or impermissible) thing to do depends on what maxim we would actually be acting on, and on whether that maxim is fit to serve as a universal law. This is a mistaken assumption. It confuses two different ways in which there can be conflicts between the ways agents conduct themselves and the universal law formula. And it also overlooks Kant's stated view of the permissibility of actions and, along with it, his corresponding distinction between the *legality* and *morality* of our actions.

As I already noted above in section 6, the universal law formula – in telling us to act on the basis of maxims that are fit to serve as universal laws – tells us *both* to act in accordance with maxims fit to be universal laws, *and* to adopt, and govern ourselves on the basis of, such maxims. This means that one kind of clash that can occur between the way in which we actually conduct ourselves and the way in which the universal law formula instructs us to conduct ourselves occurs when our actions are not regulated by self-adopted maxims that are fit to serve as universal laws. That is to say, we may fail to have a set of basic guiding principles fit to serve as universal laws that serve as the basic governing powers underlying and controlling our choices of actions; our way of conducting ourselves may, therefore and thereby, be clashing with the requirement to adopt, to maintain, and to govern ourselves on the basis of such inner laws. But this is not

to say that the actions themselves that we actually perform are impermissible in the circumstances.

In the circumstances that we face, there will always be courses of action that we can take and that would be in line with the set of basic guiding principles that all agents could adopt and be governed by, and whose following would make it possible for all to preserve, and fully realize, their humanity. If we act in these ways *because* we are indeed guided by these basic maxims that have these properties, then we would in effect be exercising full autonomy, since we would actually govern ourselves in accordance with and on the basis of maxims that are fit to serve as universal laws. Our actions are in such cases actions that, as a matter of fact, actually *do* "co-exist" with autonomy (to use a phrase whose significance will be made clear in just a moment).

But *these same acts* – e.g. be saving a drowning child, giving correct change, or keeping some promise – could also have been performed for other reasons; it could be that we are not be governing ourselves on the basis of maxims that could be universal laws, but that other motives nevertheless lead us to act in these ways. Our actions are then actions that *can* co-exist with autonomy: autonomous versions of us could have performed these actions in these circumstances, on the basis of the basic maxims that are fit to serve as universal laws. And *this* – i.e. whether actions in particular circumstances would be *in line with* maxims that could serve as universal laws, and acting in these ways, therefore, could co-exist with autonomy – is precisely Kant's measure of whether specific acts, taken in themselves, are permissible or not. Thus he writes:

That action, which can co-coexist with autonomy of the will, is *permitted*; that

which doesn't conform to it is *impermissible*.[174]

So whether particular actions are permissible or not in given circumstances depends on whether these actions are in accordance with the set of basic guiding principles (or maxims) that all could adopt, be guided by, and thereby exist and flourish in their distinctive humanity. This means that the agents performing the given acts might do so for other reasons, that they, therefore, would be performing morally permissible actions, but that they wouldn't be doing this in a way that involves being guided by the given basic guiding principles or maxims themselves.

And this is precisely why Kant in all his major moral works draws a distinction between what he sometimes calls the *legality* and the *morality* of actions: whether the action an agent performs is merely in accordance with duty, on the one hand, (legality) and whether the agent performing the action is guided by a basic principle that could serve as a universal law, on the other hand (morality).[175] The *latter* of course depends on what the agent's actual maxim is (i.e. what guiding principle she is actually governed by in the circumstances).[176] But the former, *which is what corresponds to whether the action*

---

[174] (G: 4: 439) Actions are permissible, we could thus say, if agents governed by basic guiding principles that are fit to serve as universal laws (and who are thereby fully autonomous) could perform these actions in the circumstances. (But it is not enough, from the point of view of the universal law formula, to merely aim to act permissibly. This formula also instructs us to adopt the type inner guiding principles that would themselves lead us to act in these permissible ways, and thereby becoming fully autonomous. This further feature of Kant's view – that it is a virtue-ethical theory that doesn't only concern itself with what acts are performed, but which applies to the whole of the agent's conduct – is, I believe, what causes a lot of the confusion about what Kant's views are regarding the permissibility of individual acts.)

[175] (KpV: 5: 72) It is a confusing feature of Kant's overall corpus that in *The Metaphysics of Morals*, he at one point uses "legality" to refer only to conformity with juridical laws (i.e. publicly legislated laws), and at that point contrasts it with "morality", as in inner conformity with ethical laws (MS: 6:214), whereas he in places such as the *Critique of Practical Reason* applies this distinction more generally, as I did above, to the issue of whether actions are in conformity with the moral law or whether they are also performed out of respect for it.

[176] Kant also sometimes puts this distinction in terms of whether an action merely corresponds to the "letter" of the law, by being in accordance with duty, or whether it also conforms to the

is permissible or not, does not depend on this.[177]

So, then, the actual maxims feature associated with many of the standard readings is a mistake: it overlooks the distinction between (a) potential clashes between how the universal law formula instructs us to *conduct our decision-making* and how we actually conduct our decision-making, on the one hand, and (b) potential clashes between the specific actions we perform and the actions we could have performed in the same circumstances if we were acting on the basis of guiding principles that are fit to serve as universal laws, on the other hand (the latter being what determines permissible).[178] This means that any objections that depend on interpreting Kant in line with the actual maxim feature therefore fail.

Turn next to the *highly personalized feature*, i.e. the interpretative assumption that maxims that are candidate universal laws could be very specific policies or plans stating what to do in specific situations in light of specific goals if and insofar as we and those that we interact with have particular idiosyncratic personal features. This feature of most

---

"spirit" of the law, by being performed out of respect for the law. See, for instance, (KpV: 5: 72)

[177] Note that Kant thinks that it *is* a mistake to only concern oneself with the permissibility of the acts one performs. He thinks that we have a moral duty to develop and adopt a mindset whereby we act as we do because we are governed by maxims all could act in accordance with and in accordance with which our particular human nature can both be preserved and fully realized. Only then can we make ourselves fully autonomous, be fully in control of ourselves, and thereby leave the ethical state of nature behind.

So Kant would have a strong objection to the kind of moral philosophy, which is so common these days, which only concerns itself with the mere permissibility of possible courses of action. But in interpreting his theory it is very important (in order for us to properly understand the theory) not to take this to mean that Kant takes the mere permissibility of possible acts we could perform to depend on whether, in acting in these ways, we would be exercising full autonomy. He doesn't. As we've seen, the mere permissibility of our acts instead only depends on whether these acts could possibly co-exist with full autonomy, *not* on whether, in acting in these ways, we *are* actually exercising full autonomy. Many objections to Kant's theory – such as many of the objections Parfit discusses in his chapters on Kant in (Parfit 2011a) – arise only due to inattentiveness to this distinction within Kant's work.

[178] Another way of putting this point is to say that the universal law formula gives us one moral test for our action, and another for our maxims.

of the standard readings ignores Kant's statements explaining how he means "basic principles" containing "universal determinations of the will" that could be "valid for all reason-endowed beings" when he talks about the *maxims* that could qualify as universal laws of virtue. It also, secondly, disregards what kind of examples it is, as we saw above, that Kant himself puts forward as suggested maxims that could be universal laws, and instead seems to be wholly based on Kant's examples of maxims that, in his judgment, couldn't possibly be universal laws. So this feature of many of the standard readings, and the objections that rest on it, must also be rejected.

The *either logical or practical feature* of the standard readings is the idea that to qualify as a universal law, it is *enough* for it to be logically or practically conceivable that all be able to act on a given principle. This requirement is indeed, we have seen, a necessary condition on maxims for them to be fit to serve as universal laws: these must be maxims all could adopt and govern themselves by. But this is *not*, we have also seen, a sufficient condition.

The reason for that is that there can be principles everybody could adopt and follow, but that *wouldn't* allow us to preserve and fully realize our nature as autonomous human animals. So the either logical or practical reading of what it is for maxims to qualify as universal laws fails to offer a sufficient condition, which means that objections having to do with supposedly immoral maxims that everyone could, conceivably or practically, act on are, therefore, bound to fail.

Turn next to the *rational willing feature*. This is the idea that whether some maxim could be *willed* as a universal law depends on whether the agent who is thinking of adopting the maxim could herself rationally choose to bring about a situation in which

everyone follows this principle. And to determine this we must, according to proponents of this reading, use our own ideas of what rationality-based reasons people have for doing or accepting. At no point, however, does Kant claim that whether a maxim could qualify as a universal law is a matter of whether individual agents have reasons (self-interested or otherwise) to want people to follow such principles.

Kant's "test" is instead, as we have seen, that of whether we could preserve and realize our nature as self-determining (and yet sensually affected and happiness-seeking) beings under the given maxims as universal laws of action. So unless that is itself seen as a test of whether it is rational to choose the universal adoption of given maxims (and perhaps some will want to say that it is), the rational willing feature of some of the standard readings also turns out not to track what Kant had in mind in putting forward his formula.

And as for the apparent *interpretation in isolation feature* of many of the standard readings on which the objections against Kant's formula are based, we have already seen that it is not conducive to a good understanding of Kant's theory. It instead functions as a breeding ground for misunderstandings.

Now that we know that the standard readings both involve what clearly appears to be numerous mistakes and overlook important features of Kant's overall view, we are in a good position to evaluate the seven objections against the universal law formula reviewed above. As we are about to see, these objections fail.[179]

[179] That, of course, is not to say that there couldn't be, or aren't, other objections that can be raised against Kant's universal law formula. But our interest here is in whether the standard objections, and the further objections recently added by Parfit, refute Kant's theory or not, and what I wish to argue is that all *these* objections fail.

8. Why the Seven Objections All Misfire, Continued

We can start with *Herman's objection*, which is based on the premise that we can coherently imagine a Hobbesian state of nature with a war of all against all in which everyone follows a maxim that instructs all to kill others when this is to their own benefit, and they can get away with it. This objection relies on the either logical or practical feature of the standard readings. It thus ignores the part of Kant's theory according to which basic maxims that are fit to serve as universal laws are ones whose adoption and following would allow us to both *preserve* and fully realize our nature as self-determining human animals. The universal adoption and following of a rule that says to kill and harm others when this benefits us is not conducive to the preservation and full realization of the humanity within each person. It is instead in direct conflict with both of these aims. So Herman's objection fails.

Turn secondly to the *mixed maxims objection*. This is the objection according to which the universal law formula takes the moral status of particular actions to depend on the maxims on which the agent is acting; some maxims are such that it sometimes is obviously wrong to act on them, sometimes obviously right; an according to which Kant's principle, therefore, fails. This objection fails because its first premise is false: the key question with regard to whether candidate courses of action are permitted is not, as we have just seen above, what maxim the agent is actually acting on, but instead whether the actions in question would be in accordance with basic guiding principles that all could act in on and whose following would allow all to preserve and fully realize their own human nature.

Thus if some egoist saves a drowning child for self-interested reasons, his action

indeed lacks moral worth since the egoistic maxim of his will is not a guiding principle that is fit to serve as a universal law. But whether the act of saving the child is *permissible* doesn't depend on whether the maxim actually acted upon could be a universal law or not. It instead depends on whether a person who was acting on basic guiding principles that are fit to serve as universal laws could, on the basis of those principles, save this child in these circumstances. And a person who was acting on basic principles that are fit to serve as laws in accordance with which all can exist and flourish in their human nature *could*, on the basis of these guiding principles, save the drowning child. So this life-saving act can, of course, co-exist with autonomy of the will. And performing it is, therefore, permissible no matter what our motives might be. Hence Parfit's objection fails.

Turn next to *the rarity objection*, which can be summarized as the worry that the universal law formula seems to imply that we can make immoral actions permissible by adopting eccentric maxims tailored after our own idiosyncratic situations, wishes, and whims. None of these eccentric maxims, however, meet the basic guidance condition. A maxim that says to steal wallets of women in red dresses, who are eating strawberries, etc., could, for example, not offer practical guidance to all reason-endowed human agents across situations. Nor is it a *basic* principle of action in any recognizable sense. These eccentric maxims are, furthermore, not principles whose following would allow us to preserve and fully realize our distinctive human nature (as Kant thinks of it). And in addition to all this, we have also seen that whether an action is permissible does in any case not depend on the agent's maxim. So the rarity objection fails.

Turn next to *the threshold objection*: that there are principles it would be perfectly

okay to act on, just as long as they don't become universally followed and the number of people acting on these maxims remains under some particular threshold. This objection seems plausible only when we think of maxims that are candidate universal laws as specific principles tailored after our own particular interests. The universal law formula does *not* imply that it is contrary to duty to act on such personal principles for the sake of the promotion of our own idea of happiness. It instead implies that it is permissible to act on such personal plans so long as what we'd be doing in acting on these plans is in line with basic maxims that all could adopt and in accordance with which all could exist and flourish. So if somebody wants to make it a personal plan of theirs to move to Iceland to absorb the spirit of the Nordic Sagas, then that's permissible from the point of view of the universal law formula as long as the actions she'd be performing while doing this are permitted by maxims that meet the just-specified conditions for qualifying as universal laws of the reason-endowed nature.

That is, we are not required by the universal law formula to only ever form specific intentions or personal plans that could qualify as universal laws. What we are required to do is instead to subject ourselves to *basic guiding principles* that are fit to serve as universal laws for the preservation and realization of the humanity in each person. The threshold objection fails, then, on account of how it confuses the kinds of personal plans people might adopt for personal reasons with the type of basic guiding principles Kant has in mind when he talks about maxims that are fit to serve as universal laws.

Next up is *the high stakes objection*. According to it, the universal law formula fails since there are extreme situations in which agents on self-interested grounds could

rationally will, given their options in those particular situations, that everyone be permitted to kill or harm others greatly if this greatly benefits these agents themselves under these particular circumstances. But rules for unusual situations that permit us to kill others etc. for our own personal benefit in these extreme situations are not basic guiding principles that could guide us across situations. Nor is, I believe, people's being allowed to engage in self-interested killing in extreme situations conducive to the preservation and full realization of the human nature within each person. So this objection also misfires.

Turn next to *the ideal world objections*, which are all variations on the same theme. The idea is that there are maxims that we can rationally want everyone to act on because if everyone did, then this would have various good effects; but which are nevertheless maxims such that if everyone fails to act in accordance with them, then this would have very bad effects. These objections then assume that sound moral reasoning based on the universal law formula would entail that we ought to act on these maxims (which could be maxims such as "never use violence" or, to use Parfit's more dramatic example, "never use violence unless others use it, in which case kill as many as possible"). But this, of course, is absurd.

This general type of objection assumes, then, that in our choices of the basic guiding principles that are to guide us, we are *not* to take into account the kinds of limitations and shortcomings etc. that we ourselves, and people in general, are prone to, and which can undermine our capacity to exist as autonomous human animals. I believe, however, that this is an interpretative mistake. Given the examples Kant discusses, it is clear that he thinks that we need to take into account practical issues such as those of what kinds of beings we are, what our limitations are, what kinds of situations we tend to

face, how we tend to react to these kinds situations, etc.

Some of the maxims of virtue in relation to ourselves, for example, that Kant discusses and endorses are maxims instructing us to cultivate a capacity to tame and master our own impulses and spontaneous reactions so as to be able to assume and retain self-control.[180] Such maxims apply to us precisely because it is part of our humanity that we are subject to various inclinations etc. that, if allowed to take over, can develop into tendencies towards irresistible impulses or obsessions over which we have no control.

It is also a maxim of virtue, Kant claims, that we are to act in accordance with the *principles of right* within a just juridical state[181], which are coercive laws designed to protect the freedom of the individual, using coercive force when necessary. These external laws, and the maxim of inner virtue that requires us to submit to these laws, are premised on the assumption that we cannot count on everyone's always treating each other well, but that, some of the time, we need protection from each other. Thus there are, according to Kant, both (1) maxims fit to be universal laws that are premised on our needing protection from some of our fellows (e.g. the maxim to act rightly, i.e. to subject ourselves to the coercive public laws of the juridical state) so that we don't become subject to the arbitrary wills of others; and (2) maxims fit to be universal laws that are designed to protect us from ourselves (e.g. maxims of self-control and self-governance) so that we don't become slaves of our passions.

---

[180] These are the maxims that are sorted under the general "duty of apathy," which we will discuss in chapter 4. See (MS: 6: 407-8).

[181] "...that I make it my maxim to act rightly is a demand ethics makes on me" (MS: 6: 231) The highest-order principle of externally enforceable law-giving within a just juridical state, namely that "Any action is right if it can coexist with everyone's freedom in accordance with a universal law, or if on its maxim the freedom of choice of each can coexist with everyone's freedom in accordance with a universal law" (MS: 6:230) Associated with this principle there is, Kant writes, an "authorization" to use coercive force to "hinder hindrances of freedom". See (MS: 6:231).

It is, in other words, simply *not* a part of Kant's theory that we should choose maxims that, in Hill's phrase, are made "*for* ideal law-makers" or "for another world."[182] We should instead choose maxims in accordance with which *beings like us* (with our limitations, shortcomings, etc.) can fully realize our nature as happiness-seeking human animals that, at the same time, are also capable of fully autonomous agency. And reasoning involving the universal law formula, as Kant thinks of it, thus takes into account that we are not living in a morally ideal world.[183] Since the interpretative

[182] In Rousseau's phrase, Kant's suggested maxims of human virtue (virtue normally involving, as explained in chapter one, a *struggle* against countervailing influences) do instead really "take men as they are, and the laws as they might be."

[183] Here a critic might respond in the following way: "but what about Kant's insistence on our avoiding all lies? Doesn't that show that he thinks we should choose our maxims as if we were already living in an ideal world, and not mind the possible consequences?" This critic would, of course, have the essay "On the Supposed Right to Lie from Benevolent Motives" (ÜVR) in mind in which the example from Benjamin Constant about the murderer at the door occurs, and in which Kant argues that there is not a right to tell lies from benevolent motives. I shall save detailed discussion of that issue for some other occasion, because it is a complex topic that involves sorting out various common misconceptions about that essay in particular and Kant's actually rather complex views on truthfulness in general.

I will here limit myself to briefly noting the following about that infamous essay. (1): It is almost universally discussed outside of its context (the context being the aftermath of the French Revolution with its debates about the degree to which we ought to stay true to the ideals of *freedom* and *equality* instead of sometimes letting overall expedience trump these ideals, something that Kant was strictly against but that Benjamin Constant, to whose political pamphlet Kant was responding with his essay, was in favor of (Benton 1982); and that, upon closer inspection, Kant's main concern in the essay seems to be to argue that it must necessarily be, as he puts it in the essay, "a basic principle of politics" that "Right [roughly, public justice that can possibly be enforced through coercive laws] must never be adjusted to politics, but that politics must at all times be accommodated to Right." (ÜVR: 8:429) (2): The usual English translations of this essay translate "Aussagen" merely into "statement" (or, even worse, "utterance") rather than into "testimony," and "Die Lüge, bloß als vorsätzlich unwahre Deklaration gegen einen andern Mensch definiert…" (ÜVR: 8: 426) into (e.g.) "the definition of a lie as merely an intentional untruthful declaration to another person" rather than "the lie, defined simply as an intentional untruthful declaration *against* another human being…," (emphasis added). This means that the standard translations – or so I would argue – are (i) not good and (ii) fail to pay attention to the overall discussion in the essay, for which reason the discussion of this essay has suffered greatly. (3): This is an essay that, I believe, must be read – but that often does not seem to be read – in relation to the discussions of untruthfulness of different kinds in different kinds of contexts that is featured in works such as *The Metaphysics of Morals* and Kant's *Lectures on Ethics*. These are discussions in which Kant presents an overall complex view on which a "falsification" can sometimes justifiably serve as a "weapon of self-defense" (Kant 1997: 27:448); on which the

assumption on which these ideal world objections are based is therefore mistaken, these objections, I conclude, fail.

Return lastly to the *non-reversibility objection*: the objection based on the observation that people in certain privileged social groups could sometimes on self-interested grounds rationally choose that everyone follows principles that permit people in privileged social positions to keep oppressing those in less privileged social groups. This objection gets the test Kant thinks that maxims must pass in order to qualify as universal laws completely wrong. The test for whether a basic guiding principle could serve as a universal law is, as we have seen, *not* that of whether there are some people who have self-interested reasons for wanting others to follow this rule. Since this is the interpretative assumption on which this objection rests, and it is a mistaken assumption, this objection also fails. I conclude that the seven seemingly decisive objections against Kant's universal law formula all fail.

"only kind of untruth we want to call a lie, in the sense bearing upon rights, is one that directly infringes upon another's right, e.g., the false allegation that a contract has been concluded with somebody, made in order to deprive him of what is his" (MS: 6:239); on which one of the greatest problems with a lie (in the sense of failing to be truthful) is that conflicts with our duties to ourselves as "moral" beings, (MS: 6:420, 430) etc. etc. This all shows – as (Wood 2008) also similarly notes – that Kant's views on the requirements of truthfulness are much more complex than it is normally assumed.

In light of these considerations I want to say that the common assumption that Kant has the simple view that we are under a strict duty to always tell the truth no matter what is an oversimplification, which to a large extent is based on overly quick readings of bad translations of the infamous essay. And so proponents of the ideal world objection ought to use other arguments to defend the crucial (but I believe mistaken) interpretative assumption on which their argument depends. But *even* if we were to suppose that Kant does have a much more stringent demand for truthfulness within his substantive theory than most people today would accept (as I would certainly agree that he does), this still doesn't show that he thinks that sound moral reasoning using the universal law formula does not take into account our own and other people's limitations, shortcomings, etc., because as we saw in the running text above, Kant clearly thinks that these are precisely the sort of things we should take into account when we choose our maxims. So whatever the exact degree of strictness of Kant's favored demand for truthfulness might be, the crucial interpretive premise at the center of the ideal world objections nevertheless turns out to be false.

9. Looking Ahead

On my suggested interpretation, which I have argued for on the basis of the eight interpretative steps I have suggested, maxims that are fit to serve as universal laws of the reason-endowed nature are basic guiding principles that could guide us across situations and whose following would allow for and result in the preservation and full realization of the humanity in us: i.e. the capacity we have to be autonomously governed by our own practical reason while at the same time being subject to various needs, desires, and a resulting general wish for happiness (i.e. the state where everything goes in accordance with our wishes and desires). Once we understand Kant's universal law formula in this way, we quickly see that the various standard interpretations against it reviewed above don't work. Moreover, since choosing one's maxims on the basis of whether they meet these constraints amounts to making the preservation and full realization of the humanity within each person into the most general purpose around which all our actions are based, a purpose that we would then never act contrary to, this basic moral principle can indeed also, we have found, be formulated along the lines of the humanity formula: the requirement to never through one's actions treat the humanity in any person as a means only, but instead to always at the same time also treat it as a purpose in itself.

Many influential contemporary commentators, however, claim that the humanity formula differs in content from the universal law formula because, they argue, the former asserts a substantive value, which the latter does not assert: namely, the absolute value of all human beings. Since this reading is very popular among commentators who teach at American universities (and perhaps mostly among commentators who teach in the US), I

shall call those who offer such readings members of the *American School* of Kant interpretation. In the next chapter we will first consider the American School's various variations of the same theme, namely their way of understanding Kant's reasoning leading up to and surrounding his initial statement of the humanity formula in the *Groundwork*, which they believe to importantly involve an assertion to the effect that the humanity in each person possesses absolute value. As we shall see, however, what Kant believes to have absolute value is a good will, and not humanity.

Humanity, Kant claims, has absolute value only in a derivative or conditional way, only insofar as it is related to the capacity for morality (i.e. the possession of a good will). So, the American school's reconstructions fail, and we need an alternative way of understanding the reasoning in the *Groundwork* that leads up to the statement of the humanity formula. And we also need an alternative account of the relation between what has absolute value (a good will) and what is an end in itself (the humanity within each person).

CHAPTER III

Kant's Real Argument for the Humanity Formula

1. Introduction

Discussing Korsgaard's reconstruction of Kant's argument for the humanity formula, Jens Timmermann makes the following remarks:

> Korsgaard's work stands out among recent Kantian scholarship. Her reconstruction is now widely considered the standard reading of Kant's value theory in Anglo-American moral philosophy.[184]

There is no room for disagreement here. Korsgaard's reconstruction of the argument in the *Groundwork* for the principle of humanity is at the center of all serious recent attempts to understand Kant's argument. There is, however, plenty of room for disagreement as regards whether Korsgaard's reconstruction of Kant's argument is correct, even among those who are in partial agreement with Korsgaard. Allen Wood, for example, writes:

> I think Christine Korsgaard is on the right track in suggesting that Kant's reasoning here takes the form of a "regress on conditions." It begins from the value we place on the ends we set, and infers that this value is grounded in the rational nature of the being who sets the end, which (in Korsgaard's words) possesses a "value-conferring status" in relation to that end. Because humanity or rational nature is the source of all such value, it is regarded as absolutely and unconditionally valuable.[185]

[184] (Timmermann 2006: 69)
[185] (Wood, 1999: 127)

But while Wood thinks that Korsgaard is "on the right track" in her analysis of what Kant's general strategy is – and Wood therefore is another member of what I am calling the *American School* – Wood disagrees strongly with Korsgaard's understanding of Kant's theory of value. Commenting on the "value-conference" part of Korsgaard's reconstruction, Wood writes

> The idea that any objective value could be simply *conferred* by human choice is nonsense – it contradicts the very concept of objective value… Still less should we say, as Korsgaard also has, that rational beings confer on themselves the value of being ends in themselves… Rather, the argument is that it is our basic act as rational beings, the act of setting ends and regarding them as good, that necessitates our representing ourselves as *already* ends in themselves.[186]

Wood, then, strongly disagrees with the particular theory of *value* that Korsgaard attributes to Kant. But he agrees with Korsgaard that Kant's argument for the principle of humanity is a "regress argument" that starts with the idea of the supposed objective goodness of our ends and then regresses back to something that supposedly must be assumed as valuable in itself, which also serves as the "sufficient condition" of the value of everything else. And that thing, Wood's Kant holds, is humanity.

Timmermann also disagrees with Korsgaard. According to him, what's wrong with Korsgaard's reconstruction is, to simplify, that it takes Kant to *regress* on conditions of value. What Kant's argument does is, rather, simply surveying possible candidates for absolute value and concluding that the only possible candidate for an independent value upon which all principles of morality could be founded is the value of humanity. Kant's argument, Timmermann claims, "is ontological throughout".[187]

---

[186] (Wood 2008: 95, 92)
[187] (Timmermann 2006: 75)

My disagreement with Korsgaard – and other members of the American school inspired by her work – runs much deeper. The basic mistake of the line of interpretation that Korsgaard has inspired is, I believe, that it wrongly conflates Kant's claims about the absolute value of a good will with his claim, and argument intended to show, that humanity exists as an end in itself. When Kant claims that humanity exists – and that it is in accordance with principles of practical reason to be treated – as a purpose in itself, he is making a different claim, which he supports with a different argument, than his claim that the good will has absolute value. In failing to note this key distinction, members of the American school have, I shall argue, gone fundamentally wrong in their attempts to reconstruct Kant's argument leading up to his statement of the humanity formula.[188]

As an illustration of what I mean, let's consider this recent Korsgaard-inspired attempt to summarize Kant's reasoning by Samuel Kerstein:

> First, Kant contends that if there is a supreme principle of morality (and thus a categorical imperative), then there is an objective end, something that is unconditionally good. Second, Kant claims that this unconditionally good thing would have to be humanity. In his view, therefore, if there is a supreme principle of morality, then humanity is unconditionally good. But, third, if humanity is unconditionally good, then we must always treat it not merely as a means but also as an end. Therefore, if there is a supreme principle of morality, then we ought to do just what the Formula of Humanity says. So the supreme principle of morality, if there is one, must be this formula, or at least something equivalent to it.[189]

There are at least, I believe, three major problems with this summary of Kant's reasoning. First, it is *morality* to which Kant attributes absolute value, whereas humanity only is

[188] Nor do the American school's interpretative claims fit with Kant's claims about humanity as an end in itself in the *Metaphysics of Morals* and the *Critique of Practical Reason* or his claim in the *Groundwork* that he will give the grounds for why reason requires us to regard humanity as a purpose in itself in the third part of the *Groundwork*, which is about how the universal law formula is the constitutive principle of a positively free, and therefore autonomous, will.
[189] (Kerstein 2002: 47)

something to which attributes absolute value in a derivative or conditional way, namely insofar as it is capable of morality.[190] Secondly, if this is how we understand the humanity formula, and Kant's argument for it, then we seem to paint ourselves into a corner from which we cannot explain why it is that Kant thinks that the universal law and humanity formulas "at bottom are really only different statements of the same law."[191] Thirdly, Kant's attribution of absolute value to morality (or virtuousness) is done on separate grounds than those on whose basis he, as we will see below, argues that humanity is an objective end and a purpose in itself.

The rest of this chapter is divided into seven further sections. Section two introduces the relevant bits of text in the *Groundwork* that our discussion will focus on. Sections three and four reviews Korsgaard's influential reconstruction of Kant's argument, Wood's reaction to it along with his alternative (and yet similar) suggestion, and also some other similar views held by other members of the American school. Section five argues that there is reason to reject all reconstructions on the pattern of those reviewed in section three. Section six briefly discusses Kant's concerns about the relative value vs. the absolute value of a human being, and relates these concerns to his ideas about worthiness of happiness. Section seven then offers my alternative reconstruction of Kant's reasoning leading up to the initial statement of the humanity formula, which tries to incorporate the reasoning that Kant himself puts off until the third section of the *Groundwork* into the argument. Section eight discusses the question of whether, in abandoning the recent American school of Kant interpretation and instead understanding Kant's argument as having to do with the constitutive principles of a positively free will

[190] "Morality, and humanity insofar as it is capable of it, is that alone which has dignity"[190] or an "inner worth"[190] (G: 4: 435)
[191] (G: 4: 436)

(a concept that seems metaphysically extravagant to writers such as Thomas Hill), we in effect are being uncharitable to Kant. In response to this I will argue that not only does Kant's ideas about autonomy of the will constitute the heart of Kant's ethics (for which reason moving away from these parts of Kant's theory would amount to abandoning Kantian ethics), but that many worries about Kant's focus on the freedom of the will are likely to be based on misconceptions about what Kant means by "freedom" when he talks about the freedom of the will.

2. Kant's Argument for the Humanity Formula in His Own Words

Our topic in this chapter, then, is Kant's reasoning leading up to and following the initial statement of the humanity formula in the *Groundwork*. Before getting to Korsgaard, Wood, and the other members of the American School, we will begin by locating ourselves in the text and looking at some of the main claims Kant makes along the way as he works his way towards the humanity formula.

At the point in the *Groundwork* where we will drop down, Kant has already argued his way to the universal law formula, which he believes we can argue for on the basis of two main things: (1) the very idea or concept of a *categorical imperative* of the will; and (2) the idea of the *will* as a capacity to determine oneself to action on the basis of principles: or, as Kant also puts it, reason in its practical capacity[192]. Very roughly, Kant argues that since all material principles of the will, i.e. principles presupposing some given purpose, can only give rise to hypothetical imperatives (as in "do such-and-such for the sake of the purpose P"), and a categorical imperative, if there is one, would

[192] (G: 4:412)

be an unconditional imperative, there is nothing left for a categorical imperative to be other than a requirement that the will's principles be universal and law-like.[193] Hence the most basic categorical imperative – again, if there is one – is to choose one's basic guiding principles (or "maxims") on the basis of their fitness to serve as universal laws. And this, of course, is *the universal law formula*.

But since the existence of things in accordance with universal laws is what we mean by "nature" in its *formal* sense, Kant continues (and presumably he is here taking the "things" in question to be the reason-endowed beings who'd be operating in accordance with these self-legislated universal laws), we can also state the basic categorical imperative as an imperative to "so act as if the maxims we're acting on through our wills were to become universal laws of nature".[194] And this is *the law of nature formula*.

After illustrating what he has in mind using the four famous examples, Kant then claims that because the idea of universal and unconditional laws of practical reason would assume a kind of *necessity* that we couldn't derive from any empirical facts about any contingent features of our human nature, these must simply be laws for reason-endowed beings in general in virtue of something about the concept of such a being itself (laws which, therefore, apply to human beings only because we are a type of reason-endowed beings[195]). As Kant writes:

---

[193] (G: 4:421-2)
[194] (G: 4:422)
[195] (G: 4:425-6) This is why I have said that the universal law formula must be *applied* to human beings, and thereby, i.e. in this application, becomes the human nature formula, which, strictly speaking, is what is equivalent to the humanity formula. (The human nature formula: choose your maxims of action on the basis of their fitness to serve as basic guiding principles for all human beings in accordance with which the human nature in all can be preserved and fully realized. The

> The question, therefore, is this: is it a necessary law for all reason-endowed beings to judge their own actions in accordance with such maxims that they themselves could will that they should serve as universal laws? If such it is, then this must already (completely a priori) have to do with the very concept of the will of a reason-endowed being.[196]

And what Kant does next, therefore, is precisely to more carefully examine the concept of a *will* and the way that the will relates to action to see if we might be able to argue our way to a categorical imperative starting with an analysis of the concept of the will (or the practical reason) of a reason-endowed being. In a very dense paragraph, Kant presents the following analyses and distinctions:

> The will is thought of as a capacity to determine oneself to action *in accordance with the representation of certain laws*. And such a capacity can only be found in reason-endowed beings. Now that which serves the will as the objective ground for its self-determination is the *purpose*, and if this is given by reason alone, then it must hold equally for all reason-endowed beings. What, in contrast, merely contains the basis of the possibility of action, whose result is a purpose, is called a *means*. The subjective basis of desires is the *enticement*, the objective [basis] of the will [is] the *reason for action*; hence the difference between subjective purposes, which derive from enticements, and objective [purposes], which derive from reasons for action, which hold for all reason-endowed beings. Practical principles are *formal* when they abstract from all subjective purposes; they are, however, *material* when they put these, [and] hence certain enticements, at their basis.[197]

Kant then immediately asserts that:

> The ends that a reason-endowed being sets itself as *effects* of its actions in accordance with what appeals to it (material purposes), are all only relative; because it is only the relation to the particularly constituted faculty of desire of the subject that gives them their value, [a relation] which therefore cannot provide any necessary universal principles, i.e. practical laws, that are valid for the wills

---

humanity formula: so act that you always at the same time treat the (preservation and full realization of the) humanity in each person as a purpose, and never as a means only.)

[196] (G: 4:426)

[197] (G: 4: 427-8)

> of all reason-endowed beings. Therefore are these relative ends only bases for hypothetical imperatives.
>
> Supposing, however, that there were something whose existence has absolute value, which, as *a purpose in itself*, could serve as the basis of certain laws, then in it, and in it alone, the basis of a categorical imperative, i.e. a practical law, would lie.[198]

That Kant brings up the issue of *value* here comes, in a way, as a surprise. Because while the first section of the *Groundwork* – which is about what Kant calls (what can roughly be translated into) "commonsense morality" – is about the *absolute value* that he thinks that we all place on a good will (and the good will alone), the section of the *Groundwork* that we are currently discussing is presented as having been a "transition from commonsensical thinking about morality to moral philosophy". And the discussion has now moved away from commonsensical value judgments and instead moved on to the question of what norms and principles it is in accordance with which a good will operates.

What we are about to find, however, is that Kant is at this point slowly but surely starting to lead his discussion back towards its starting point.[199] I will save my own

---

[198] (G: 4:428)

[199] The starting point, again, having been what he regards as the commonsense judgment that the only thing we can imagine that has an absolute value and that is unconditionally good is a good will. The section second of the *Groundwork* tries to vindicate that judgment by arguing, among other things, that a good will is: (1) a will that "couldn't be evil" since it is a will "whose maxim could not possibly contradict itself" (G: 4: 437); and (2) a will that possesses a special *dignity* (as in an exalted rank or elevated status) since it is ultimately only subject to laws that it gives to itself (G: 4:434, 440).

As Kant himself puts it with reference to this latter point: "From what has just been said it is now easy to explain how it happens that, although in thinking of the concept of duty we think of subjection to the law, nevertheless at the same time we thereby represent a certain sublimity and *dignity* in the person who fulfills all of his duties. Because there is indeed no sublimity in him insofar as he is the *subject* to the moral law, but there certainly is insofar as he is at the same time *law-giving* with respect to it and only for that reason subordinated to it. We have also shown how it neither is fear nor inclination but simply respect for the law that is the incentive that can give actions a moral worth. Our own will insofar as it would act only under the condition of a possible giving of universal law through its own maxims – this will [that is] possible for us in idea – is the

reading of why exactly Kant brings up concerns about the idea of absolute value here till later, but I can already briefly mention that I believe that this has to do with the fact that one of the central ways in which Kant thinks of genuine moral motivation is as a concern to be worthy of happiness, which means, he thinks, that there must be something that could give us an inner worthiness that doesn't depend on what external desires and needs we might satisfy by having the non-moral properties that we do.[200] And we find such a thing when we eventually come to explain why and in what sense it is that a good will is unconditionally good.[201]

At any rate, having momentarily returned to the topic of what has value and having also reasserted that all our objects of desire as well as all that belongs to non-rational nature can only have conditional value, Kant then claims that

> …in contrast [to all these things] reason-endowed beings are called *persons*, because their [very] nature already makes them out to be purposes in themselves, i.e. thing[s] that may not be treated as means only, [a nature that] hence constrains all choices, (and is an object of respect). These are thus not merely subjective purposes whose existence has value *for us* as effects of our actions; but [rather] *objective purposes*, i.e. things whose existence is a purpose in itself, in whose place no other purposes can be put for which they are to serve as mere means, because without them nothing of *absolute value* could be found…[202]

---

proper object of respect; and the dignity of humanity consists just in this capacity to give universal law, though with the condition of also being itself subject to this very same law-giving." (G: 4: 440) (For a very useful discussion on Kant's use of the term "dignity" throughout his works, which is of great help when interpreting passages such as the just-quoted one, see (Sensen 2009).)

[200] In both the *Groundwork* and *The Metaphysics of Morals* Kant distinguishes the relative value we and our talents have on the market and in relation to our own and people's affections, on the one hand, from the absolute (or inner) value we have on account of our inner character, on the other hand. In *The Metaphysics of Morals* Kant adds that we have a duty to ourselves to judge ourselves by our inner character, or the quality of our will, and never on the basis of the degree to which we please others or meet their needs. (G: 4: 434-5); (MS: 6: 434-5, 441)

[201] (G: 4: 437, 439-40)

[202] (G: 4: 428)

What we have at this point, then, is not only the claim that *if* it is a necessary law for all reason-endowed beings to choose their maxims on the basis of their fitness to serve as universal laws, then this has to do with very concept of the will of a reason-endowed being. We also now have an assertion to the effect that the *nature* of a reason-endowed marks the existence of the reason-endowed being out as a purpose in itself. And since wills are themselves only found within reason-endowed beings, and the good will is the only thing that has absolute value, it is of course the case, if Kant is right that only the good will can have unconditional value, that without reason-endowed beings, nothing of absolute value would be anywhere to be found.

Kant continues as follows:

> If there thus is to be a highest practical law, and, in relation to the human will, a categorical imperative, then it has to be such [a law] that it makes an *objective* principle of the will out of the representation of what is an end for everyone since it is a *purpose in itself*, [which] hence can serve as a universal law. The basis of this principle is: *the reason-endowed nature exists as a purpose in itself*. The human being necessarily conceives of its own existence in that way: thus far is it, then, a *subjective* principle of human action. But [it is] also [the case that] each reason-endowed being conceives of its own existence in this way, on the basis of the same reason that also holds for me: thus it is also an *objective* principle, out of which, as a highest practical basis, all laws of the will must be derivable. The practical imperative thus becomes the following: *so act that always treat the humanity in your own person, as well as that in each other person, always at the same time as an end, and never as a means only*.[203]

Now as regards the claim that each reason-endowed being, on the basis of the same reason, must conceive of its own existence as a purpose in itself because it has the nature that it does[204] – why are we to accept this claim (whatever it might mean exactly)?

---

[203] (G: 4:428-9)

[204] Recall here that just a few paragraphs before the just-quoted passage, Kant has said that "reason-endowed beings are called *persons*, because their [very] nature already makes them out to be purposes in themselves, i.e. thing[s] that may not be treated as means only" (G: 4: 428)

Realizing that the reader must be asking him or herself just this question, Kant attaches a crucial footnote to this claim, which reads as follows:

> I here put forward this proposition [i.e. that each reason-endowed being possessing a will, on the same basis as all other reason-endowed being, must conceive of its own existence as a purpose in itself] as a postulate. The grounds for it are to be found in the last section.[205]

That last section of the *Groundwork*, in turn, is a continuation of Kant's analysis of the concept of the *will* of a reason-endowed being. And the aspect of this concept that Kant there turns his attention to is the *freedom* we all attribute to our wills when, as decision-making agents, we think of ourselves as responsible agents who are not determined or ruled by their desires and impulses.

We can only secure this independence from the rule of our inclinations and impulses, Kant goes on to argue, if we make our wills positively free by ourselves laying down the laws in accordance with which our wills are to operate. And this, Kant argues, is something we can only do by choosing our basic guiding principles on the basis of their fitness to serve as laws for all reason-endowed beings. Hence we end up with the claim that the moral law is nothing less than the constitutive principle of the will of an autonomous being. But what exactly just happened here?

We started with the universal law formula and eventually ended up returning to the universal law formula. But what exactly went on in the middle part when the argument had to do with how the will always uses purposes as "the ground[s] of its self-determination" and when it also had to do with the different types of purposes that there are? It is with regard to this middle part of the argument, i.e. with the reasoning there that

[205] (G: 4: 429)

leads up to the statement of the humanity formula, that Korsgaard and, following her, the American school of Kant interpretation make a distinct suggestion about what exactly Kant's argument is. Let us now review their suggestions, with a special emphasis on Korsgaard's reading.

3. The Reconstructions of Korsgaard, Wood, and other Members of the American School: Preliminaries

The *first* and perhaps most distinctive feature of the American school as a whole is that it takes what I above just called the "middle part" of Kant's reasoning to constitute a *separate* argument, which itself is not a part of the overall argument of the *Groundwork* that only gets completed in the third part of the book.[206] That is to say, the reconstruction of Kant's reasoning leading up to the statement of the humanity formula (and its relation to what comes after the statement of this formula in the *Groundwork*) that Korsgaard and those following her present does not at all make use of the fact that what Kant does in the third part of the *Groundwork* is to argue that the universal law formula is the constitutive principle of an autonomous will.[207]

As for how to understand Kant's claim that the reasons why each reason-endowed being must regard its own existence as a purpose in itself will be given in the third part of

[206] An observation about writers such as Korsgaard and Wood, I could add here, also made in (Darwall 2006). (Darwall himself does not, however, endorse that interpretative idea, but makes the same assessment that I also make, namely that Kant's reasoning surrounding the initial statement of the humanity formula is just one part in the overall argument of the *Groundwork* as a whole.)

[207] This, of course, is intimately related to another view shared by most members of the broad American school of Kant interpretation: namely, their shared view that the universal law and humanity formulas differ in their content, whereby the former is something not even considered as having any substantive implications and the latter is thought to supply those. See section 4 of chapter 2 above.

the *Groundwork*, Korsgaard understands this as having to do with how Kant at the very end of the *Groundwork* in his "concluding remark" says that it is typical of human reason that it responds to its conception of something (whatever it might be) as *conditioned* by "seeking the unconditioned".[208] This idea – i.e. that the just-mentioned search for the unconditioned is what Korsgaard thinks Kant hints at in his crucial footnote – is something about the American school's reconstruction we will return to in just moment, but the first distinctive thing to emphasize is that this reconstruction doesn't think of the preconditions for the realization of the autonomy of the will as having anything to do with the reasoning that leads Kant to the initial statement of the humanity formula. And this being so is, at least in part, a consequence of the *second* distinctive feature of the American school's reconstruction, which is the following.

Korsgaard, Wood, and company understand Kant as *not* drawing any distinction between *having something as a purpose*, on the one hand, and *valuing* or *attributing goodness to something*, on the other. Thus just as one example of this feature of the American interpretation, we can note that in commenting on an example of Kant's – which is about people with different motivations for promoting the welfare of others (doing it out of a sense of duty, doing it out of enjoyment in the welfare of others, etc.) – Korsgaard writes that "[e]ach of these characters genuinely has the welfare of others as his end – *that is, each values it for its own sake*."[209] Or, for an even blunter statement of this interpretative idea, consider Richard Dean's declaration that on his reading,

[208] (G: 4:463); (Korsgaard 1986a: 192)
[209] (Korsgaard 1986a: 184)

"anything that is an end in itself is also good without qualification, and anything that is good without qualification also is an end in itself."[210]

In the *Groundwork* when Kant himself defines what a purpose is, he says (as we saw above) that it is something that serves as an objective ground of the will's self-determination. In the *Critique of the Power of Judgment*, Kant defines a purpose as an effect through causation via concepts.[211] And in the *Metaphysics of Morals*, he writes that an end is the "object" of a choice (from which he concludes that all our actions must "contain" ends, since we choose our actions)[212]. Kant, then, doesn't himself explicitly define an end as something that we value for its own sake in any of his own main analyses of what purposes are. But it is a distinctive feature of the American school that it nevertheless understands Kant as viewing having something as our end as equivalent to placing a value on it.

Indeed, that the members of the American school of interpretation understand Kant as not drawing a distinction between something's being a purpose for which we're acting (or ought to be acting) and something's being a value directly explains their unwillingness to understand Kant's argument that the universal law formula is the constitutive principle of autonomous agency as being what explains why each reason-endowed agent necessarily thinks of its own existence as a purpose in itself. Why? *Because none of the main arguments for this claim in the third part of the Groundwork are about value*. And so the only thing that seems left in the third part of the *Groundwork* that could possibly be relevant to the claim that the reason-endowed being thinks of its

---

[210] (Dean 2011: 307)
[211] "A *purpose* is the object of a concept, in so far as the concept as seen as the cause of the object" (KU: 5:220)
[212] (MS: 6:381)

own existence as a purpose in itself is Kant's small set of remarks towards the very end of the book in his "concluding remark" about how reason inevitably seeks the unconditioned. Thus Korsgaard writes that, on her reading,

> the argument for the Formula of Humanity depends upon the application of the unconditioned/conditioned distinction to the concept of goodness. This follows from the fact that good is a rational concept. In any case where anything is conditional in any way, reason seeks out its conditions, not resting until the "unconditioned condition" is discovered (if possible).[213]

Now since Korsgaard (and others along with her) do not distinguish between something's having an absolute value and something's existing as a purpose in itself, it also follows, fourthly, that *what* is understood as having absolute value in Korsgaard's reconstruction is *the same thing* as that which is understood as existing as a purpose in itself. And the thing that Korsgaard understands "humanity" to refer to is our capacity to set our own ends of action, which is thus also that which she understands Kant as attributing an absolute and unconditional value to.[214] So the success of Korsgaard's reconstruction is going to hang, among other things, on whether the thing to which Kant attributes absolute value really is the same thing he claims to exist as a purpose in itself.

4. The Reconstructions of Korsgaard, Wood, and other Members of the American School, Continued

With these various preliminaries in place, we can now return to Kant's text and consider the American School's reading of it, starting with Korsgaard. Recall first that

[213] (Korsgaard 1986a: 192)

[214] This, of course, clashes with the understanding of Kant's use of "humanity" that we are using in this dissertation, which is that Kant uses this term to refer to our distinctive dual nature as animals with various needs, desires, etc. who nevertheless at the same time are capable of autonomous agency.

Kant asserts that the ends we set ourselves on the basis of finding things enticing can only be said to have relative value and additionally that the only principles of reason such ends can be associated with are hypothetical imperatives that tell us what we must do in order to attain these goals. Having made those claims, the very next thing Kant writes is:

> Supposing, however, that there were something whose existence itself has an absolute value, which, as *a purpose in itself*, could be a basis of certain laws, then in it, and it alone, the basis of a possible categorical imperative, i.e. a practical law, would lie.[215]

On Korsgaard's understanding of this crucial sentence, Kant is not here relating two different properties to the thing he has in mind (those of *possessing absolute value* and *existing as a purpose in itself*), but is instead simply asking us to imagine there to be something that possesses absolute value (which, on the American reading, is the same as existing as a purpose in itself). On this reading, the main task of the rest of the argument therefore becomes to argue that there is something that possesses absolute value. And the way Kant does this, Korsgaard suggests, is by starting with what possesses merely conditional value and then "regressing" back on conditions towards something that serves as the basic condition upon which everything else's having value depends.[216] The idea is

---

[215] (G: 4: 428) What Korsgaard and company fail to see, I believe, is that what Kant is after here is not just one thing, but *two* things: first, something that can serve as the end or purpose contained in all maxims fit to serve as universal laws, which is necessary since all actions are organized around purposes; and, second, something on whose basis the human being could acquire a kind of absolute value (a value not determined by the needs and desires that she might answer to), which could make her worthy of happiness. The latter turns out being virtue or morality in the sense of self-government through a good will, whereas the former, in contrast, turns out being the preservation and full realization of the humanity within us, which is something that contains, but is *not* exhausted by self-governance through a good will (a will, as Kant puts it, "that couldn't possibly contradict itself"). As I will argue below, having a good will can be regarded as the *formal* aim of somebody who is virtuously trying to be worthy of happiness, whereas having the humanity in each person as her most general purpose can be considered as the *material* or substantive aim of somebody who is striving to be worthy of happiness.

[216] (Korsgaard 1986a: 122-3)

that whatever serves as a "sufficient condition" for there being any value in the world must itself, and for this very reason, have unconditional value.[217] That thing would, as Korsgaard understands the argument, exist as an end in itself. (Again, this would follow since Korsgaard understands having an absolute value as being the same as existing as a purpose in itself.)

In light of these interpretative assumptions that Korsgaard suggests, let us next consider the following passage, which is part of the paragraph immediately following the above-quoted sentence:

> All the objects of our desires only have a conditional value; because if it were not for these desires, and the needs based on them, these objects would be without any value. The desires themselves, however, as sources of needs [thereby] have so little absolute value that being rid of them must be a universal wish among all reason-endowed beings. Thus the value of all the objects/states of affairs we're trying to *bring about* through our actions is at all times conditional. The beings whose existence doesn't depend on our wills, but that instead are parts of nature, have but a relative value as means if they are beings without reason, and are therefore called *things*; [while] in contrast reason-endowed beings are called *persons* because their nature already makes them out to be purposes in themselves, i.e. things that may not be treated as means only … [and] without them, nothing of *absolute value* would be anywhere to be found...[218]

What Korsgaard takes Kant to be doing here is regressing upon conditions in the sense laid out above: that is to say, she understands Kant as starting with what he thinks has merely conditional value (the ends we set on the basis of our desires and needs, these

---

[217] As Korsgaard puts it elsewhere: "An end provides the justification of the means; the means are good if the end is good. If the end is only conditionally good, it in turn must be justified. Justification, like explanation, seems to give rise to an indefinite regress; for any reason offered, we can always ask why. If complete justification of an end is to be possible, something must bring this regress to a stop; there must be something about which it is impossible or unnecessary to ask why. This will be something unconditionally good. Since what is unconditionally good will serve as the condition of the value of other good things, it will be the source of value." (Korsgaard 1986b: 227)

[218] (G: 4:428)

desires and needs themselves, any and all parts of nature that lack the ability to reason, etc.) and then trying to find a sufficient condition for why we would attribute any value to these things at all. And the only reason Korsgaard reads Kant as seeing why we would attribute any value to these others things is that they are *either* things that appeal to us that we therefore have chosen to make into ends of our actions *or* things that can serve as means in the pursuit of the ends we set for ourselves.

But this also explains, Korsgaard thinks, why it is that Kant claims that as reason-endowed beings, we all necessarily see ourselves as having absolute value: it is because (1) we in effect treat our having chosen to pursue some end as a sufficient condition for this desired end to be of value or for any means to the achievement of this end to be valuable, and (2) anything that is a sufficient condition for something else to acquire value must itself possess unconditional value. Thus Korsgaard explains her reading of this part of the argument in the following way:

> When Kant says: "rational nature exists as an end in itself. Man necessarily thinks of his own existence in this way; thus far it is a subjective principle of human actions," … I read him as claiming that in our private rational choices and in general in our actions we view ourselves as having a value-conferring status in virtue of our rational nature. We act as if our own choice were the sufficient condition of the goodness of its object: this attitude is built into (a subjective principle of) rational action.[219]

This then leads directly to Korsgaard's reading of the next major step in the argument. She writes:

> When Kant goes on to say: "Also every other rational being thinks of his existence by means of the same rational ground which holds also for myself; thus, it is at the same time an objective principle from which, as a supreme practical

[219] (Korsgaard 1986a: 123)

> ground, it must be possible to derive all laws of the will," … I read him as making the following argument. If you view yourself as having a value-conferring status in virtue of your power of rational choice, you must view anyone who has the power of rational choice as having, in virtue of that power, a value-conferring status.[220]

And summarizing her reconstruction, Korsgaard explains the transition from the supposed assertion of an absolute value to a categorical imperative of practical reason in the following way:

> …regressing upon the conditions, we find that the unconditioned condition of the goodness of anything is rational nature, or the power of rational choice. To play this role, however, rational nature must itself be something of unconditional value – an end in itself. This means, however, that you must treat rational nature wherever you find it (in your own person or in that of another) as an end.[221]

In the simplest possible terms: attributing importance to things because they are important *to us* in effect amounts (Korsgaard understands Kant as arguing) to attributing a basic importance to ourselves. And if we really have this type of importance whereby our choosing things is enough to make these objects of choice important, then we ourselves (or, rather, the rational capacities involved in our choices) must be things that are always to be treated as ends in themselves, and never as mere means to other ends.[222]

---

[220] (Korsgaard 1986a: 123)

[221] (Korsgaard 1986a: 123) What does this mean, according to Korsgaard, with regard to how we should act? It means, she writes, that, "no choice is rational which violates the status of rational nature as an end… It is an unconditional end, so you can never act against it without contradiction. If you overturn the *source* of the goodness of your end, neither your end nor the action which aims at it can possibly be good, and your action will not be fully rational." (Korsgaard 1986a: 123)

[222] Here's how Korsgaard herself roughly summarizes her own reconstruction in *The Sources of Normativity*: her own argument in that book is, Korsgaard writes, "just a fancy new model of an argument that first appeared in a much simpler form, Kant's argument for his Formula of Humanity. … He started from the fact that when we make a choice we must regard its object as good. His point is the one I have been making – that being human we must endorse our impulses before we act on them. He asked what it is that makes these objects good, and, rejecting one form of realism, he decided that goodness was not in the objects themselves. Were it not for our desires

That, in a nutshell, is how Korsgaard thinks we should understand Kant's argument. Allen Wood, who we'll turn to next and whose reconstruction can be more quickly summarized, thinks that Korsgaard is indeed on the right track (in taking Kant's argument to be about the conditions under which things acquire their value), but that her idea of our choices as "conferring value" on our objects of choice cannot be made to cohere with the premise that we choose things we regard as good. And this, Wood thinks, is a premise that Kant uses as a part of his overall argument.

The objects of our choices, when we choose wisely, are indeed (Wood's Kant thinks) objectively good independently of our decisions to choose these things. But what is good about such ends, Wood suggests, is usually (or perhaps always) that they relate in certain ways to us. As Wood himself puts this point:

> …setting an end is an exercise of *practical reason* only to the extent that we think there is already *some good reason* for us to set that end. The value of the end is to be located in that reason, which must have existed already prior to our rational choice. Of course, if it is true that the sole fundamental and unconditional value is the value of rational nature as an end in itself, then the goodness of any other end must somehow be grounded in this value. Ends to be produced will usually have value, for instance, because they fulfill the needs, or enrich the lives, or contribute to the flourishing and happiness of rational beings, and so setting and achieving these ends shows respect and concern for the value of those rational beings.[223]

---

and inclinations – and for the various physiological, psychological, and social conditions which gave rise to those desires and inclinations – we would not find their objects good. Kant saw that we must therefore take ourselves to be important. In this way, the value of humanity itself is implicit in every human choice. If complete normative scepticism is to be avoided – if there is such a thing as a reason for action – then humanity, as the source of all reasons and values, must be valued for its own sake." (Korsgaard 1996a: 122)

[223] (Wood 2008: 92) In between (Wood 1999) and (Wood 2008), Wood seems, perhaps under the influence of (Regan 2002), to have changed his view in the direction of a view that sounds much more like some type of mind-independent value realism than the Kantian view that "[a]ll the objects of our desires only have a conditional value; because if it were not for these desires, and the needs based on them, these objects would be without any value. The desires themselves, however, as sources of needs [thereby] have so little absolute value that being rid of them must be a universal wish among all reason-endowed beings. Thus the value of all the objects/states of affairs we're trying to *bring about* through our actions is at all times conditional." (G: 4: 428)

Hence the main problem with Korsgaard's reconstruction, Wood thinks, is the theory of value that Korsgaard attributes to Kant. According to Wood, Kant accepts something along the lines of mind-independent realism about the non-moral value of ends, i.e. the view that some possible ends have an irreducible value quite independently of whether these are objects of desire and/or things that appeal to us.[224] Using that basic assumption, how does Wood reconstruct Kant's argument?

Wood first claims that in acting on what we see as independently good ends, we must necessarily regard our "own rational capacities as authoritative for what is good in general."[225] But to do so, Wood then claims, is "to esteem oneself – and also to esteem the correct exercise of one's rational capacities in determining what is good both as an end and as a means to it."[226] And esteeming oneself and the "correct exercise of one's rational capacities"[227] is already to in effect regard oneself as *an end in itself* and to do so on the basis of our possession of these rational capacities.

Having come that far, the "next step in Kant's argument" is then, Wood claims, to extrapolate to all other beings with practical reason and to claim that they all have the "same rational ground" for representing their own existence as an end in itself. It follows, Wood concludes, that each person is committed to the claim that "not only my rational nature, but the rational nature in every person, is an end in itself."[228] And if that is so, then we ought rationally to govern ourselves on the basis of the humanity formula, i.e.

[224] Thus Wood calls the non-naturalist realism about reasons for action that Parfit propounds in (Parfit 2011a-b) "good Kantianism". (Wood 2011)
[225] (Wood 2008: 91)
[226] (Wood 2008: 91)
[227] (Wood 2008: 91)
[228] (Wood 2008: 92)

never treating the "rational nature" in any person as a mere means, but always at the same time treating it as an end in itself.[229]

The main disagreement between Korsgaard and Wood, then, is about whether our choices *make* our objects of choice good due to our own inherent importance (Korsgaard's view) or whether our choices are themselves value-tracking, but things nevertheless are good just because they relate in essential ways to us, who are inherently important (Wood's view). This disagreement aside, there is nevertheless, between Korsgaard and Wood, a broad agreement that Kant's reasoning has to do with how our own value (which we are supposedly somehow committed to simply in virtue of valuing our chosen ends) is what makes us into ends in themselves, and thus what grounds the humanity formula.[230]

And many other contributors to the literature seem to be in broad agreement, too, even though there of course is lots of disagreement about the exact details of how to understand Kant's reasoning. Julia Markovits, for example, seems to flatly accept Korsgaard's reconstruction without any reservations.[231] Samuel Kerstein, as we saw above, is in broad agreement with Korsgaard (though perhaps leaning a little more in the direction of Wood's reconstruction) about how to reconstruct Kant's argument (see quote of his summary in the introduction above), but argues, unlike Korsgaard and Wood, that

[229] Kant's argument, says Wood, is a little bit like John Stuart Mill's "proof" of the principle of utility, because what Kant tries to do, in arguing for a basic value, is to show that it is something that we are already committed to "in theory and practice". (Mill 1861/1963)

[230] That is to say, the universal law formula's being the constitutive principle of autonomous agency – as Kant argues that it is in the third part of the *Groundwork* – is not, according to this line of interpretation, any part of Kant's argument for the humanity formula. The argument instead only has to do, they argue, with what valuing our chosen ends necessarily commits us to, and how these commitments ought rationally to be honored in our actions in relation to ourselves as well as in our interactions with those around us.

[231] (Markovits 2011)

Kant's argument fails.[232] Like Wood, Geoffrey Sayre-McCord thinks that Kant's reasoning is of the same general sort as John Stuart Mill's (as in the reasoning that Mill uses in his "proof" of the principle of utility), which means that Kant is trying to show that the intrinsic value of rationality (what Sayre-McCord understands by Kant's claim that humanity exists as an end to amount to) is something we are "already committed to in theory and in practice."[233] Timmermann, in turn, may disagree with the idea that Kant is "regressing" on conditions of value towards something that, in being the condition that must hold for anything to have value, is of unconditional value. But he agrees that what Kant's argument does is trying to locate something of unconditional value, which is to be treated as an end. And in short what Timmermann takes the argument to do is simply survey different options before finding that rational beings are the only things to which we could attribute this kind of value. Only beings with practical reason, Timmermann takes Kant to be arguing, are "fit" to be regarded as absolutely and unconditionally valuable (thus Timmermann's analysis attributes something along the lines of a so-called "fitting attitudes" analysis to Kant[234])[235]. Rae Langton, in turn, summarizes Kant's reasoning as follows:

> Kant thinks [that] human beings have an intrinsic worth that has its basis in our capacity for rational choice. Human beings are ends in themselves, who have a dignity not a price. The moral law is the requirement to recognize and respect this dignity, and to act in a way consistent with it.[236]

---

[232] (Kerstein 2008)
[233] (Sayre-McCord 2001: 331)
[234] On "fitting attitude" analyses of different types of value, see, for example, (Rabinowicz & Rønnow-Rasmussen 2004) and (Scanlon 1998).
[235] (Timmermann 2006:)
[236] (Langton 1992: 487)

Paul Guyer takes Kant's argument to simply be a philosophical sharpening of the looser claim at the beginning of the *Groundwork* according to which the good will is the only thing that is unconditionally good: what Kant does, Guyer takes it, is to argue that a good will is really an autonomous will and that what has unconditional value, and is therefore to be treated as an end, is more precisely the autonomy of rational beings. Thus Guyer writes that:

> That the human will does not depend on anything else is intended, I believe, to be a necessary condition of its suitability for stopping the regress of merely conditional goodness. But that freedom has an "inner value, i.e. dignity," is the fundamental normative fact that makes this necessary condition sufficient for this purpose. It is, as I understand it, the premise on which Kant's argument rests, not the conclusion to be drawn from it.[237]

Guyer's main argument for why we should accept his reading is that when Kant gave his academic lectures "to the teenagers" at the Königsberg University, and therefore needed to simplify the exposition of his views, he was transcribed as having argued in a way that suggests this view.[238]

Resounding these same kinds of themes, Hill similarly writes that the formula of humanity appears to "go beyond the famous first formula" since it, unlike the universal law formula, appears to declare "a rather substantive value judgment with significant practical implications."[239] This means, I take it, that Hill agrees with the general idea that the reasoning leading up to the statement of the formula of humanity most importantly involves a basic assertion of a substantive value, and that it thus is distinct from the

[237] (Guyer 1998: 33)
[238] Guyer doesn't think, he says, that analyzing Kant's published texts, or trying to "massage any single expression in the *Groundwork*" to fit one's favorite interpretation, is sufficient to settle how exactly to understand his reasoning. (Ibid.)
[239] (Hill 1980: 98)

overall reasoning that continues into the third part of the *Groundwork*, which there importantly involves the claim that the universal law formula is the constitutive principle of an autonomous and, therefore, positively free will.

I regard these writers as all being members of a more or less unified school of Kant interpretation (with regard to how to understand Kant's reasoning surrounding the humanity formula), namely what I am calling the American school of Kant interpretation. (I'm using the term the "American school" because with the exception of Timmermann, all these writers teach at American universities.) The most distinctive feature of this school of interpretation is that it understands the argument for the humanity formula as using a premise that attributes absolute value to the reason-endowed being as its key premise, a premise that, according to the school, is not defended with the arguments having to do with the universal law formula's being the constitutive principle of autonomous agency that Kant gives in the third part of the *Groundwork*, but which instead is supposed to gain support on other grounds (and here members of the American school disagree amongst themselves).

So ought we to sign up for the American school's general understanding of Kant's argument for the humanity formula (at least initially setting aside the exact details of how Kant's argument goes)? I believe not. In the next section below I offer three arguments for rejecting all reconstructions along the lines of those of Korsgaard and Wood (and the others) that we have just reviewed above. In sections six and seven I offer an alternative reconstruction, which builds on the interpretative claims I have already argued for in chapters one and two.

5. Three Objections to the Just-Reviewed Reconstructions of Kant's Reasoning

If the American school's closely related reconstructions of Kant's argument for the humanity formula (and the reasoning surrounding it) were correct, then we should be able to confirm its correctness by checking the main features and implications of these reconstructions against major themes and claims in Kant's main texts and finding that there is a match between the main features or implications of these reconstructions and the given themes and claims in Kant's texts. In this section I will argue, however, that when we do this we quickly find that these reconstructions have features and implications that clash with major themes and claims in Kant's texts. I will therefore conclude that the reconstructions associated with the American school of Kant interpretation fail.

We will consider three arguments of this just-described form. Putting them together into one main argument, here is what we get:

Premise 1: If the reconstructions of Kant's argument for the formula of humanity reviewed above were on the right track, then (1) we would expect Kant to *not* take the universal law and humanity formulas to be equivalent; (2) we would expect Kant to attribute absolute value to the very same thing that he claims to exist as a purpose in itself; and (3) we would expect there to be a nice match between the reasons why humanity exists as an end in itself defenders of these reconstructions thinks of Kant as giving in the *Groundwork* and the various claims Kant makes here and there in his other main ethical works about why our humanity is a purpose in itself.

Premise 2: We find, however, that (1) Kant *does* of course take the universal law and humanity formulas to be equivalent (at least as these apply to human beings); (2) Kant *doesn't* attribute absolute value to the very same thing he claims exists (and ought

to be treated) as a purpose in itself; and (3) the explicit reasons Kant himself gives for humanity's being an end, and something not to be treated as a means only, in the *Critique of Practical Reason* and *The Metaphysics of Morals* do not match up well at all with the reasons that the American school supposes that he has for this view.

In light of these premises we must draw the *conclusion* that a review of some of the main themes and claims that Kant makes does *not* confirm the readings offered by members of the American school, but instead seems to refute their reconstructions. Now that we have a rough overview of what our three arguments are going to be, we can go through them more carefully, one by one.

Our first objection, then, is that if the American reconstructions were correct, then we would expect Kant to not make his equivalence claim between the universal law and humanity formulas, but that since he, of course, famously *does* make this very claim, we here have an argument that counts strongly against these reconstructions. Thus Kant writes, for example, that:

> The principle, so act with reference to every reason-endowed being (yourself and others) that in your maxim it is at the same time a purpose in itself, is thus at bottom the same as the basic principle, act on a maxim that at the same time contains in itself its own universal validity for every reason-endowed being. For, to say that in the use of the means to any end I am to limit my maxim to the condition of its universal validity as a law for every subject is tantamount to saying that the subject of [all] ends, i.e. the reason-endowed being itself, must be made into the basis of all maxims of action, never merely as a means, but as the supreme limiting condition in the use of all means, i.e. always at the same time as an end.[240]

[240] (G: 4: 438) The way I am reading this idea, to repeat, is that whether some maxim (or basic guiding principle) "contains in itself its own universal validity for every reason-endowed being" is a matter of whether each reason-endowed being could adopt and act on the basis of this maxim and its at the same time being possible for each reason-endowed being to preserve and fully realize their own distinctive nature. This means that following this principle would be to choose one's maxims in a way that amounts to making the preservation and full realization of the reason-endowed nature into the most general purpose around which all of our actions are organized. And

Why do I say that if these reconstructions were correct, then we would expect Kant *not* to make this so-called equivalence claim? Because according to the reconstructions we've looked at, the humanity formula asserts that humanity has absolute value and instructs us to treat it accordingly, whereas – according to these same writers – the universal law formula does *not* contain an assertion of any value.

As a reminder: Korsgaard thinks that the formulas cannot be equivalent because as she interprets them, they differ in their substantive implications. If that is so, Korsgaard does not (or so it would seem) need to worry about claims of the just-quoted sort: because these claims must be mistaken. We have also already seen that Wood thinks that Kant didn't actually seriously mean that these formulas are equivalent. Since the humanity formula, as Hill puts it, asserts "a rather substantive value" whereas the universal law formula doesn't, Wood claims that the universal law formula is but a mere step in a search for the basic principle of morality that is not concluded until Kant gets to the humanity formula. And so, on Wood's reading, just as on Hill's reading, the two formulas are not two sides of the same coin, and claims such as the one I have just quoted above are not to be taken seriously.

There is a better – and indeed more charitable – way of reading Kant. And that is, of course, to instead reject these reconstructions, take Kant's word for these formulas' being two sides of the same coin – a claim that he repeatedly makes! – and for this reason look for a different way of understanding the reasoning surrounding the original

---

so the requirement to act on maxims that are fit to serve as universal laws is equivalent to the requirement to always treat the reason-endowed nature as a purpose, and never as a means only. That the members of the American School cannot, in contrast, explain the equivalence of the two formulas is, I take it, itself enough reason to reject their readings. But as we are seeing, there are also further reasons to do so.

statement of the humanity formula. If we instead, as I think we should, understand the reasoning surrounding the initial statement of the humanity formula as being a part of a larger argument that stretches all the way into the third part of the *Groundwork* – not least since Kant says he is going to defend one of his main premises in the argument for the humanity formulation there! – then we are free to accept Kant's claim that the humanity and universal law formulas (at least as the latter applies to human beings) are two different statements of the same law. And since we have already presented an interpretation of the universal law formula on the basis of which we can explain this claim of Kant's, and we have supported it with a number of arguments in chapter two, we have all the more reason to reject these reconstructions, and instead look for an alternative.

Turn next to the second objection. This is the objection that if these other reconstructions were right, then we'd expect Kant to attribute absolute value to the same exact thing that he claims to exist as an end in itself, but that Kant doesn't do this, which suggests that these reconstructions cannot be right. What Kant attributes absolute value to is what he calls "morality" (by which he means *a good will* operating in accordance with the moral law). As he famously puts it at the very opening of the *Groundwork*:

> There is nothing in this world – or indeed even outside of it – that can be held to be unconditionally good, expect for a good will.[241]

But "humanity", which is said to exist as a purpose in itself, is *not* identical to a good will. And members of the American School do themselves recognize this. According to Korsgaard, for example, our humanity consists in our capacity to set our own ends of

[241] (G: 4: 393)

action[242], and this is not the same thing as a good will. According to Wood and Hill, our humanity is the capacity to set our own ends of action along with the various rational capacities we have that are intimately related to the capacity to choose our ends.[243] This set of capacities is also not identical to a good will (since they can be used in ways that don't amount to having a good will). And so on.

This means that defenders of the American way of reconstructing Kant's argument for the humanity formula are left in an awkward situation: they are claiming (and this is one of the key features of their reconstructions) that possessing absolute value and existing as an end in itself are one and the same property, and yet they understand our humanity as being something different than that to which Kant attributes absolute value. This tension, of course, hasn't been lost on everyone, and Richard Dean even uses it to develop an argument to the effect that writers such as Korsgaard, Hill, and Wood are wrong about what Kant means by "humanity".

What Kant means by "humanity," Dean argues, is simply the possession of a good will. Thus on this "good will reading", as Dean calls it, we can make sense of Kant's claiming both that the good will is the only thing of unconditional value *and* that we are to treat humanity as an end in itself: "humanity," on this reading, just refers to a good will. And so, it seems, the American school's way of reconstructing Kant's argument for the humanity formula could be saved: all its defenders have to do is to follow Dean in taking "humanity" to refer to good wills.[244]

Now I follow Dean in rejecting Korsgaard, Hill, and Wood's understanding of what Kant means by "humanity", since I don't think it simply refers to our capacity to set

[242] (Korsgaard 1986a: 187)
[243] (Wood 1999; 2008); (Hill 1980)
[244] (Dean 2006; 2011)

ends (nor even this capacity taken together with the rational capacities involved in end-setting)[245]. But I don't think Dean solves the interpretative problem that he discusses in the right way. The right way to go is instead to recognize that Kant *doesn't* understand *something's existing as a purpose in itself* to be the same as its possessing absolute value, which means, if I am right, that there is no interpretative awkwardness to be dealt with in the first place since – as I am suggesting – what Kant attributes absolute value to is not the same things as those that he claims to exist as purposes in themselves.

On my reading of what Kant means by our *humanity*, our humanity consists in our distinctive nature as reason-endowed beings that despite being animals with certain needs, desires, etc., nevertheless are also capable of autonomous agency. This reading – unlike the other readings we're discussing – helps to explain how it is that when Kant goes through what he regards as our main duties to ourselves in the *Metaphysics of Morals*, he divides up into duties that we have towards ourselves *as animals* of the particular kinds that we are *and* duties that we have to ourselves as (what he calls) "moral beings" (by which he means beings capable of autonomous agency).[246] With regard to the former, we have, for example, basic self-directed duties to preserve ourselves and the distinctive capacities associated with our animal nature, Kant claims. This claim makes perfect sense if our animality, as I believe it is, is a *part* of what Kant means by our humanity: we are to treat the humanity within ourselves as an end, and if the animal side of our nature is part of our humanity, then treating it as an end is one part of treating the humanity within us as a purpose in itself.

---

[245] I understand our humanity as involving these things, but as not being exhausted by them. Because "humanity" also, as I am about to also say in the running text, refers to our partial nature as animals with desires and needs, and our capacity to autonomously act in accordance with self-adopted principles fit to serve as universal laws.

[246] (MS: 6: 419-20)

The claims that Kant makes about duties to ourselves with respect to the animal side of our nature do *not* make sense if – following Korsgaard, Wood, and Dean – we understand Kant as only meaning our rational or autonomous capacities when he talks about our "humanity".[247] But these claims do, as we have just seen, make good sense if we understand our humanity to also involve the animal side of our nature, and we furthermore also remember that Kant often points out that the human will is such that it is affected by our animal side, but not determined by it. So we should join Dean in rejecting Korsgaard and Wood's readings of what Kant means by "humanity," but we should do so for a reason that also shows that Dean is also wrong about what Kant means when he uses this term.[248]

Return now to how, on the one hand, Kant claims that only a good will has an absolute and unconditional value at the same time as he also claims, on the other hand, that humanity, which is *not* the same as a good will, is and ought to be treated as a purpose in itself. The only way to make sense of this pair of claims, I am arguing, is to distinguish between Kant's claims about absolute value and Kant's claims about

---

[247] See also the argument against the readings of the humanity formula offered by Korsgaard, Langton, and O'Neill in section 6 of chapter 4 below.

[248] In response to this, Korsgaard might point to the sentence on which her reading of our humanity as exhausted by the capacity to set ends hangs, namely Kant's claim that "the capacity to at all set ends is what's characteristic of *humanity* (in contrast to *animality*)" (MS: 6: 392) Given the context – the context being that Kant argues that we have duty to preserve ourselves in the particular nature we have as the animals that we are – we should not, however, read this sentence in the way that Korsgaard does. We should read it as saying that what is characteristic of humanity is not its animal side, but rather the side of humanity that other animals don't share, namely our practical freedom. This reading allows for the animal side of our nature to be one part of our humanity, as I think Kant thinks it is, without its being the distinct characteristic that sets us apart from other animals (that characteristic being our capacity for autonomy).

(Korsgaard uses a different translation of the just-quoted sentence – namely, Gregor's translation that renders the sentence "the capacity to set any end whatsoever is the characteristic of humanity (as opposed to animality)" – but as Jens Timmermann also points out, that translation is not accurate. For Timmermann's discussion of this point, see footnote 37 at (Timmermann 2006: 91).)

humanity as a purpose in itself. Since the American school's reconstructions crucially depend on *not* making such a distinction, and not doing so makes Kant come off seeming as if he is flatly contradicting himself – i.e. claiming that a good will both is, and is not, the only thing that has unconditional value – we should, I again suggest, reject these readings.

And if the two already given objections are not enough, turn lastly to our third objection against the readings of Korsgaard, Wood, and the other members of the American school. This is the argument that in order to be acceptable, these reconstructions of Kant's *Groundwork* argument should fit with the reasons that Kant gives in other major works for why humanity exists as a purpose in itself; but that they don't match up with those claims, since these other claims in the second *Critique* and in the *Metaphysics of Morals* are *not* about what valuing our ends commits us to, but instead about other things; and that these readings therefore – and thus for yet another reason in addition to the two already adduced above – are not acceptable.

Consider first this crucial passage from the *Critique of Practical Reason*:

> The moral law is *sacred* (inviolable). The human being is certainly unholy enough, but the *humanity* within her must be sacred to her. In the whole of the creation, anything that in relation to which one has a will and over which has power may be treated *as a means only*; only the human being and with him every reason-endowed being is *a purpose in itself*. For he is the subject of the moral law, which is sacred, in virtue of the autonomy of his freedom. Precisely for this reason is every will, included one's own self-directed will, constrained under the condition of agreement with the autonomy of the reason-endowed being, namely, to not subject him to any purpose that isn't possible in accordance with a law that could spring out of the affected subject's own will; thus never to treat [this subject] as a means only, but always at the same time as an end. This condition … rests on the *personality*, which is the only thing that makes [reason-endowed beings] into purposes in themselves.[249]

[249] (KpV: 5: 87)

By our "personality", as Kant explains in the paragraph right before the one containing the passage above, he means "the freedom and independence from the mechanism of the whole of nature, … [i.e.] the capacity of a being considered as peculiarly subject to laws she gives herself."[250] Our "personality", in other words, is another name for our *autonomy*: the capacity we have as reason-endowed beings to operate in accordance with self-adopted guiding principles that are fit to serve as universal laws for all beings that have the same nature as we do. According to the passage we've just looked at, then, it is our autonomy, and our autonomy alone, which makes us into beings that exist as purposes in themselves. But according to the Korsgaard-Wood readings, our being ends in ourselves depends on our being sources of value, and our reason to think this, they add, has to do with what valuing our ends commits us to. These readings, therefore, clash with this crucial passage in the second *Critique*, which is one of two passages in that work that argues that we are purposes in ourselves in virtue of how we are the subjects of the moral law and thereby capable of autonomy.[251]

Consider next this rather dense, but nevertheless crucial passage from the *Metaphysics of Morals*:

> The highest principle of the teaching of virtue is this: act in accordance with a maxim of *purposes* the having of which can be a universal law for everyone. – In accordance with this principle the human being is a purpose for himself as well as for others and it is not enough that he is not permitted to treat either himself or others as a means only (whereby he can also be indifferent to [himself and to others]), but to make the human being into an end is in itself a duty of human beings.
>
> This basic principle of the teaching of virtue admits, as a categorical imperative, of no proof, but nevertheless [admits] of a deduction out of pure

[250] (KpV: 5: 87)
[251] The other passage is at (KpV: 5: 132)

> practical reason. – What *can* be a purpose within the relation of human beings to themselves and to others, that *is* a purpose before pure practical reason, because it [i.e. our humanity[252]] is a capacity for purposes in general; to be indifferent to this same thing, i.e. not to take an interest in it, is thus a contradiction; because in that case [our practical reason] would also not determine the maxims of our actions (which always contain purposes), [and] hence not be any practical reason. Pure reason, however, cannot command any purposes a priori, other than insofar as it at the same time announces these as duties; which duties are then called duties of virtue. [Therefore, it is a duty of virtue to treat the human being as a purpose, and never as a means only.][253]

This is one of the passages of the whole of the *Metaphysics of Morals* that are the hardest to understand precisely, but a few things seem clear, and these things suggest that Korsgaard, Wood, and the others are wrong about in what sense and why Kant takes the human being to exist as an end in itself (at least if Kant means to be using a similar argument here for this claim as the one he uses in the *Groundwork*). What is clear here is that Kant is intending to give a constitutivist argument according to which the humanity formula is a constitutive principle of autonomous agency: he is arguing that our capacity to be governed by maxims that we ourselves adopt can only be secured if we make the humanity in ourselves and in others into something that we always treat as an end, and never as means only.

Whatever the exact details of the argument Kant is sketching are, it is clear that they don't have to do with what valuing our chosen ends commits us to. The argument, which is supposed to show that the humanity within us exists as something that we must

---

[252] As the way that I am quoting this passage suggests, one needs to fill in things here in there (when one translates it) in order for it to make good sense in English. Why am I, for example taking the "it" here as referring to our humanity? It is because Kant says, in the above-quoted passage, that the "it" here is a capacity for purposes in general, because he has just said that this (i.e. the capacity to set ends) is the most characteristic feature of humanity (in contrast to our animality) a few paragraphs before, and because the German "sie" (translated here into "it") signals that Kant is referring to something whose definite article is "die", which is the case for "die Menscheit" (i.e. "humanity" with its definite article).

[253] (MS: 6: 395)

treat as a purpose in itself, instead has to do with the conditions that must obtain for us to securely possess and robustly be able to exercise our capacity to be governed by our own practical reason.

Note also that this passage is relevant to the first of our three objections, since this passage claims that *in accordance with the universal law formula* (in the version that Kant uses here), the human being is an end in itself. But the reconstructions of Korsgaard, Wood, Hill, and company involve the idea that the humanity formula asserts a value judgment that is *not* contained with the universal law formula. That clashes with the idea Kant expresses here, which is that humanity's being a purpose, and something that is not to be treated as a means only, is a direct consequence of the universal law formula itself (as it applies to human beings). And so we again see that we must abandon the interpretative suggestion that the humanity formula goes beyond the universal law formula by introducing or containing a substantive value judgment that is not associated with the latter.

Now if we understand the reasoning surrounding the initial statement of the humanity formula in the *Groundwork* as being part of a larger argument which is not concluded until the third part of the *Groundwork* – in which Kant argues that the universal law formula is the constitutive principle of autonomous agency – then the different reasons and arguments why humanity is a purpose in itself that we have looked at suddenly start to look like variations on the same theme. And that theme is that our being capable of autonomous agency is the reason why the humanity within us (which itself includes our capacity for autonomy) exists as a purpose in itself from the point of view of the formal principles of pure practical reason (i.e. principles of practical reason

abstracting away from our various subjective ends). Since that theme is *not* part of the reconstructions of Korsgaard, Wood, and other members of the American school, these reconstructions cannot offer a unified account of Kant's theory of humanity as an end in itself, and ought, therefore, be replaced with an alternative that can.

On the basis of the three arguments that I have now presented, we ought, I conclude, to reject the American School's reconstructions. What we need is instead to somehow understand Kant's reasoning surrounding the humanity formula in a way that relates it to Kant's argument in the third part of the *Groundwork* according to which the moral law is the constitutive principle of a fully autonomous will.

6. On the Absolute Value of a Good Will and the Worthiness to Be Happy

Return first to this claim:

> Supposing however, that there were something whose existence has absolute value, which, as *a purpose in itself*, could serve as the basis of certain laws, then in it, and in it alone, the basis of a categorical imperative, i.e. a practical law, would lie.[254]

I have claimed that Kant distinguishes between something's possessing an absolute value and something's existing as a purpose in itself; that the humanity formula itself does not contain the assertion of an absolute value (but does instead prescribe an "objective end"); and that we shouldn't understand the reasoning leading up to the initial statement of the humanity formula as making use of a premise asserting that humanity has absolute value.

If this is all correct, then there needs to be an alternative story about why it is that Kant in the above-quoted sentence not only requires there to be a purpose in itself in

[254] (G: 4: 428)

order for there to be a categorical imperative, but also additionally requires there to be something that possesses absolute value. Before getting to my own alternative reconstruction of the reasoning surrounding the humanity formula, I will, therefore, first (but more briefly) offer an account of why Kant claims that there needs to be something of absolute value in order for there to be a categorical moral imperative (and why it is not enough that there be something that exists as a purpose in itself). This has to do, I believe, with one crucial aspect of Kant's view of what moral philosophy is and also with his closely related view of how to understand moral motivation.

What I have in mind is a theme that runs through all Kant's major writings. Thus summarizing his own overall theory in 1793's "On the Saying: That May Work in Theory, but Not in Practice", Kant, for example, writes that moral philosophy is "not the teaching of how to be happy, but of *how to be worthy of happiness.*"[255] And already in the first *Critique* – which was originally published in 1781, and thus four years before 1785's *Groundwork* – Kant distinguishes non-moral motivation (which on the most general level amounts a desire for happiness) from moral motivation: motivation to be worthy of happiness.[256] Not only that; the *Groundwork* itself, of course, famously starts off from the distinction between the mere conditional goodness, or relative value, that our various qualities may have, on the one hand, and the unconditional and absolute moral worth of our character, which we may exhibit through the ways in which we conduct

[255] (TP: 8:278) The same claim, and indeed the exact same sentence, also occurs in the *Critique of Practical Reason*, at (KpV: 5:130).
[256] (KrV: B840-2; A812-4)

ourselves if our will is good, on the other. And the moral worth of our character is, of course, a measure of our worthiness to be happy.[257]

Now to make sense of this idea there is, I believe, an important, and perhaps somewhat subtle, distinction that we can draw here between being worthy of happiness and, as I will put it, deserving happiness. To bring out this distinction, one of the things we can first note is that on the substantive view Kant subscribes to, we are all under a requirement of virtue to make the happiness of others into an end of ours without thereby making the promotion of other people's happiness conditional upon whether or not these others are worthy of happiness.[258]

That is, the virtuous person makes the happiness of others into an end of hers quite independently of whether these others "deserve" happiness or not, and also quite independently of whether she reaps any direct personal benefit or reward for this. So we all, in other words, stand to benefit from the fact that those who are virtuous have as their aim to promote our happiness without any expectation of a reward, without taking into account whether we deserved to be treated in such a way or not.

This raises the question, from the point of view of us as recipients of this "practical love", of *whether we are worthy* of that happiness that these virtuous agents

---

[257] Thus Kant rounds off the very first paragraph of the first chapter of the *Groundwork* – i.e. the paragraph that starts with the famous assertion that nothing can be thought of as unconditionally good expect for a good will – by writing that "a pure and good will" appears to "constitute the indispensible condition of the worthiness to be happy". (G: 4: 393) And in the *Critique of Practical Reason*, Kant writes that a person's "virtuousness" is a measure of "the value of the person and her worthiness to be happy" (KpV: 5: 111)

[258] "In accordance with the ethical law… "love your neighbor as yourself." The maxim of benevolence (practical love of human beings) is a duty of all human beings toward one another, whether or not one finds them worthy of love." (MS: 6:451) This, of course, is not the only requirement of virtue, but it is one of the main requirements. With regard to our relations to others, all the requirements of virtue that we are under can, Kant writes, be "traced back to the duties of love and respect," (MS: 6: 488) the former being the requirement to "make the happiness of others into an end of ours" or, in the Christian phrase Kant also occasionally uses, "to love one's neighbor".

have as their aim of trying to help us achieve. The virtuous person, therefore, tries to be, or become, worthy of this happiness that others are under a requirement of virtue to try to help her achieve. These others are not supposed, then, to make their help conditional upon her being deserving of their help, but she herself, if she is virtuous, will try to be or become worthy of it.

As this shows, this idea of morality as the worthiness to be happy makes sense even outside of a theological context, quite independently, that is, of whether we believe there to be a higher power that will provide everyone with all the happiness of which they are worthy in an afterlife. And correspondingly Kant draws a distinction between the "realm of purposes"[259] (the possible state of affairs in which everyone always treats the humanity in each other person, as well as in themselves, as a purpose in itself; a state in which all enjoy "moral friendship"[260] on terms of "love and respect"[261]); and, on the other hand, "the highest possible good": the realm of ends with the additional feature of everyone's also actually getting all the happiness they are worthy of.[262]

The latter, Kant thinks, is only possible if there is a higher power and an afterlife in which this higher power could give everyone all the happiness they are worthy of.[263] But the former – i.e. the realm of ends, or, as Kant also puts it, a "realm of virtue"[264] – does *not* require the existence of these things. And the moral law requires us to try to bring about a realm of ends quite independently of whether the highest good will ever be

[259] (G: 4:433-4)
[260] (MS: 6: 439-74)
[261] Ibid.
[262] (KpV: 5: 110-1)
[263] (KpV: 122-32)
[264] (R: 6:95)

achieved. We may *hope* for the existence of a higher power and an afterlife. And *belief* in these things may indeed make virtuousness easier for us. But, Kant thinks, because:

> morality [is] based on the concept of the human being as a free being that is, however, precisely for this reason bound to unconditional laws through her own reason, it neither needs any other being above him for him to recognize his duty nor any other enticement than the law itself for him to observe his duty. [265]

This all raises the question of when and under what conditions a person is worthy of happiness. There needs, Kant thinks, to be some state of being – some way of conducting oneself – that has an absolute, unconditional value that (at least in theory) could serve as a measure of whether we are worthy of the happiness, which could function as a standard to try to live up to in our pursuit of the worthiness of happiness, and that, therefore, could serve as a basis of a categorical imperative.[266] And that way of being – and conducting oneself – is, of course, being in possession of, and acting on the basis of, a good will: "a good will appears to constitute the indispensible condition for the worthiness of happiness"[267].[268] So in what we might call a "formal" sense, what the person who is

---

[265] (R: 6: 3)

[266] Note that the distinction that Kant draws between mere "relative value" and "absolute" or, as he sometimes also puts it, "inner value" mainly seems to depend on whether the value of something depends on its own inner constitution (as the absolute value of something does) or on its external relations (as relative values do). Relative value is either had as a market price, which means meeting certain needs or being the objects of certain desires, or as an affection price, which means having affective significance for somebody relative to their individual sensibilities. The absolute value of the good will, Kant in contrast thinks, derives or consists in its inner constitution, quite independently of what its effects, whether it meets anybody's needs, or whether it strikes a fancy in those with particular likes and dislikes. (G: 4: 434-5) Thus Kant writes "The essence of things is not changed by their external relations; and that which, without taking account of such relations, alone constitutes the worth of a human being is that in terms of which he must also be appraised by whoever does it, even the supreme being." (4:439)

[267] (G: 4: 393)

[268] Again, Kant thinks that this is a commonsensical enough idea that he rounds off the first paragraph of the *Groundwork*'s first section by noting that "a pure and good will appears to constitute the indispensible condition for the worthiness of happiness." The philosophical challenge here, Kant thinks, is, first, to seek out the operating principle of such a wholly good will and then, second, to show that this principle is a necessary law that applies to us simply in

trying to be virtuous – who is trying to be worthy of happiness – is aiming for is simply the acquisition of a good will.

But this earnest seeker of worthiness of happiness, who is thus trying to have a good will, cannot merely curb her own desire for happiness on the basis of the purely formal aim of being worthy of happiness; she must also have a *material* or substantive purpose in whose service to act. Otherwise her attempts to be worthy of happiness would be moral equivalents of fumbling in the dark, because human action, Kant thinks, always "contains" an end for whose sake our particular actions are chosen.[269] Her *having* this material purpose must be able to give her an "inner worth" on whose basis she can be worthy of happiness – since, again, it is part of the concept of morality, on Kant's view, that it has to do with the attempt to be worthy of happiness, not simply the pursuit of happiness. But any philosophical attempt to establish what that particular material end is must avoid all simple appeals to the idea of moral perfection itself (i.e. an absolute or inner value that makes us worthy of happiness) because otherwise we end up in a never-ending circle, or an empty theory of the sort Kant associates with some of the rationalists whose theories he sees himself as offering an alternative to (i.e. theories whose most basic principles are precepts such as "always do what is right," etc).[270]

So there has to be something that serves as the material end that is the most general purpose around which a virtuous person organizes all her actions: a purpose her

---

virtue of how we are beings possessing a will, or practical reason, which means, he thinks, that its being a necessary law for us must somehow be derivable from the very concept of the will of a being possessing practical reason.

[269] (MS: 6:381)

[270] Thus in his lectures on ethics, Kant writes that "If [for example] the question is, What am I to do in regard to my obligation?, and the answer is, Do the good and abstain from the bad, that is an empty answer," and, as such answers are often put forward by moral philosophers, "[t]here is no science so filled with tautologies as ethics; it supplies for an answer, what was actually the question, and question and answer to the problem form a tautology." (Kant 1997: 27: 264-5)

having of which amounts to her having a good will. And what has absolute value here is, therefore, *her having* this material end as her most general guiding purpose in life, not the material purpose itself. Having this end as our most general purpose around which our actions are organized amounts to, or constitutes, having a good will. The argument we use when we try to establish what that purpose is can, therefore, not involve a mere assertion that there is something that is such that having it as an end makes us worthy of happiness, since we need to know why it is that having this as our most general purpose makes us qualify as having an unconditionally good will (the having of which makes us worthy of happiness).

Now to understand the reasoning in the second part of the *Groundwork* that we're currently analyzing it is not enough to know that Kant is, in part, trying to find a purpose the having of which amounts to having a good will (or which is the most general purpose around which an agent who chooses her maxims on the basis of their fitness to serve as universal laws organizes her actions). We must *also* keep in mind that at this point in his discussion, Kant is equally concerned with whether we actually have any duties or, in other words, with whether the idea of unconditional requirements that apply to us simply in virtue of how we are reason-endowed beings actually makes sense at all. And when he is trying to solve *that* problem, what he needs to do, Kant thinks, is to show that there is something about "the very concept of the will of a reason-endowed being" on the basis of which we can argue that she is under the requirement to choose her maxims on the basis of their being fit to serve as universal laws.

It is when he starts engaging with this question by starting to analyze the concept of the will of a reason-endowed being that Kant gets to the point where he asserts that at

the same time as the will is, firstly, the capacity to act on the basis of representations of principles, it is also the case that the will always has *purposes* that serve as the "objective grounds for its self-determination" or, as I am putting it, in whose service it is that we're acting. So if there is a moral law that applies to all reason-endowed beings simply in virtue of the fact that they possess practical reason – and, in relation to human beings, a categorical imperative – then there also has to be some basic purpose that can serve as the "objective ground of self-determination" for the wills of all these beings, a purpose, as I will argue in the next section, which is tied up with the most basic constitutive principles of reason in accordance with which any will must operate.

I understand Kant's idea of something that exists as a purpose in itself, then, as the idea of something that is a purpose from the point of view of the principles that are constitutive of the autonomous will of a reason-governed agent (and I will, as I just said, develop this reading in the next section). And I understand Kant as taking this purpose to have an absolute value in the sense that *having this as our most general purpose* makes us worthy of happiness, since our having this end as our most general purpose constitutes having a good will, which is the only thing that is in itself unconditionally good. With a rough sketch of an alternative story of why Kant thinks there needs to be something that can give us an absolute and unconditional worth in place, we can now return to the question of how to understand the reasoning surrounding Kant's initial statement of the humanity formula.

7. How to Understand Kant's Argument for the Humanity Formula

Return, then, once more to the following passage:

> If there thus is to be a highest practical law, and, in relation to the human will, a categorical imperative, then it has to be such [a law] that it makes an *objective* principle of the will out of the representation of what is an end for everyone since it is a *purpose in itself*, [which] hence can serve as a universal law. The basis of this principle is: *the reason-endowed nature exists as a purpose in itself*. The human being necessarily conceives of its own existence in that way: thus far is it, then, a *subjective* principle of human action. But [it is] also [the case that] each reason-endowed being conceives of its own existence in this way, on the basis of the same reason that also holds for me*: thus it is also an *objective* principle, out of which, as a highest practical basis, all laws of the will must be derivable. The practical imperative thus becomes the following: *so act that always treat the humanity in your own person, as well as that in each other person, always at the same time as an end, and never as a means only*. [Footnote: I here put this proposition forward as a postulate. The grounds for it will be given in the last section.][271]

Once we (1) discard the American school's interpretative idea that the just-quoted reasoning – specifically the sentence asserting that *the reason-endowed nature exists as an end in itself* – involves "the assertion of a basic value"[272] (as Wood and others claim that it does), *and* we also (2) take a look at what exactly the main point Kant is arguing for in the last section of the *Groundwork* is, making sense of the reasoning above actually becomes much less difficult than it can otherwise seem to be. We have, of course, already rejected the American school's interpretative premise, so let us therefore now have a look at what, in broad outline, it is that Kant argues in the third section of the *Groundwork*.

Now Korsgaard, we have seen, interprets Kant's discussion as if the point Kant makes in this third part that matters to the argument above is contained in the brief reflections Kant makes at the very end about how reason seeks the unconditioned. But those are comments that come after all the main action is over and Kant is wrapping things up with some general remarks. We should instead look to the main claims and arguments Kant gives in the third part of his book as being the ones that matter to his

[271] (G: 4:428-9)
[272] (Wood 1999)

assertion that we necessarily must conceive of the reason-endowed nature as a purpose in itself. So let us do just that.

The main claims Kant argues for, as I understand them, are: (1) Each reason-endowed being with a will cannot but act *under the idea of freedom*, which means that each such being thinks of its own will as *not* being completely ruled by impulses, inclinations, or other influences external to her own practical reason.[273] (2) To secure this kind of independence from outside rule, the will – which nevertheless needs to operate in accordance with some laws – must operate in accordance with laws that it gives to itself.[274] (3) To achieve this kind of *autonomy* of the will, what the reason-endowed being can and must do is to choose its guiding principles on the basis of no other criterion than simply their fitness to serve as laws for all reason-endowed beings.[275] (4) But this – i.e. a principle that says to choose one's maxims on the basis of their fitness to serve as universal laws – is just the highest principle of morality or virtuousness.[276] So (5) to robustly secure the kind of independence of the will that we act under the idea of, and thus achieve the kind of autonomy of the will we are capable of enjoying, we must make the moral law (understood on the model of the universal law formula) into the basis of our choices of guiding principles.[277] And this means that (6) the moral law (to always choose one's maxims on the basis of their fitness to serve as universal laws) is the constitutive principle of an autonomous will, and thus a standard we must live up to in

[273] (G: 4:448-9)
[274] (G: 4:447)
[275] (G: 4:447)
[276] (G: 4:447)
[277] (G: 4:447)

order to robustly secure and fully realize our own nature as self-governing beings with a will of our own.[278]

To put things in the terms Kant uses in the *Religion*, which will take us back to the Platonic and Hobbesian themes from chapter one[279]: unless we subject ourselves to maxims that are fit to serve as universal laws for all self-governing beings, we shall remain in an "ethical state of nature" in which we may not necessarily be ruled by our impulses (since we may still possess different degrees of control over our desires and impulses), but in which there is an internal war of all against all among the different motivational influences and impulses within us, all of which are trying to assume power over us. The only way to securely and robustly accomplish a state of inner lawfulness in which we can be truly responsible and in full control of ourselves – and in which the kind of peace and rule of law that we associate with a well-functioning juridical state will rein within us – is by subjecting ourselves to basic guiding principles of a sort all self-governing beings could follow and at the same time preserve and fully realize their nature as beings capable of this kind of autonomy even though they may also be subject to desires, needs, and other motivational states planted in them by outside influences.

Return now also to Kant's claim that the universal law formula and the humanity formula "at bottom are the same"[280]. We explained this claim in the following way in

[278] (G: 4:453-4)
[279] As we saw in the first chapter, the similarity between Kant on positive freedom of the will (and thus on human virtue) and Plato on the virtue justice in the *Republic*, in how both use an analogy between the *inner state* of a person's mind and the *outer state* of a polity, hasn't been lost on Korsgaard (quite the contrary). But Korsgaard herself hasn't yet related this aspect of Kant to the formula of humanity, for the reason (I think) that she doesn't believe that the universal law and humanity formulas are equivalent, but also (I also believe) because Korsgaard ascribes to, and is one of the main authors behind, the American school of how to understand the humanity formula and Kant's argument for it.
[280] (G: 4: 436)

chapter two. A maxim that is fit to serve as a universal law of the human variety of the reason-endowed nature is a basic guiding principle all human beings could be guided by and whose following would enable them to preserve and fully realize their own nature. But if we choose our guiding principles on the basis of their fitness to serve as such universal laws for the preservation and full realization of our own distinctive human nature, then we are in effect making the preservation and full realization of our own distinctive nature (our humanity) into the most general purpose around which all our actions are organized. So it follows that, as it applies to human beings (in which case we can also put things in terms of what I have called the human nature formula above[281]), the universal law formula is equivalent to the humanity formula: the requirement to always so act that we treat the humanity in each person as a purpose in itself, and never as a means only.

If this is all correct, then it follows that anything that is true of the universal law formula (as it applies to human beings) must also be true of the humanity formula. So if the universal law formula (again, as it applies to human beings) is the constitutive principle of the kind of autonomous agency we are capable of as beings possessing a will, then the same holds of the humanity formula. That is to say, the humanity formula is, in that case, an alternative way of stating the constitutive principle of autonomous human agency. So since Kant thinks that the universal law *is* the constitutive principle of autonomous agency, he therefore also thinks that the humanity formula, for this very reason, is another way of stating the constitutive principle of autonomous human agency.

[281] The human nature formula: act on the basis of a basic guiding principle (or maxim) all could act on and whose following would allow for and contribute to the preservation and full realization of our distinctive human nature. Using this application of the universal law and law of nature formulas to the human nature, it is easier to see how, as Kant claims, those more general formulas (when applied to us) are equivalent to the humanity formula. (See chapter one, section 4.)

This means – in other words – that in order to secure the kind of autonomous self-governance or internal rule of law we are capable of as human beings, we must in effect make the preservation and full realization of the humanity within each person into our most general purpose constraining all of our choices, and never treat the humanity in any person as a means only.

But this, of course, means that insofar as the human being cannot but act under the idea of freedom of the will, then human being must upon reflection realize that she can only fully and robustly secure the realization of her own nature as a being capable of autonomy by making humanity into something she always treats as an end, and never as a means only: only then can she emerge from the ethical state of nature. Or so, I believe, Kant means to be arguing.

In other words, when Kant writes that each reason-endowed being necessarily conceives her own reason-endowed nature as existing as a purpose in itself, I understand him as saying that upon reflection each reason-endowed being must come to realize that its own nature as self-governing being can only be robustly secured and fully realized on the condition of her adoption of basic guiding principles in accordance with which the realization of the self-governing nature is made into a purpose for everyone and a mere means for no one. Now if we "make" an objective guiding principle out of this purpose, which is fit to serve as a universal law, then what we get is just the humanity formula. And that, I believe, is precisely the argument Kant means to be making in the passage under consideration. [282]

[282] In case putting things in slightly different terms would be helpful, here's another attempt at summarizing this overall reading: As agents possessing practical reason we act and make our decisions under the idea of freedom: under the idea of not being determined by our impulses and desires, but as enjoying independence from the determining rule of such influences. But if our

Now if we understand Kant in this way, we can see why it is that he at one point claims that, "morality is the condition under the human being can be a purpose in itself." He means, I believe, that it is only in accordance with the basic principles of morality that the human being, or indeed anything, can be thought of as a purpose in itself (i.e. in virtue of its own nature). From the point of view of what Kant calls heteronomous agency, where our ends are set in accordance with or on the basis of influences coming from outside of our own practical reason, nothing is an end in and of itself: everything is instead a purpose only relative to some specific desire, need, or other external relation. From the point of view of autonomous agency, however, which is agency in accordance with self-adopted maxims of virtue that can serve as universal laws, there can be – and, Kant thinks, there are – things that exist as purposes in themselves since these maxims contain purposes, which are purposes, thus, in accordance with the principles that are constitutive of autonomous agency. Since the reason-endowed being, in this way, is a

---

practical reason (our will) is to be self-governing, then the laws in accordance with which it operates – it needs laws since everything that has a distinct nature, according to Kant, operates in accordance with laws – need to be laws that our practical reason gives to itself. What we can do in order to operate in accordance with self-given laws is to adopt maxims (i.e. basic guiding principles) that are fit to serve as universal laws in accordance with which all reason-endowed beings can exist and fully realize their nature as self-determining beings. But if we subject ourselves to such maxims, then we make the preservation and full realization of our own nature into the most general purpose around which all our actions are to be organized. So in order to fully realize our own nature as beings capable of being governed by their own wills, we must, in other words, subject ourselves to guiding principles whose following would amount to making the humanity within us (i.e. our distinct type of potentially self-governing nature) into our most general end. For this reason the humanity formula – so act that you always treat the humanity in each person as an end, and never as a means only – is one way of stating the constitutive norm of the kind of the kind of autonomous agency we are capable of. From the point of view of fully autonomous human agency, then, the humanity within each person exists as a purpose in itself, and must never be treated as a means only.

Its being a necessary law for all reason-endowed being to always judge its own actions on the basis of maxims that could serve as universal laws does, then, turn out being something that we defend on the basis of an analysis of the concept of the will of a reason-endowed being: because associated with this concept is the idea of the capacity for self-governance, and that can only be achieved through action in accordance with self-given laws, which in turn is only possible through the subjection to self-given maxims fit to serve as laws for all beings of our kind.

purpose in virtue of (a) its own nature and (b) the principles constitutive of the preservation and full realization of its nature it is, in that sense, a purpose *in itself* (and not, in other words, due to any external relation to that it has to the contingent desires or choices of any particular other agent outside of itself).[283]

This reading respects, and indeed makes use of, Kant's claim that the universal law and humanity formulas (as the former applies to human beings) at base are the same. It also fits with the observation that we have made that what Kant attributes absolute value to (i.e. morality or virtuousness) is not the same as what he claims to exist as an end itself (namely the reason-endowed nature, and, in our case, our humanity). Let's also see if our reconstruction fits with the passages from the second *Critique* and *The Metaphysics of Morals* we briefly considered above. According to the former passage, the human being exists as a purpose in itself in virtue of being subject to the moral law ("in virtue of the autonomy of his freedom"). This fits with our reading of the *Groundwork* argument, since it takes Kant's view that the moral law is the constitutive principle of autonomous

[283] Here's another way of explaining this point, which instead spells out the analogy between moral laws and laws of nature Kant uses. Recall that as Kant understands the term "nature" in what he calls its formal sense, a nature is a particular kind of constitution that exists in accordance with certain general principles, which are laws of its nature. This means that all laws have something they are laws of, which are the things that exist in accordance with these laws. Now Kant claims that the laws of nature are to serve as the "type" on the basis of which to model the maxims we are to adopt on account of their fitness to serve as laws. These laws therefore need to have certain things that exist in accordance with them. And it is also the case, Kant thinks, that maxims (i.e. basic guiding principles) contain or imply ends. So there needs to be some end that can be contained in all those maxims that are to serve as universal laws for all reason-endowed beings (or at least all human reason-endowed beings). This end is the thing that will exist in accordance with these laws of nature. And in our case it is, Kant suggests, we ourselves who are supposed to exist in accordance with these laws, and who, therefore, are the ends contained within all these maxims that can serve as universal laws. This, I believe, is another way of explaining why it is that, on Kant's view, *the reason-endowed nature exists and operates as a purpose in itself.* As Kant puts it in the *Critique of Practical Reason*, "the moral law is, in effect, a law of causality through freedom, or of the possibility of a supersensible nature" (a "supersensible nature" meaning a nature that, from the point of view of practical decision-making, is not subject to determination by the non-rational laws of natural necessity). (KpV: 5:47)

agency as the key to understanding why, as Kant argues in the *Groundwork*, the humanity within us exists as a purpose in itself. And the passage from the *Metaphysics of Morals* sketches an argument according to which the moral law is the constitutive principle of autonomous agency, which seems to be a rough version of the *Groundwork* argument, as we are reconstructing it. So none of the three objections we have used against the American school's way of reconstructing Kant's argument work against my alternative reconstruction.

In fitting so well with Kant's discussion of humanity as a purpose in itself across his different main works in ethics – and also not only respecting but making use of Kant's equivalence claim about the relation between the universal law and humanity formulas – my reading, I believe, ought to be considered as superior to the American School's closely related readings. We will end this chapter, however, by considering and responding to an objection having to do with what might appear as an advantage the American school's reconstruction has over our suggested reconstruction.

8. Is Our Reconstruction Uncharitable to Kant?

As we have been understanding Kant's argument for the humanity formula, when Kant claims that each reason-endowed being necessarily conceives of her own existence as a purpose in itself, what he means is, roughly, that the independence from the rule of her desires, fears, impulses, etc. that she thinks of herself as enjoying as an agent can only be robustly secured if she subjects herself to basic maxims that are fit to serve as universal laws of our nature and whose following, therefore, would amount to treating the preservation and full realization of the humanity in each person as a purpose in itself. We

cannot, in other words, securely enjoy the kind of governance over ourselves we self-attribute when we think of ourselves as agents fully responsible for their own actions – as agents who, in Korsgaard's terms, are the sole authors of our own actions – unless we exit the ethical state of nature we're in when there isn't an internal rule of law within our decision-making faculties. And the only way to achieve this is to ourselves adopt basic guiding principles – whose successful following might indeed involve some struggling, which Kant calls virtuousness – that have properties that make them fit to be laws of our own nature: basic principles in accordance with whose following our practical reason can govern us and all beings like us even though we – as human beings – are also subject to various desires, fears, impulses, passions, etc.

But to adopt basic guiding principles on the basis of their fitness to serve as laws for the preservation and full realization of our own nature as self-governing agents of the distinctively human variety is to make the humanity within us – indeed within each person – into the most general purpose around which we organize our actions. Because if the maxims we adopted and followed were only designed to help protect and realize our own capacity for self-governance, then our maxims would not have properties that would make them fit to serve as *laws* for the realization of the particular type of beings that we are. So to make an objective principle out of this end – i.e. the end of the humanity within us – we must make the protection and full realization of the humanity within each person into an end, and also thereby adopt the maxim never to treat humanity within any person as a means only (i.e. as something we only need to concern ourselves with if this serves our own personal goals). So choosing our maxims on the basis of their fitness to serve as universal laws of the reason-end nature (and in particular of the human variety of it) is

really the same as to make it our highest-order end to always treat the humanity in our own person, as well as that in each other person, as a purpose in itself, and never as a means only.

In understanding the argument Kant uses in the discussion which leads to his initial statement of the humanity formula in this way, we are rejecting the idea that Kant's reasoning has to do with what we commit ourselves to regarding our own value when we value the various ends we're pursuing, which is how members of the American school of Kant interpretation (in various different ways) think that Kant's reasoning should be understood. Our reconstruction instead follows Kant's own way of summarizing his own view, which most essentially involves understanding "morality as based on the concept of the human being as a *free being* and, however, precisely for this reason bound, through her own reason, to unconditional laws."[284] But this suggested way of understanding what is most central in Kant's reasoning – even if it accepted that this indeed is something that Kant himself regards as essential – might nevertheless strike some as uncharitable towards Kant. Indeed it might strike some as more charitable to understand the take-home lesson from Kant's works in moral philosophy to be something along the lines of the view that members of the American school attribute to him.

That, at least, might very well be how Thomas Hill – judging from the discussion in one of his Tanner Lectures – might object to our reconstruction. For in the published version of these lectures, Hill – likely representing the sentiment of many, though perhaps not all the members of the American school – writes that "associated" with Kant's views about respect for persons (the topic of Hill's lectures) there is "a radical

[284] (R: 6:3, emphasis added)

"two perspective" metaphysics that few philosophers today can accept."[285] To avoid embarrassment we should, therefore, treat "these as associated ideas that are inessential to Kant's central moral insights." What is essential is, instead, Hill thinks,

> the idea that the ultimate source of human values is humanity itself, rather than Platonic forms, natural teleology, God's commands, universal human sentiments, or particular social conventions.[286]

Hill seems to be suggesting here that in order to reconstruct Kant's overall reasoning in the only possible way in which it can be made plausible, we must shun everything Kant argues for regarding freedom of the will and instead follow Korsgaard's 1986 paper and see Kant's most important reasoning as essentially involving the argument that, as Korsgaard puts it,

> when we choose things because they are important *to us* we are in effect taking *ourselves* to be important. Reflection on this fact commits us to the conception of our humanity as a source of value. This is the basis of Kant's Formula of Humanity… [287]

Hill's claim seems to be that it is only if we take *that* kind of reasoning to be what is at the heart of Kant's discussion that we can extract any important ideas from Kant's claims and arguments.

Thus in response to our reconstruction of Kant's reasoning surrounding the statement of the humanity formula in the *Groundwork*, Hill might object that in taking this reasoning to essentially involve Kant's views and arguments about freedom of the will, we are focusing on radical ideas within Kant's metaphysics – particularly a "radical

[285] (Hill 1994: 37)
[286] (Hill 1994: 37)
[287] (Korsgaard 1996b: ix-x)

"two perspective metaphysics" – that, as Hill puts it, "few philosophers today can accept." What we are doing, Hill might conclude, is focusing on bits in Kant that should be ignored, where it would have been possible to instead, as he and other members of the American school do, focus on the Kantian claims about value (again, as he and members of the American school understand them).

In making this kind of objection, what would Hill (or somebody taking this line) be referring to? They would be referring to an implication of Kant's view that in practical decision-making and all deliberation we necessarily present different *options* as open to us; that in representing candidate options as ones we *ought* to take, we necessarily thereby represent these options as within our power (however strongly we may desire to act in other ways instead); and, in short, that all decision-making presupposes at least negative freedom of the will in the sense that our decision-making faculty enjoys independence from the external rule of irresistible desires and impulses. The implication, which Kant famously points out, is that in all our decision-making, we take up a different *standpoint* on ourselves than we do if we passively reflect on or observe our own psychology as an external bystander might do.[288] In thinking of ourselves as the authors of our own actions, we thus place ourselves, Kant thinks, in an *intelligent order*[289] in which we can ourselves ultimately decide and take responsibility for which laws are to govern our decision-making.[290]

---

[288] (G: 4: 450)
[289] (G: 4:452)
[290] And the argument Kant then makes, on the basis of what he regards as these necessary presuppositions of practical thought, is that we can only securely establish ourselves as members of such an intelligent order – in which we are the responsible authors behind our own actions, rather than the drives, desires, impulses and so on that we are subject to – if we subject ourselves to guiding principles that could serve as universal laws for all members of such a community, which would be basic principles in accordance with which all these members could fully realize

These are, I suspect, the kinds of claims that writers like Hill take to suggest that Kant thinks that there exists two wholly distinct realities both of which we as human beings are somehow mysteriously part of even though what happens in one reality has nothing to do with what happens, or is necessary, in the other. And such claims are, another critic complains, "not even vaguely intelligible."[291]

In response to such an objection to the reconstruction of Kant's reasoning offered above, I would like to put forward the following four points. First, the features of Kant's arguments about human agency that I have focused on leave open different interpretative options regarding how to more precisely understand Kant's claims about the different standpoints we can take up. Given the aspects of Kant's overall line of reasoning in the *Groundwork* that I have focused on – such as the idea that the way to securely guard against an uncontrollable inner struggle between different motivational forces is to adopt and continually try to follow basic guiding principles in accordance with which all could preserve, harmonize, and fully realize the main aspects of their dual nature as human beings (our animal nature and our capacity for self-governance through practical reason) – it is, of course, possible to further develop this view as one which involves the postulation of dual realities: a "world" of senses and a distinct "world" of pure reason (or something along those lines).

But nothing about the general features of Kant's view I have focused on requires us to think of human beings as inhabiting multiple realities. Nor is, I believe, this the best

---

their nature as potentially self-governing beings. Only then can we make our partial nature as animals with certain drives, desires, needs – and an accompanying desire for happiness: the state in which everything goes in accordance with our desires and wishes – harmonize with the other part of our nature: the part that can enable us not be to the slaves of these passions, namely our practical reason.

[291] (Parfit 2011a: 269)

way to understand Kant in the first place. As Korsgaard writes, the relevant part of Kant's view "is not, as so many have supposed, an ontological or metaphysical theory according to which we exist simultaneously in two different "worlds", one somehow more real than the other." It is rather that,

> insofar as we are rational, we also regard ourselves as *active* beings, who are the authors of our thoughts and choices. We do not regard our thoughts and choices merely as things that *happen* to us; rather, thinking and choosing are things that we *do*.[292]

Kant's most essential point, on this reading, is that practical thinking is a radically different *mode of thinking* than thinking that merely tries to understand something[293]: it

---

[292] (Korsgaard 1996b: xi)

[293] When it comes to merely trying to understand something (as opposed to trying to make some decision), our ways of thinking, Kant interestingly claims, divide into two kinds: one kind that is completely separate from all practical thinking, which is made possible in cases in which we are able to subsume phenomena under general principles or laws we are already operating with and another way of thinking, which is *modeled* on practical thinking, applying in cases where we are not able to subsume particulars under already given general principles or laws. This latter kind of thinking, which Kant calls *teleological*, functions by trying to organize and relate our objects of thought into means and ends, whereby we try to relate things we don't immediately understand to certain ends or purposes in relation to which they can serve some function.

This means that even when we take up the purely theoretical standpoint, we sometimes, if Kant is right, make use of the kind of thinking we use from the practical standpoint, from which everything divides into practical categorical such as those of *means* and *purposes*. But the idea is *not*, as I understand it, that we are justified in regarding the world as actually being teleologically ordered (insofar as we are unable to understand it using our best basic non-teleological theories and principles), but rather that it is a feature of the human mind that when we don't understand something, we are naturally prone to try to make sense of it using the categories of thought that come most naturally to us as agents, namely those involved in practical rather than purely theoretical thinking.

Thus Kant writes, "that things of nature serve one another as means to ends, and that their possibility itself should be adequately intelligible only through this kind of causality, for that we have no basis at all in the general idea of nature as the sum of he objects of the senses," but rather that "teleological judging is [sometimes] rightly drawn into our research into nature … but only in order to bring it under principles of observation and research in analogy with causality according to ends, without thereby presuming to explain it." (KU: 5: 359, 361) Practical thinking is indeed thus, strictly speaking, a radically different mode of thinking than theoretical thinking, although the former can sometimes by applied in the relation to the latter in order to get the latter started, or for the purpose of making useful analogies. For Kant's fascinating discussion of how

uses a different set of concepts, different principles, different ways of representing our own relation to possibilities (such as viewing them as options rather than things that might merely happen etc.), and so on. Accepting this important point does not require us to adhere to the mysterious dualism about worlds or realities that Hill and others worry about. But most of all: the features I've highlighted as central to Kant's reasoning allow for different more or less "metaphysical" interpretations. (And I myself, following Korsgaard, favor a "practical" interpretation: i.e. an interpretation about what is necessarily presupposed from the practical point of view of the decision-maker.)

Second, the worry that a writer like Hill has in mind might be that Kant's views about freedom of the will involves an unacceptable so-called "libertarianism" whereby free choices are understood as wholly uncaused and unbound occurrences that are not subject to any laws.[294] But this is not how we ought to understand Kant, and the reason is in part that, as we already saw in chapter one, that Kant thinks of freedom of the will on analogy with the freedom enjoyed by members of a Republic that enjoy a share in the making of the laws that are to govern them: that is to say, the will (i.e. our own practical reason) is free if the laws it operates in accordance with are not imposed on it by outside forces, such as our or other people's arbitrary whims, but instead derive from our own capacity to think about action in terms of general and universal principles. That's the first thing to note.

---

he thinks our power of judgment functions in cases in which it is unable to take already given principles and laws as its starting point, see the second book of the *Critique of Judgment* (KU).

[294] For "libertarianism" as a kind of philosophical theory of freedom of the will, see, for example, (Kane 2007).

The second thing to note is that Kant thinks of choice as being associated with a special kind of causality ("causality through concepts"[295]) – meaning that we assume ourselves to have the ability to bring about results in the world through our own decisions – and that since we cannot imagine causality without governing laws, the idea of a free will that is not subject to any laws would be something close to an absurdity or a non-existent thing.[296] Kant would, in other words, regard the libertarian idea of "free" choices as choices not bound by any laws at all as transforming our supposedly free choices into random occurrences, which, due to their randomness, are therefore not free.

So many (or perhaps all) of the main issues that critics of libertarianism about free will raise against that theory don't apply to Kant's theory (although some of the language he uses, such as talk about "spontaneity" might suggest that his position is a form of libertarianism). If the Hill-style worry about our reconstruction were that it brings up aspects of Kant's theory of a sort that many see as problems with libertarianism about free will (lack of laws applying to choice, etc.) then the worry would fail on account of how Kant's theory doesn't seem to be a libertarianism of the sort in question.

It should be noted in this connection that just as Kant himself was influenced by his way of understanding political freedom in his thinking about freedom of the will, so are, most likely, many contemporary readers of Kant's works. But what writers like Philip Pettit[297] and Quentin Skinner[298] call the "republican" conception of freedom – on

[295]This Kant writes, in the *Critique of the Power of Judgment*, that "…we have in the world only a single sort of beings whose causality is teleological, i.e., aimed at ends and yet at the same time so constituted that the law in accordance with which they have to determine ends is represented by themselves as unconditioned and independent of natural conditions but yet as necessary in itself. The being of this sort is the human being…" (KU: 5:435) Cf. footnote 293 above.
[296] (G: 4:446) I use this disjunction here since it is rather hard, I think, to directly translate the German word Kant himself uses when he makes this point, i.e. "Unding", into English.
[297] (Pettit 1997)

which freedom is the independence from an external dominating master, which in political contexts can be achieved through the rule of law – was much more in the foreground of the thinking of a writer in Kant's days than it is for readers today. Many contemporary readers instead naturally think of political freedom as consisting either in not being hindered from doing whatever one desires to do, or as the enjoyment of many different options and opportunities in life.

So the associations many contemporary readers have when they read about "freedom of the will", insofar as these are affected by their concept of *political* freedom, are likely to be different in important respects from those that are salient to Kant in his thinking about the concept of *freedom* in general. This, I suspect, is one of the reasons why people are sometimes more skeptical of Kant's views on freedom of the will than they need to be, and also why they are often inclined to think of Kant as offering something along the lines of a libertarian theory of freedom of the will.

Third, a view like Hill's – that suggests that we should ignore Kant's claims about autonomous agency, and instead focus on what members of the American school believe to be his views about value – is likely to ignore many of the key substantive normative issues that Kant and those that he draws and builds on (perhaps most importantly Plato, Aristotle, the main Greek Hellenistic schools, the Roman stoics, Christianity, as he interprets its ethics, etc.[299]) take a great interest in: namely, those essentially having to do with how to deal in our practical lives with the internal conflicts that can occur between, as Plato might put it, the "different parts of our soul".[300] One of Kant's main concerns is precisely that of what we can do as human beings to assume the kind of control over

[298] (Skinner 1998)
[299] See, for instance, (KpV: 5: 126-9)
[300] See the whole second part of (MS) and the whole of (AP).

ourselves that is required in order to be able to interact with others (and with ourselves) in a way that allows us to flourish as the particular kind of beings that we are, but without thereby vainly trying to lead overly rational lives that are not possible for human animals of our kind.[301] This is a classical virtue-ethical theme that runs through many of Kant's publications in moral philosophy that we might wholly miss out on and not be able to make sense of if we think that we should understand Kant's ethics as based on the idea of humanity "as the ultimate source of all value", and not as a theory that takes the preconditions of autonomous human agency as its ultimate basis.[302]

Fourth, Kant's constitutivism – i.e. his views and arguments about how to derive the basic principles of ethics from the preconditions of autonomous human agency – seems to precisely be the key to how Kant can offer a distinctive alternative to ethics based on the "Platonic forms, natural teleology, God's commands, universal human sentiments, or particular social conventions" that writers like Hill desire to avoid appealing to. So rather than taking Kant's claims about the two different standpoints we can take on ourselves – and the different norms and principles we can use to reason from these different standpoints – to imply some metaphysical extravaganza that takes away from Kant's theory's capacity to offer an alternative to these other kinds of moral

[301] Thus Kant thinks, as we have already noted, that the Stoics went too far in their attempt to free themselves of their feelings and inclinations, writing, as we've already seen, that, "Natural inclinations are, *considered in themselves, good*, i.e. unobjectionable, and to have a will to exterminate them would not only be futile, but also harmful and reprehensible; one must rather only tame them, so that they will *not* wear each other out, but instead can be brought to harmonize into a whole, called happiness." (R: 6:58)

[302] In the next chapter we will look at another classic virtue-ethical theme from Kant's distinctive point of view, when we consider what it is to flourish as a human being in accordance with Kant's idea of the highest possible good for a human being (as well as the highest possible good for a world as a whole). Just as Kant has a view that builds on the Ancient idea of ethics as partly being about how to achieve harmony within the soul, he also builds, in formulating his theory, on the Ancient idea that ethics should give us an idea of what the highest possible good would amount to.

theories, Kant's claims about the constitutive principles of fully autonomous human agency seem to me to be exactly what to turn in order to find an alternative to these other types of theory that writers like Hill wish to avoid.

On the basis of these four considerations, I submit that (1) we don't have good reason to think that it is more charitable to Kant to see his most important ideas as having been ideas about humanity as the "source" of all value; and also that (2) we rob ourselves of important opportunities to learn from Kant's works if we read his theory in that way (as the American school does), rather than reading Kant's theory as a kind of constitutivism of the general sort described in chapter one.

When Kant claims that the reason-endowed nature exists as a purpose in itself, he is not, then, meaning to assert a basic value; nor is he claiming that humanity is the source of all other value. Nor is he, as I have just argued, even best understood in this way for his theory to be of moral-philosophical relevance for contemporary ethics. In offering us a way of approaching the basic principles of ethics that derives these from the basic preconditions for the possibility of fully responsible self-governing agency, while at the same time recognizing that the human being also has (if you will) a "non-rational" side (which means that many of our personal values tied to our individual or communal conceptions of happiness are based in our "particularly constituted faculties of desire"[303] or our personal sensibilities), Kant offers us an important alternative to the intuitionist approach to moral theorizing that plays such a dominant role in much contemporary moral philosophy.

Another important aspect of Kant's ethics that sets him apart from much contemporary discussion and which – because this is often not how many contemporary

[303] (G: 4:428)

writers think of ethics – is often overlooked, forgotten about, or even ignored is Kant's oft-repeated distinction between what he calls *morality* and *legality*: between acting on the basis of maxims that are fit to serve as universal laws, on the one hand, and merely acting in accordance with maxims that are fit to serve as universal laws, on the other hand. The *permissibility* of different courses of action is a matter, on Kant's view, of whether these are courses of action that would be among the options that a fully virtuous person could regard as open in the circumstances on the basis of the principles of autonomy in accordance with which she is operating, not a matter of whether the agent performing the given actions *would* in fact be acting on the basis of virtuous maxims.

So the moral permissibility of some given course of action is not, on Kant's theory, determined by on what motives or maxims the agent performing the action would be acting. But Kant himself is most often not only concerned with the mere permissibility or impermissibility of actions, but also rather with whether acting in the ways he is discussing is compatible with virtuousness or "morality" (as opposed to mere legality, in his terms); and so his treatment of some of his main examples is not merely about permissibility, but also about the thicker notion of the *morality* of the actions in question. This has to do, of course, with how one of the main ways in which Kant thinks of moral philosophy, as we saw above, is as the teaching of how to be worthy of happiness. But this is often overlooked, and many contemporary discussions fail to understand Kant's views about what is permissible and what is impermissible as a result. And so many critical engagements with Kant's principles and examples end up involving objections to what are believed to be claims about permissibility on Kant's part, but which are really

instead claims about the morality or virtuousness of the imagined agents in the examples Kant discusses.

Other common misunderstandings – which also give rise to seemingly strong objections – have to do with what kind of moral ideal Kant sets out for a human life and the virtuous agent. It is often thought that the morally best life – the kind of life in which we would flourish morally as human beings – is a life in which we are indifferent to everything except for the moral law. And this has been thought to give rise to various objections, the simplest one being that this is not a very attractive ideal. But these objections, as we shall see, do not fit with Kant's own account of what it is for a human being to flourish in accordance with the moral law, set out in his vision of what he calls the "highest good."

It is furthermore also the case, as I will explain, that some standard interpretations of the humanity formula lack the resources to explain the relation between this formula and Kant's view of what constitutes the highest possible good. The last thing I will do in the next chapter – and in my overall attempt to offer a better understanding of the universal law and humanity formulas – is therefore to discuss and to try to throw new light on the relation the humanity formula and the highest good, as Kant conceives of it.

CHAPTER IV

Permissibility, Virtue, and the Highest Good

1. Some Distinctions

Our main topic in this dissertation is how to understand the universal law and humanity formulations of the categorical imperative, as well as the relation between the two. And we have seen that the moral law – which in relation to human beings takes the form of a categorical imperative, which can be expressed in either of these two ways[304] – is the constitutive principle of fully autonomous agency: agency whereby our ultimate governor is simply our own capacity to represent and act on principles, i.e. our practical reason.[305] The topic of this last chapter is how to understand Kant's views on human flourishing, virtue, and permissibility in relation to the humanity formula. As I will explain, some of the objections against Kant's theory in the recent literature depend on (what I believe to be) interpretative mistakes with regard to these very basic issues.

[304] These, however, are, if Kant is right, only two of many different ways in which it is possible to formulate the most basic categorical imperative that applies to us. Different commentators differ in how many different formulations they think Kant presents us with. I will not enter into that debate here.

[305] This doesn't mean that an autonomous human agent, as Kant thinks of her, neglects or ignores other aspects of her mental life (such as her inclinations and feelings), but rather that she tries to bring these other aspects into harmony with her partial nature as a being capable of autonomous agency, so that she can possibly achieve both autonomy and happiness at the same time – Kant would add – as being worthy of all the happiness that she desires. The goodness of her happiness, Kant thus thinks, is conditional upon its being consistent with her possession of a good will, but it is nevertheless, as we will see below, an intrinsic part of the whole and complete good of a human being.

Before describing the topic of this chapter in a little more detail, I will begin by first very briefly summarizing some of the main features of the above-presented account of how and in what sense the moral law is the constitutive principle of fully autonomous human agency. We have seen that Kant thinks that we have a dual nature because although we have the capacity to think about action in terms of general and universal principles (and also to be governed by this capacity), we are also animals with various desires, needs, passions, hopes, fears, and – insofar as these are brought into a unified whole – a general desire for happiness: the conceivable state where everything, insofar as it is possible, goes in accordance with our desires and wishes (many of which are social in nature[306]). Our nature is, therefore, most fully realized, Kant thinks, if we manage to bring our pursuit of happiness into harmony with our "moral" capacity to be governed by our own practical reason (i.e. our capacity choose and think about our actions in terms of general and universal principles).[307]

And we can only do that by subordinating our pursuit of happiness to the condition that we always act on the basis of basic guiding principles (or "maxims") that all human agents could act on and that could serve as universal laws for the preservation

[306] The human being, Kant furthermore writes, is also, with regard to the social side of her nature, under a "duty to [herself] and to others not to isolate [herself] (*separatistam agere*) but to use [her] moral perfections in social intercourse (*officium commercii, sociabilitas*)," and that the highest possible union between people is "*Friendship* (considered in its pefection) [which] is the union of two persons through equal mutual love and respect." (MS: 6: 476, 469) These ideas about sociability and friendship as being part of ethics can sound strange if one thinks of *duty* as many writers in the contemporary discussion do – i.e. as something that primarily has to do with what norms one ought to conform to in order to avoid being an appropriate object of blame, one's own guilt, or other kinds of internal and external punishments – but not so strange if one, like Kant and the Ancients he is drawing on, thinks of duty as a matter of what would allow us to both preserve and fully realize our own nature in accordance with the constitutive principles of autonomous human agency.

[307] …so that we stand a chance of achieving both happiness and autonomy. We would then, on Kant's view, completely flourish as a human being, achieving the highest possible good conceivable for a human being in accordance with the principles of pure practical reason. See sections 5 and 6 below.

and full realization of our particular nature. These moral laws take laws of nature (as Kant thinks of them) as their model: a law of a particular nature is a basic principle in accordance with which a certain kind of thing or phenomenon can exist and operate. But since these are "laws" that we must subject ourselves to, action on the basis of moral laws amounts to autonomy, i.e. self-governance on the basis of laws we give to ourselves, just as a group of people can constitute an autonomous republic by living in accordance with laws they have themselves decided upon.

Now if we choose our basic guiding principles on the basis of their fitness to serve as universal laws for the preservation and full realization of our particular human nature (as Kant thinks of it), we thereby make the preservation and full realization of the humanity within us into the most general purpose around which all our actions are organized, and we also in effect subordinate the pursuit of our own happiness to the preservation and full realization of the humanity within each person. This has the following implication: that because the basic law of morality is to choose one's maxims on the basis of their fitness to serve as universal laws of our nature, this law can also, as it applies to human beings, be described as an imperative to so act that we always treat the humanity within each person as an end, and never as a means only. As applied to human beings, then, the moral law can either be described using the universal law formulation of the categorical imperative or using the humanity formulation. And so we find that if Kant is right, the constitutive aim of fully autonomous human agency is nothing less than the preservation and full realization of the humanity within each person.[308]

[308] And so Kant's theory has the implication that it is only in accordance with the moral law that we can completely flourish as human beings. More on this in sections 5 and 6 below.

Now we have already looked – in chapter two – at various objections to the universal law formula. Those objections all misfire, I argued, because they are based on the various standard ways of understanding that formula, which are plagued with interpretative mistakes. The humanity formula is much more popular than the universal law formula, and the objections to it tend to take a less dramatic form. They tend to take the following form: it is argued that whereas Kant's idea that humanity is always to be treated as an end, and never as a means only, captures something intuitively compelling and important about morality, Kant is either (a) wrong to think that this serves as a good basic criterion of the moral permissibility of candidate courses of action or (b) too dramatic or extreme in his own views about how we should understand the practical implications of this imperative.

As before, I think that many of these objections to Kant's views depend on misunderstandings about what exactly his views are. This partly has to do, I believe, with how Kant's own basic concerns in his moral theorizing are slightly different from those of many of the most influential writers within contemporary Anglophone moral philosophy. And related to that issue is the further issue of keeping track of the most basic distinctions Kant draws within his overall theory as they relate to his discussion of the humanity formula in particular. By overlooking these distinctions at the same time as not being attentive to what Kant's basic concerns are within his moral philosophy, many writers are, I believe, led to objections that don't fully engage with the view Kant presents, but that instead attack a straw man.

So rather than trying to settle the foundational question of whether to accept the humanity formula, what I will do in this final chapter is to try to steer the discussion of

Kant's humanity formula in the right direction. And I will do this by focusing particularly on Kant's views about permissibility and virtue; Kant's view of the highest possible good for individual human beings as well as for possible worlds as wholes; the relation between these views and the humanity formula; and some objections against Kant's views that, as I believe, fail to take these views and these relations into account.

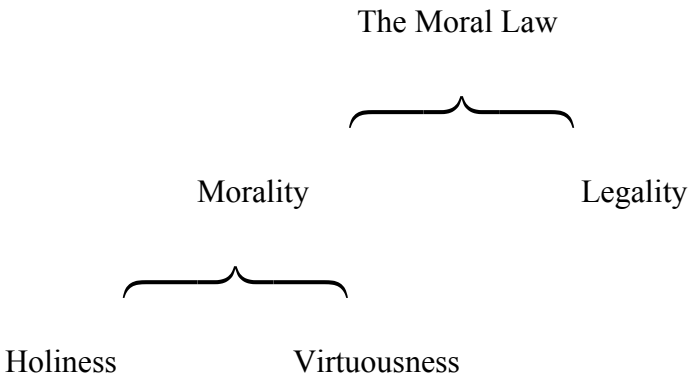
Start with the following distinctions. The moral law, we have seen, is (on Kant's view) to choose one's basic guiding principles on the basis of their fitness to serve as universal laws. And in relation to the particular type of reason-endowed beings that we are, this moral law can also be said to be to act so that we always treat the humanity in each person, ourselves included, as an end in itself, and never as a means only. This allows for a distinction between what Kant calls "morality" and "legality".[309] The former is action directly based on the moral law or, as he also puts it, action out of respect for the moral law. The latter is action merely in accordance with, but not taken on the basis of the moral law.[310]

With regard to morality (in Kant's sense), there is then a further distinction between "holiness" and "virtue".[311] A holy will is a will (or a decision-making faculty) that by its own nature necessarily operates in accordance with the moral law. But given our above-described dual nature as human beings (i.e. as beings who in addition to possessing a will also are animals of a particular sort), holiness is not, Kant thinks, possible for us. What the human being instead is capable of is virtue. And this is the exercise of our capacity to subordinate our pursuit of happiness to the condition that the

[309] (KpV: 5:70-2); (MS: 6:214,225)
[310] This, of course, is a distinction we have already had occasion to discuss above, since some of the main objections that have been raised against the universal law formula also depend on ignoring it.
[311] (G: 4:439); (KpV: 5: 128-9)

maxims we act on that can be willed as universal laws, or, as we could also put it, the active and efficient concern to always at the same time as we are pursuing happiness make sure that we treat the humanity in each person as an end, and never as a means only. Virtue, in other words, is the form of morality (again, as Kant uses the term) that human beings are capable of. As for permissibility, all actions that have the property of legality are permissible. Hence there is no direct inference to be drawn, on Kant's view, from some agent's taking a permissible course of action to the conclusion that the agent, in acting in this way, is thereby being virtuous.

These distinctions can be illustrated as follows:



To repeat: the first division depends on whether actions are performed on the basis of the moral law (morality) or on whether they merely are in accordance with it (legality). The second division on the left depends on whether the will of the agent necessarily and naturally operates on the basis of the moral law (holiness) or on whether this requires continual effort (virtuousness).

Most of the time in his writings Kant is interested in human virtue and thus neither holiness (which isn't possible for humans anyway) nor mere permissibility. But Kant is often understood as concerning himself with either holiness (as if it were a

possible ideal for human beings on his view) or permissibility. And it is on the basis of these interpretative presuppositions that many of the objections to Kant's humanity formula (or Kant's discussion of it) are based. As our case studies of objections to Kant's humanity formula as a formula for determining permissibility (which it *can* indeed also taken to be) that overlook the fact that Kant's *discussion* is mostly concerned with the humanity formula insofar as it relates to human virtue, we will look at some recent objections against the humanity formula offered by T.M. Scanlon and Derek Parfit. To the end of illustrating how not paying sufficient attention to the distinction Kant draws between human virtue and (what is for us impossible) holiness, we will consider one of the most fascinating engagements with Kant's humanity formula in recent literature, namely Rae Langton's discussion of the Kant's brief correspondence with his 22 year old admirer Maria von Herbert.

Although Langton says that her paper is *not* meant as a "cautionary tale about a philosopher's inability to live by his own philosophy"[312], what is intriguing about Langton's paper is to a large extent precisely that it does challenge Kant's own conduct on the basis of Langton's understanding of the humanity formula. Langton's paper also discusses what, using a phrase of Susan Wolf's, Langton understands a "Kantian moral saint" to be like. Langton argues that Maria von Herbert exemplifies a Kantian moral saint, but that her life helps to illustrate that this is *not* an attractive ideal; and also, as just mentioned, that Kant himself fails to act in accordance with the humanity formula in his interaction with von Herbert.

[312] (Langton 1992: 501)

I shall not take it upon myself to defend Kant's conduct[313] (though I will in footnotes point out a certain lack of fairness with regard to at least one aspect of Langton's account of the interaction[314]), but will instead focus on whether Langton is right to think that Kant's view implies that von Herbert is, as Langton puts it, a "Kantian saint." And what I will argue is that the rather tragic state that von Herbert was in before she eventually committed suicide does not, contrary to Langton's argument, fit with the picture Kant's works describe as the moral ideal for human beings[315].

---

[313] …since I am in broad agreement with Langton's moral objection to how Kant, as we shall see, forwarded the letters he had received von Herbert to a third party, using them as an "example of warning". What complicates matters is that I am only in partial agreement (as I will explain below) with Langton as to how to understand Kant's humanity formula. In other words, insofar as I agree with Langton's moral concerns over Kant's forwarding of the letters, this agreement is not based on full agreement about how a moral evaluation of the von Herbert-Kant interaction would be done on the basis of the humanity formula, as I believe Kant intends this formula and its application to be understood.

[314] This has to do with how Langton abridges von Herbert and Kant's respective letters, *leaves out* passages in Kant's letter that speak to von Herbert's request that Kant comment on her decision not to commit suicide, but at the same time claims that Kant doesn't speak on this issue. Langton does point out in a footnote that the letters are heavily abridged, but doesn't indicate where she leaves things out, and some of the things she leaves out speak – at least in my judgment – to some of the concerns that she raises.

[315] And this – i.e. that Kant doesn't see a life like von Herbert's in which one acts in accordance with duty while being indifferent to everything (except for the prospect of each day getting closer to death) – can somewhat ironically be seen to be so in the *un*abridged version of Kant's own letter to von Herbert. That is to say, one of the bits that Langton has edited out says that, "life, insofar as it is cherished for the good that we can do, deserves the highest respect and the greatest solicitude in preserving it and cheerfully using it for good ends" (Kant 1967: 190), whereas Langton presents Kant's view as one on which one is ideally supposed to be without any feeling at all (an ideal that, as we shall see, Kant thinks is impossible and even "harmful" for human beings (R: 6:3)).

The idea of the ideal way of discharging one's duties being doing so in a *cheerful* manner is also brought up in the *Religion*, where Kant – in a striking set of paragraphs that presents what Kant regards as the heart of Jesus' moral teachings on the basis of which Kant thinks he shows that Jesus can justifiably be revered as a great moral teacher (6:159-63) – expresses strong approval of (what he interprets as) the following aspects of Jesus' view. Namely, the parts of these teachings according to which "he sums up all duties … into a *particular* rule, one namely that concerns the human being's external relation to other human beings as universal duty, Love everyone as yourself, i.e. promote his welfare from an unmediated good-will, one not derived from selfish incentives…" and according to which "he wants these works to be performed also in public, as an example for imitation, in an attitude of cheerfulness, not as actions extorted from slaves." (R: 6:161)

To this end we will need to discuss what Kant *does* think of as the moral ideal for human beings. We will also compare this ideal against the interpretations of the humanity formula offered by Langton, O'Neill, and Korsgaard, argue that those readings cannot satisfactorily make sense of the relation between the humanity formula and the highest good (as Kant conceives of it), and, finally, relate the discussion back to where it started in the first chapter. Thus this dissertation will end up not only having related Kant's ethics to the ancient philosophical theme of inner harmony among the different parts of the soul, but also to the classic theme of what it is to flourish as a human being in accordance with the highest possible good.

2. Scanlon and Parfit's Objections to the Humanity Formula as a Test for Permissibility

Our first topic, however, is *permissibility*. Does whether some possible course of action is morally permissible depend on what attitudes or principles the agent performing the action would be acting on? If it can convincingly be argued that it does *not*, and Kant's humanity formula implies that it *does*, then this would ground a seemingly decisive objection against Kant's formula. T.M. Scanlon and Derek Parfit have both recently presented arguments for thinking that permissibility does not depend on the attitudes and policies of the agent. And they both offer interpretations of Kant's humanity formula on which this formula implies that it does. So on the basis of these arguments, Scanlon and Parfit both reject the humanity formula as a criterion of permissibility.

This is not to say that they think the humanity formula fails altogether. Scanlon thinks that whether (in the sense he has in mind) we are treating others as ends rather than

mere means is part of what determines the moral *meaning* of our actions: the significance our ways of conducting ourselves have for what kinds of relationships others can have with us. And Parfit thinks that while it is not always wrong to treat others as mere means (as he understands this idea), it is a moral fault to *regard* others as mere means.

But Scanlon and Parfit's discussions use interpretations of Kant's formula that depend on (a) their own intuitive analysis of what it makes sense to mean by the phrase "treating somebody merely as a means" and (b) the assumption that Kant's own discussion of the humanity formula (especially as it applies to his *Groundwork* example of the lying promise), which they take as their starting point, only concerns the *legality* of our actions. Our own intuitions about what it is reasonable to mean by "treating the humanity in a person as a means only" are, however, not a good guide to what Kant means by this phrase. And what Kant is mostly concerned with in the *Groundwork* is the morality of our actions, not merely their legality. So the approach to understanding Kant's humanity formula and its relation to permissibility that Scanlon and Parfit take is, I believe, not a good one. Their interpretations are, as I will argue below, mistaken, and their objections to the humanity formula do, therefore, fail.

For something by way of context before we get to Scanlon and Parfit's objections[316], let us first consider the remarks Kant makes about the lying promise

[316] For some more context, compare the stated projects that Scanlon, Parfit, and Kant are engaged in their respective books (Scanlon and Parfit are almost exclusively concerned with the *Groundwork*, so by "Kant's book" I here mean the *Groundwork*). Scanlon's book, *Moral Dimensions*, is about different categories used in moral thinking, such as those that Scanlon calls the "permissibility" and the "meaning" of particular courses of action (Scanlon 2008). The former has to do with whether a given course of action is one that a deliberating agent ought to rule out as morally wrong or not. The latter has to do with what kind of impact a person's acting in a certain way on the basis of certain attitudes, aims, or principles would have for what kind of relationships others can have with this agent: whether this person could function as a good friend, business partner, mentor, member of the general moral community etc. Parfit devotes much of his

example in the *Groundwork* when he returns to this example after having just made his initial statement of the humanity formula. Recall that in this example, a person who is in financial need is considering whether or not to try to lend (or, rather, acquire) the money he needs by making a false promise that he will return the money later while he really has no intention of doing so since he knows he won't be able to do so anyway. Commenting on this example, Kant writes:

> …he who has in mind to make such a promise will immediately realize that he has a will to make another human being serve *merely as a means*, without its being possible for [this other human being] to at the same time contain the end within himself. Because he that I have a will to use for my purposes through such a promise cannot possibly agree with my way of treating him and himself contain the purpose of this action. Even clearer does this conflict with the principle of other people appear when one puts forward examples of assaults on the freedom and property of others. Because then it is clearly so that the offender against the rights of human beings is of a mind to make use of the person of another as a means only, without taking into account that as reason-endowed beings, they ought at all times also be assessed as a purpose, i.e. as such a thing, which itself must be able to contain the end of the same action.[317]

Relying heavily on the just-quoted remarks as well as on what it seems to them most plausible to mean by "treating somebody as a means only", Scanlon and Parfit suggest that what makes the difference here are the attitudes, motives, or underlying policies on the basis of which we act. They – i.e. these underlying attitudes etc. - are what determine

---

book, *On What Matters*, to what he regards as following Kant in searching for "the supreme principle of morality". And since Parfit regards Kant as the greatest philosopher since the Ancient Greeks, much of Parfit's book is devoted to investigating whether Kant has already discovered the "supreme principle of morality". (Parfit 2011a: 174) Kant's *Groundwork*, in turn, has as its most basic aim to "seek out and establish the highest principle of morality." And this, of course, makes it sound as if whereas Scanlon and Kant may perhaps be interested in slightly different questions, Parfit and Kant are at least engaged in exactly the same enterprise. Parfit clearly conceives of things in that way. So it seems that the stage is set for the possibility of genuine disagreement. But – alas! – we will see that Parfit and Kant are really talking past each other (or, rather: that Parfit's objection is based on a misunderstanding about what exact project Kant is mainly engaged in).

[317] (G: 4:429-30)

whether we are treating humanity *merely* as a means or not (or whether we at the same time are also treating the humanity in each person as a purpose).[318] As Scanlon puts it, on "any plausible construal, whether in acting a certain way I treated someone merely as a means […] depends at least in part on the reasons I saw as governing my action."[319]

What, more specifically, are Scanlon and Parfit suggesting? On Scanlon's interpretation,

> To see something as a mere means is to see it as something that provides us with reasons only derivatively – that is to say, only insofar as the reasons are provided by something else.[320]

The idea is that treating something as an end, in contrast to a means only, requires also seeing it as an independent or "non-derivative" source of reasons for action. This formulation, Scanlon writes, "puts us in a better position to capture … what Kant had in mind."[321]

Parfit offers us a somewhat more informative interpretation. He writes:

> we treat someone merely as a means if we both use this person in some way and regard her as a mere tool, someone whose well-being and moral claims we ignore, and whom we would treat in whatever way would best achieve our aims. …[And] we do *not* treat someone merely as a means, nor are we even close to doing that, if either (1) our treatment of this person is governed in a sufficiently important way

[318] The interpretative challenge – as regards the mere means-clause of the humanity formula – is to explain what distinguishes treating somebody as a means, which Kant thinks is permissible if we at the same time treat their humanity as a purpose in itself, from treating somebody *merely* as a means (or as a means *only*), which Kant thinks is contrary to morality.
[319] (Scanlon 2008: 89)
[320] (Scanlon 2008: 92)
[321] (Scanlon 2008: 92)

> by some relevant moral belief, or (2) we do or would relevantly choose to bear some great burden for this person's sake.[322]

What makes the difference, then, between treating somebody as a means in a permissible way and *im*permissibly treating the humanity in somebody *merely* as a means are the attitudes we take towards this person, what kinds of reasons we see ourselves as having in relation to her, or what principles actually govern our interaction with her. Or so Scanlon and Parfit claim. And they argue that understood in this way, it quickly becomes clear

---

[322] (Parfit 2011a: 227) And commenting on the suggestion that writers such as John Rawls and Onora O'Neill make, which is that Kant is using the phrase "merely as a means" in some technical or special sense, Parfit writes, in defense of his own reading, that:

> The phrase 'merely as a means' has, I believe, an ordinary sense that is both fairly clear, and morally significant. Though Kant may sometimes use this phrase in a special sense, he also uses it, I believe, in the ordinary sense. It is not misleading to say that, according to Kant's Formula of Humanity, we must never treat people merely as a means [in the sense described above]. And this is the version of Kant's formula that is most worth discussing. (Parfit 2011a: 227)

Rawls and O'Neill, in contrast, think that we should interpret Kant in a way that takes into account his claim that the universal law and humanity formulas are two different ways of stating the same law, and that this requires us to interpret Kant as using "merely as a means" in, as Parfit puts it, some "special sense." Parfit somewhat unconventionally responds to this interpretative suggestion by claiming that Kant's texts are obviously "inconsistent" and that not only that, but that Kant's inconsistency is one of Kant's greatest assets: were it not for his inconsistency, Parfit claims, Kant could not have provided us with so many different and new interesting ideas, which are, Parfit thinks, what makes Kant the greatest moral philosophers since the Ancient Greeks (Parfit 2011a: 183). But given that Parfit's interpretations of these various ideas, which are virtually all based on his own intuitions about what it is plausible to mean by the formulas Kant puts forward, lead him to objections against virtually all the ideas of Kant's he discusses that makes Kant's allegedly inconsistent ideas all seem rather crazy (given the outrageous implications Parfit argues that they have), Parfit's interpretative strategy seems rather questionable.

From the point of view of Parfit's interpretation, in other words, the greatest moral philosopher since the Ancient Greeks presented a whole range of inconsistent ideas that are claimed to be part of a consistent theory, all of which also have highly counter-intuitive implications when we test them against different hypothetical scenarios. From the point of view of interpretative charity, as argued in chapter two, this strongly suggests that it is Parfit's interpretations that we should be doubtful of, and not Kant's claims about the relations between the different formulas that he presents. (For Rawls's and O'Neill's discussions of how to understand the relations among Kant's different formulas, (Rawls 2000) and (O'Neill 1989), and especially "Universal Laws and Ends-in-Themselves" in the latter. There is, I think, much to be agreed with in these more careful readings of Kant's ethics. We'll return to O'Neill's interpretation of the humanity formula towards the end of this chapter.)

that we must reject Kant's formula as a criterion of whether candidate actions are permissible or not, which is the main issue Scanlon and Parfit are concerned with.

Start with Scanlon's argument, which is extremely simple. The argument's first premise is that, as a general matter, whether possible courses of action are impermissible depends on what morally relevant reasons there actually are for or against these courses of action, not on what reasons agents think there are for or against them. The second premise is that the humanity formula, as it is most "plausible" to construe it, takes whether actions are permissible to depend on what their agents see and treat as reasons. It immediately follows, if these premises are correct, that the mere means-clause cannot be right.[323]

Parfit's argument, which is of a similar kind, is also very simple. It is, as I understand it, that we *cannot* make permissible courses of action impermissible by adopting or acting with objectionable attitudes; that the mere means principle implies that we *can* do so; and that the mere means-clause must therefore be a mistake.

Parfit gives his argument through examples, but that is essentially the argument he offers with the help of these examples. He writes

> Consider some gangster who […] regards most other people as a mere means, and who would injure them whenever that would benefit him. When this man buys a cup of coffee, he treats the coffee seller just as he would treat a vending machine. He would steal from the coffee seller if that was worth the trouble, just as he would smash the machine.
>
> …though this gangster treats the coffee seller merely as a means, what is wrong is only his attitude to this person. In buying his cup of coffee, he does not act wrongly.[324]

[323] (Scanlon 2008: 104-5)
[324] (Parfit 2011a: 216)

Parfit also imagines an egoist who saves a drowning child, at a great risk to himself, but whose only aim is to be rewarded. This imagined egoist treats others only in ways that he "believes would be best for him", which also means that when he, for example, "keeps some promise to somebody whose help he will later need, he wants to make use of this other human being, and treats him merely as a means." Commenting on these further examples, Parfit writes

> Since this man treats these other people merely as a means, Kant's principle implies that, in keeping his promise and saving this child's life, this man acts wrongly. That is clearly false.[325]

If Kant's mere means-principle can be so easily refuted, what, according to Scanlon and Parfit, explains its wide appeal? Parfit writes:

> Kant's claim contains an important truth. It is wrong to *regard* anyone merely as a means. But the wrongness of our *acts* never or hardly ever depends on whether we are treating people merely as a means.[326]

Scanlon in turn claims that whether we see others as non-derivative sources of reasons or not bears on the *meaning* of our actions.[327] What kinds of reasons we see ourselves as having in relation to those around us directly bears, that is, on what kinds of relations it makes sense for others to have with us: whether we can sensibly be chosen as friends, business partners, advisors, etc. etc. When it comes to this issue, what reasons we see ourselves as having in our interactions with others is, Scanlon takes it, of crucial

[325] (Parfit 2011a: 216)
[326] (Parfit 2011a: 232)
[327] (Scanlon 2008: 100-5)

importance. But it is *not* essential when it comes to whether actions we might choose are permissible or not.

3. Why Scanlon and Parfit's Objections Fail

Do these objections refute Kant's humanity formula? This depends on whether it really is Kant's view that the *permissibility* of candidate courses of action depends on what attitudes, motives, etc. we would be acting on in performing these actions. The remarks that Kant makes about the lying promise case that we looked at above can make it seem as if it is indeed Kant's view that our attitudes and ways of deliberating determine whether we are acting permissibly or not. But if we were to interpret Kant's remarks about that example in this way, then we would be working under the assumption that what Kant is mainly interested in when he is making these remarks is the *legality* of the way in which the person in his example is conducting himself.

But Kant's main concern in the *Groundwork* is to discuss what he calls the *morality* of our actions. The morality of our actions is a matter of whether they are performed on the basis of, or out of respect for, the moral law. And whether actions are performed on the basis of, or out of respect for, the moral law *is* a matter of how we deliberate and what our attitudes are. Since this – i.e. the "morality" of our actions – is Kant's main concern in the *Groundwork*, we should understand his remarks about his four main examples as *not* only being about the mere *legality* of the actions of the agents in these examples, but also importantly about the *morality* of these actions (or lack thereof).

Scanlon and Parfit's respective discussions (and especially Parfit's discussion) suggest that they understand Kant as only concerning himself with the legality or permissibility of the actions he discusses when he gives his four famous examples. Scanlon and Parfit don't, therefore, distinguish between (a) whether the performance of some action would be *compatible* with treating the humanity in each person as a purpose in itself, and (b) whether, in deciding to act in this way, the agent would deliberate in a way that exemplifies aiming to always treat the humanity in each person as a purpose, and never simply as a means only. But since Kant distinguishes between merely acting in accordance with the moral law (legality) and acting out of respect for, or on the basis of, the moral law (morality), Kant *would*, I believe, draw this distinction. [328]

And when Kant explicitly offers an analysis of *permissibility* (which is the concept Scanlon and Parfit are interested in) in the *Groundwork*, Kant does *not* define permissibility in a way that suggests that he thinks that it is determined by what attitudes, aims, or policies the agent would be acting on the basis of. This is not an analysis that Scanlon and Parfit seem to take note of. But it is an analysis we must take note of in evaluating their arguments against Kant's humanity formula, which crucially depend on the premise that Kant thinks permissibility is partly determined by the agent's attitudes etc. Kant writes:

[328] I say "would" here for no other reason than that Kant never puts his distinction between acting out of, and acting in accordance with, duty directly in terms of whether we are actually treating humanity as an end, or whether our actions conforms to treating humanity as an end, but he does of course already make this distinction; it is just that he never *formulates* it in terms of the humanity formula in particular. That is, every time Kant formulates the distinction between acting from duty (morality) and merely acting in accordance with duty (legality) he does so on a high enough level of abstraction that he does not mention what particular moral principles are at stake.

> The action, which can coexist with the autonomy of the will, is *permitted*; that which doesn't conform to it is *impermissible*.[329]

This analysis does itself, of course, not directly comment on the relation between the humanity formula and its relation to the permissibility of actions. But we already know that Kant thinks of the humanity formula as being one way of stating the constitutive principle of an autonomous human will. And in the *Critique of Practical Reason* Kant *does* directly relate whether we are treating somebody as a means only directly to the condition of whether our action conforms to the autonomy of the will of the person affected by our action (where this affected party might be *either* the agent herself *or* some other agent). Thus Kant writes:

> In the whole of the creation, anything in relation to which one has a will and over which has power may be treated *as a means only*; only the human being and with him every reason-endowed being is *a purpose in itself*. For he is the subject of the moral law, which is sacred, in virtue of the autonomy of his freedom. Precisely for this reason is every will, included one's own self-directed will, constrained under the condition of agreement with the autonomy of the reason-endowed being, namely, to not subject him to any purpose that isn't possible in accordance with a law that could spring out of the affected subject's own will; thus never to treat [this subject] as a means only, but always at the same time as an end.[330]

When we put this together with the *Groundwork*'s analysis of permissibility, what we get is this. Whether we are acting in a way that conforms to at the same time treating somebody as a purpose, and we are thereby acting permissibly, depends on whether our action is possible in accordance with a law that could spring out of the affected party's own will. And this, we have seen in previous chapters, is a matter of whether we are acting in accordance with basic guiding principles that the affected party could also adopt

[329] (G: 4: 439)
[330] (KpV: 5: 87)

for herself, be governed by, and in accordance with which her humanity could be preserved and fully realized. But this all boils down to the following: whether our actions conform to the requirement not to treat the humanity in any person as a means only, but always at the same time as an end, depends on whether our way of acting in relation to this person (which might be ourselves) is compatible with the preservation of this person in her nature as a potentially autonomous human animal, and on whether our action is compatible with the possibility of the full realization of her nature as such a being.

To test this interpretation we can first look at the other three *Groundwork* examples before returning to the lying promise case. Turn first to the suicide example. Regarding this example Kant writes that the human being is not a "thing", but a purpose in itself and that she is, for this reason, not something we can "maim, ruin, or kill."[331] And maiming, ruining, or killing human beings is, of course, incompatible with preserving them as beings with the particular kind of nature that they have. With regard to the example of leaving one's talents and capacities uncultivated, where one could instead have developed these so as to make oneself better equipped to accomplish the ends one might set oneself, Kant explicitly writes that doing so "may be compatible with the *preservation* of humanity as a purpose in itself, but not with its furthering."[332] And with regard

to the example of being helpful to others, Kant similarly writes:

> Now humanity could indeed be preserved if nobody contributed to anybody else's happiness, but there would only be a negative and not a positive agreement with *humanity as a purpose in itself* if everyone didn't also try to promote the ends of others, whatever they might be. Because the ends of the subject, who is a purpose

[331] (G: 4: 429)
[332] (G: 4: 430)

> in itself, must also as far as possible be *my* ends, if this conception is to have its full effect on me.[333]

As these last two examples illustrate, acting in ways that are compatible with the preservation of each person in their human nature is sufficient to qualify as *not* treating anybody as a means only. But also always treating the humanity as an end in a way that positively harmonizes with this idea furthermore requires also acting in ways that allow the human being to flourish in accordance with her own conception of happiness.

Return next to the example of the lying promise, in which Kant claims that making this lying promise would be in "conflict with the principle of [the other]", who couldn't "contain the end of this action within himself." To understand these remarks we should, I believe, read them in light of the above-quoted remarks from the *Critique of Practical Reason*. We should understand these remarks, in other words, as indicating that Kant thinks that being the victim of a lying promise is not compatible with any law that could arise out of the affected party's own will. And this, we have seen, means that being treated in this way is, as Kant thinks of it, incompatible with our being preserved as agents capable of acting in accordance with principles we ourselves adopt and in accordance with which we can both persist and flourish in our nature as beings of the particular kind that we are (again, animals who desire happiness, but who are also capable of fully autonomous agency).

What we find when we read Kant's examples in light of (a) his analysis of permissibility and (b) his various other claims (such as the *Critique of Practical Reason*-claim about what it is to treat humanity as an end in itself) is that whether we are acting

[333] (G: 4: 430)

in a way that conforms to the principle of always treating humanity in each person as a purpose, and never as a means only, is *not* a matter of what our attitudes etc. are in acting as we do. It is instead – as we already saw in the previous chapters – a matter of whether we are acting in ways that are compatible with the preservation and full realization of the particular nature of the human beings with whom we're interacting. But as regards whether we're not merely acting permissibly (i.e. in ways that are compatible with this principle), but we are *also* acting out of respect for this moral law: that *is* a matter of what reasons we are acting on the basis of, what our attitudes are, and what our basic guiding principles and aims are. This is what determines whether our actions also have the property Kant calls *morality* in addition to mere *legality*.

With regard to Scanlon's objection, what this all means is that contrary to what Scanlon is assuming to be the case, the humanity formula does *not* imply that whether a course of action is permissible depends on what reasons the agent doing the acting actually takes there to be for or against the action. Since this is the interpretative premise on which Scanlon's objection rests, it therefore fails.

As regards Parfit's objections, we can first return to the example of the coffee-buying gangster. Concerning the permissibility of this action (evaluated on the basis of the humanity formula), it depends, we can now see, on whether buying this cup of coffee in the given circumstances would be compatible with the preservation of the humanity within the coffee-seller, not on whether the gangster is acting with bad attitudes. And saving the life of a drowning child, or keeping some promise, are, to also return to Parfit's other examples, also examples of actions whose performances are compatible with the preservation of the humanity within these people, no matter how egoistic the

motives that inspire us to do these things might be. So it follows that Parfit is *not* right that Kant's humanity formula would rule out the actions in the cases he discusses as impermissible. And since Parfit's objection hangs on that interpretative premise, it too fails.

To sum up this evaluation of the objections to Kant's humanity formula that Scanlon and Parfit raise against it: these objections are based on Scanlon and Parfit's overlooking Kant's distinction between the legality and the morality of our actions, and also on how they overlook that in discussing his examples, Kant is partly interested in the morality of the actions performed by the agents in these examples, and not just their legality. With regard to whether our actions are *permissible* in accordance with the humanity formula, this itself does not depend on what our attitudes and aims are in acting in these ways, but only on whether acting in these ways is compatible with the preservation and realization of the humanity within all those affected. Since Scanlon and Parfit's objections depend on the interpretative assumption that the humanity formula takes the permissibility of our actions to partly depend on the attitudes, aims, policies, etc. on which we are acting, and that is *not* correct, these objections fail.

4. A Kantian Moral Saint?

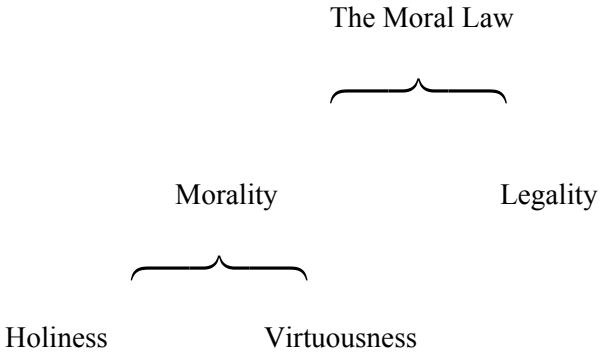
We have now seen that insufficient attentiveness to Kant's distinction between the *morality* and *legality* of our actions can lead to objections against the humanity formula that more careful attentiveness to this distinction shows to be unfounded. I also believe that insufficient attentiveness to Kant's *other* above-mentioned distinction between *holiness* and *virtuousness* – and his accompanying discussions in his major ethical works

of what kind of morality human beings are capable of – can give rise to objections that will be found invalid once we take a closer look at how Kant thinks this distinction relates to the type of *morality* (in Kant's technical sense of the term) that human beings are capable of.[334] To illustrate this, and to evaluate one such objection, I will now discuss Rae Langton's intriguing philosophical engagement with Kant's brief correspondence with Maria von Herbert, a young Austrian woman who participated in an intellectual salon centered around discussion of Kant's philosophy, and who wrote to Kant in a state of despair.

Langton cites from the letters between von Herbert and Kant; applies Kant's moral philosophy (as she interprets it) to Kant and von Herbert themselves; and in the process *both* criticizes Kant the person and his theory *and* highlights features of the theory that she agrees with. (This all, of course, makes for unusually engaging philosophical reading!) In this section I will first cite the letters in the heavily abridged

---

[334] To quickly see what I have in mind here, recall that Kant divides things up in this way:



The first division has to do with whether our actions merely are in conformity with the moral law (legality) or whether we're acting on the basis of, or out of respect for, the moral law (morality). The second division has to do with our wills are so constituted that we necessarily operate in accordance with the moral law (holiness) or whether acting out of respect for the law – as we might do if we always try to make sure to treat the humanity in each person as a purpose in itself – requires continual effort (virtuousness).

versions used by Langton.[335] Then I will review Langton's understandings of (a) the humanity formula and (b) what she calls "Kantian sainthood", and also explain the criticisms Langton offers of Kant himself as well as of Kant's theory based on the example of von Herbert. In the next section I will critically evaluate these arguments, and argue that the life of von Herbert (with her state of indifference to everything, her only wish being to die) does neither match the moral ideal of human life that Kant paints in his writings nor fit with the type of "morality" that Kant thinks human beings are capable of.[336]

Here, in the version Langton uses, is the first letter from von Herbert to Kant:

1: To Kant, From Maria von Herbert, August 1791

> Great Kant,
> As a believer calls to his God, I call to you for help, for comfort, or for counsel to prepare me for death. Your writings prove that there is a future life. But as for this life, I have found nothing, nothing at all that could replace the good I have lost, for I loved somebody who, in my eyes, encompassed within himself all that is worthwhile, so that I lived only for him, everything else was in comparing just rubbish, cheap trinkets. Well, I have offended this person, because of a long drawn out lie, which I have now disclosed to him, though there was nothing unfavorable to my character in it, I had no vice in my life that needed hiding. The lie was enough though, and his love vanished. As an honourable man, he doesn't refuse me friendship. But that inner feeling that once, unbidden, led us to each other, is no more – oh my heart splinters into a thousand pieces! If I hadn't read so much of your work I would certainly have put an end to my life. But the conclusion I had to draw from your theory stops me – it is wrong for me to die because my life is tormented, instead I'm supposed to live because of my being. I've read the metaphysic of morals, and the categorical imperative, and it doesn't

[335] As regards to Langton's fascinating criticism of Kant, a complicating factor – which in fact also complicates the evaluation of Langton's criticism of Kant's *theory* – is, as already noted above, that rather than quoting the letters between von Herbert and Kant in full, Langton uses heavily abridged versions that *leave out* things that bear on the evaluation Langton makes of Kant's reply to von Herbert (and that also make a difference to the overall impression one gets of the two correspondents). (Cf. footnote 315 above.)

[336] And whereas there is much to agree with in Langton's discussion of Kant, there is also much to disagree with, I believe, in how she thinks Kant's own principles apply to the case.

help a bit. My reason abandons me just when I need it. Answer me, I implore you – or you won't be acting in accordance with your own imperative.

My address is Maria Herbert of Klagenfurt, Carinthia, care of the white lead factory, or perhaps you would rather send it via Rheinhold because the mail is more reliable there…[337]

Being "much impressed" by this letter, and having consulted a friend about what to do, Kant eventually responded to von Herbert in the following way (and note that this is a version of Kant's letter that leaves out significant portions of it[338]):

2. To Maria von Herbert, Spring 1792 (Kant's rough draft)

…Your deeply felt letter comes from a heart that must have been created for the sake of virtue and honesty, since it is so receptive to instruction in those qualities. I must do as you ask, namely, put myself in your place, and prescribe for you a moral sedative. I do not know whether your relationship is one of marriage or friendship, but it makes no significant difference. For love, be it for one's spouse or for a friend, presupposes the same mutual esteem for the other's character, without which it is no more than perishable, sensual delusion.

A love like that wants to communicate itself completely, and it expects of its respondent a similar sharing of heart, unweakened by distrustful reticence. That is what the ideal of friendship demands. But there is something in us which puts limits on such frankness, some obstacle to this mutual outpouring of the heart which makes one keep some part of one's thoughts locked within oneself, even when one is most intimate. The sages of old complained of this secret distrust – 'My dear friends, there is no such thing as a friend!'

We can't expect frankness of people, since everyone fears that to reveal himself completely would be to make himself despised by others. But this lack of frankness, this reticence, is still very different from dishonesty. What the honest but reticent man says is true, but not the whole truth. What the dishonest man says is something he knows to be false. Such an assertion is called, in the theory of virtue, a *lie*. It may be harmless, but it is not on that account innocent. It is a serious violation of a duty to oneself; it subverts the dignity of humanity in our own person, and attacks the roots of our thinking. As you see, you have sought counsel from a physician who is no flatterer. I speak for your beloved and present him with arguments that justify his having wavered in his affections for you.

Ask yourself whether you reproach yourself for the imprudence of confessing, or for the immorality intrinsic to the lie. If the former, then you regret having done your duty. And why? Because it has resulted in the loss of your

[337] (Langton 1992: 481-2)
[338] For the full letter Kant wrote to von Herbert, see (Kant 1967: 188-90)

> friend's confidence. This regret is not motivated by anything moral, since it is produced by an awareness not of the act itself, but of its consequences.[339] But if your reproach is grounded in a moral judgment of your behavior, it would be a poor moral physician who would advise you to cast it from your mind.
>
> When your change in attitude has been revealed to your beloved, only time will be needed to quench, little by little, the traces of his justified indignation, and to transform his coldness into a more firmly grounded love. If this doesn't happen, then the earlier warmth of his affection was more physical than moral, and would have disappeared anyway – a misfortune which we often encounter in life, and when we do, must meet with composure. For the value of life, in so far as it consists of the enjoyment we get from people, is vastly overrated.
>
> Here then, my dear friend, you find the customary divisions of a sermon: instruction, penalty and comfort. Devote yourself to the first two; when you have had their effect, comfort will be found by itself…[340]

Here, in turn, is the reply von Herbert eventually sent to Kant's letter:

---

[339] This is the kind of claim that must be read in light of Kant's distinction between the morality and the legality of actions. When Kant here claims that what's moral is related to the act itself, not itself consequences, he means that it is the way the act was decided upon that determines whether the agent was being virtuous, not the consequences of the action. Whether von Herbert's actions, for example, were displays of her virtuousness depends, not on what the consequences of her actions were, but on what lead her to perform these actions. Or so Kant thinks.

[340] (Langton 1992: 482-4) Commenting on this letter, Langton writes, that "[c]onspicuously absent is an acknowledgement of Herbert's more than theoretical interest in the question: is suicide compatible with the moral law?" (Langton 1992: 484) But this is because Langton omits much of the letter, *including* a long a sentence that directly comments on the value of "life, insofar as it is cherished for the good that we do." For Kant immediately follows his above-quoted claim that "the value of life, in so far as it consists of the enjoyment we get from people, is vastly overrated" with the further claim that "life, insofar as it is cherished for the good that we can do, deserves the highest respect and the greatest solicitude in preserving it and cheerfully using it for good ends." (Kant 1967: 190) And just a few lines earlier in the letter – i.e. in the longer, actual version – Kant has also just said that

> to brood over one's remorse and then, when one has already caught on to a different set of attitudes, to make one's whole life useless by continuous self-reproach on account of something that happened once upon a time and cannot be anymore – that would be a fantastic notion of deserved self-torture (assuming that one is sure of having reformed). It would be like many so-called religious remedies that are supposed to consist in seeking the favor of higher powers without one's even having to become a better human being. That sort of thing cannot be credited in way to one's moral account. (Kant 1967: 190)

Rather than ignoring von Herbert's "more than theoretical interest" in the issue of suicide, then, what Kant actually does is to not only tell her that one must move on and not continue beating oneself up after one has reformed oneself. He also tells her that *the good we can do in life gives us reason to cherish life, and that a life in which we do good deserves the highest respect and the greatest solicitude in preserving it and cheerfully using it for good ends*.

3. To Kant, from Maria von Herbert, January 1793

Dear and reverend sir,
Your kindness, and your exact understanding of the human heart, encourage me to describe to you, unshrinkingly, the further progress of my soul. The lie was no cloaking of a vice, but a sin of keeping something back out of consideration for the friendship (still veiled by love) that existed then. There was a struggle, I was aware of the honesty friendship demands, and at the same time I could foresee the terribly wounding consequences. Finally I had the strength and revealed the truth to my friend, but so late – and when I told him, the stone in my heart was gone, but his love was torn away in exchange. My friend hardened in his coldness, just as you said in your letter. But then afterwards he changed towards me, and offered me again the most intimate friendship. I'm glad enough about it, for his sake – but I'm not really content, because it's just amusement, it doesn't have any point.

My vision is clear now. I feel that a vast emptiness extends inside me, and all around me – so that I almost find myself to be superfluous, unnecessary. Nothing attracts me. I'm tormented by a boredom that makes life intolerable. Don't think me arrogant for saying this, but the demands of morality are too easy for me. I would eagerly do twice as much as they command. They only get their prestige from the attractiveness of sin, and it costs me almost no effort to resist that.

I comfort myself with the thought that, since the practice of morality is so bound up with sensuality, it can only count for this world. I can hope that the afterlife won't be another life ruled by these few, easy demands of morality, another empty and vegetating life. Experience wants to take me to task for this bad temper I have against life by showing me that nearly everyone finds his life ending much too soon, everyone is so glad to be alive. So as not to be a queer expectation to the rule, I shall tell you of a remote cause of my deviation, namely my chronic poor health, which dates from the time I first wrote you. I don't study the natural sciences or the arts anymore, since I don't feel I'm genius enough to extend them; and for myself, there's no need to know them. I'm indifferent to everything that doesn't bear on the categorical imperative, and my transcendental consciousness – although I'm all done with those thoughts too.

You can see, perhaps, why I only want one thing, namely to shorten this pointless life, a life which I am convinced will get neither better nor worse. If you consider that I am still young and that each day interests me only to the extent that it brings me closer to death, you can judge what a great benefactor you would be if you were to examine this question closely. I ask you, because my conception of morality is silent here, whereas it speaks decisively on all other matters. And if you cannot give me the answer I seek, I beg you to give me something that will get this intolerable emptiness out of my soul. Then I might become a useful part of nature, and, if my health permits, would make a trip to Königsberg in a few years. I want to ask permission, in advance, to visit you. You must tell me your story then, because I would like to know what kind of life your philosophy has led you to – whether it never seemed to you to be worth to bother to marry, or to give

> your whole heart to anyone, or to reproduce your likeness. I have an engraved portrait of you by Bause, from Leipzig. I see a profound calm there, a moral depth – but not the astuteness of which the *Critique of Pure Reason* is proof. And I'm dissatisfied not to be able to look you right in the face.
>
> Please fulfill my wish, if it's not too inconvenient. And I need to remind you: if you do me this great favour and take the trouble to answer, please focus on specific details, not on general points, which I understand, and already understood back when I happily studied your works at the side of my friend. You would like him, I'm sure. He is honest, goodhearted, and intelligent – and besides that, fortunate enough to fit this world.
>
> I am, with deepest respect and truth, Maria Herbert.[341]

About the time Kant received this letter from von Herbert he also received another letter from their mutual acquaintance, J.B. Erhard, who says that von Herbert has "capsized on the reef of romantic love." Erhard further writes that in order to "realize an idealistic love", von Herbert had given herself to "a man who misused her trust." And then, "trying to achieve such a love with another, she told her new lover about the previous one." That, Erhard concludes, "is the key to her letter." Von Herbert is not, Erhard adds, in a state in which she is receptive to moral instruction, but has lost all "delicacy" and instead entered into a kind of fantasy world.[342] Trusting Erhard's judgment, Kant decides not to respond to von Herbert's second letter. He instead passes von Herbert's two letters as well as Erhard's letter on to the young daughter of a friend as an "example of warning" of "sublimated fantasy."[343]

Now first responding earnestly to von Herbert's letter in a way that engages with it philosophically and then not responding to the second letter, but instead passing it on to

---

[341] (Langton 1992: 493-4)
[342] (Kant 1967: 203-4)
[343] (Kant 1967: 204) Not ever learning about this, but having heard around that time (or earlier) from Erhard that Kant had asked about her, Von Herbert wrote Kant one more time in 1794, this time expressing her admiration for the *Religion Within the Boundaries of Reason Alone*, and once more mentioning her wish to eventually visit Kant. In 1803 she eventually committed suicide. Kant himself died the year after of old age.

a third party as an example of warning constitutes a shift on Kant's part, Langton argues, between treating von Herbert as a purpose in itself and then – in conflict with Kant's own humanity formula – treating von Herbert as a mere means. And it also, secondly, shows, Langton further argues, that, "Kant is unaware that he has received a letter from a Kantian saint." "Indeed, it is," Langton continues, "hard to believe that he has read her second letter,"[344] since, Langton argues, it so clearly is written by somebody who so clearly exemplifies Kant's own ideal for a moral saint.

Here, first, is what Langton means with her first criticism. She writes:

> The important thing is not to treat a person *merely* as a means. I treat a person merely as a means when I act towards her in a way that blocks her ability to form her own ends and act on them. I do this when I make it impossible for her to assent to my action towards her, impossible for her to share the goal I have in acting. I treat her as a means when I act in a way that prevents her from choosing whether to contribute to the realization of my end. This is a violation of duty, in Kant's terms.[345]

Treating somebody as an end, in contrast, most importantly amounts to relating to her in the sort of way that P.F. Strawson describes in his influential paper[346] that argues that holding people responsible involves interacting with them on the basis of "involved" social emotions, thereby taking up an "interactive" standpoint in relation to them. Langton writes:

> This standpoint is manifested in more than our attitudes … [and] shows up in the way we act when we communicate and co-operate with others. …you are prepared to work with them, view them as someone who has goals of their own that might come to share, or as someone who might come to share your goals. […] The standpoint we take towards human being is [thus] interactive, and it is different from the standpoint we take with things. Kant thinks this is because

[344] (Langton 1992: 499)
[345] (Langton 1992: 489)
[346] "Freedom and Resentment", which is reprinted in (Strawson 1971)

> human beings have an intrinsic worth that has its basis in our capacity for rational choice. Human beings are ends in themselves, who have a dignity… The moral law is the requirement to recognize and respect this dignity, and to act in a way consistent with it.[347]

This all means, Langton thinks, that when Kant responded to von Herbert's first letter, and tried to apply his philosophical theory to her situation, explaining how he thought it relates to von Herbert's relationship with her friend, Kant was indeed initially taking up such an "interactive standpoint" in relation to von Herbert, thus treating her as an end in itself. He thereby acted, Langton thinks, in accordance with the humanity formula. But when Kant failed to reply to her second letter, and instead passed it on to a third party as an "example of warning", Kant did no longer retain this interactive standpoint towards von Herbert. He instead acted in a way that made it impossible for von Herbert to share his goal: he didn't ask for her permission to pass on her letters, he in effect treated her as somebody who was not fully responsible for her own actions, and therefore "blocked" von Herbert's agency, thus treating her as a mere means. Kant thereby acted, Langton concludes, in a way that conflicted with the humanity formula.

This first criticism of Kant – that he himself failed to live up to his own moral standard, as Langton understands it – is based on a part of Kant's theory that Langton *agrees* with: the part that, as Langton understands the theory, says to always interact with people (a) in ways that involve taking up the "interactive standpoint" and (b) that doesn't block their agency, by making it impossible for them to share our goals. I shall return to Langton's reading of the humanity formula using the framework of Strawson's influential theory of responsibility-attributions below, but am even more interested, for our present purposes, in Langton's second major criticism of Kant (and this time Kant's theory). And

[347] (Langton 1992: 487)

this is criticism that involves (1) identifying von Herbert, as Langton does (but she thinks Kant fails to do), as an example of a Kantian moral saint, and (2) arguing, on the basis of the example of von Herbert, that this surely isn't a very attractive ideal. The "example of warning" that Langton uses von Herbert as, in other words, is not for young women in danger of falling prey to "sublimated fantasy", but for philosophers who might be tempted to accept Kant's philosophical theory of the morally ideal human being.

So what is Langton's basis for identifying von Herbert as an example – and a rather frightening one at that – of a Kantian moral saint? Why is it that "her life constitutes a profound challenge" to Kant's moral philosophy? It has to do with how Kant's substantive theory of human virtue contains what he calls a "duty of apathy".[348] Thus Kant writes (drawing on the Stoics and their ideal of *apatheia*[349]):

---

[348] "Apathy", Kant is quick to point out, is *not* to be understood as indifference. For indifference, he thinks, is *in*compatible with our duty to make the humanity in each person into the most general purpose around which all our actions are based. Indifference to human beings, Kant writes, is incompatible with treating humanity as a purpose in itself. Thus he writes: "The highest principle of the teaching of virtue is this: act in accordance with a maxim of *purposes* the having of which can be a universal law for everyone. – In accordance with this principle the human being is a purpose for himself as well as for others and it is not enough that he is not permitted to treat either himself or others as a means only (whereby he can also be indifferent to [himself and to others]), but to make the human being into an end is in itself a duty of human beings." (MS: 6:395) Rather than being indifferent, we are instead under a requirement of virtue, he argues, to cultivate our natural "love of human beings" and "self-respect", which are "moral feelings" all human beings have a predisposition towards, but that we may need to cultivate in order for them not to be overshadowed by other "affects" and "passions" (such as anger and hatred) that can lead us to vice. (MS: 6: 399-402)

[349] Anybody familiar with the Roman Stoics will see that the passage in Kant we're about to look at could almost have been directly lifted from their writings, such as Seneca's writings on anger. These Stoics were, of course, in turn under the influence of Plato, and we here once more see the aptness of Korsgaard's comparison between Kant's moral philosophy and the theory of the virtue of justice that Plato puts forward in the *Republic*. Phrases such as "reason's taking up the reins controlling us" seem lifted straight out of Plato, whereas vocabulary such as "apathy" and "affects" (used in the way we're about to see that Kant uses it) is lifted out of the way many Stoics – again, such as Seneca who, writing in Latin, used *adfectus* for what Kant seems to mean by "affects" (*Affekte*) – put their doctrine. That Kant puts his theory forward as a "teaching of virtue" of course also has to do with the great influence Ancient Greek ethics has on his thinking. Indeed, in the *Religion* he even writes that, "the ancient moral philosophers … have pretty well

> …the teaching of virtue [is based on] the principle of inner freedom. […] For inner freedom, however, two things are required: to be one's own *master*… and to *rule* over oneself… i.e. so *tame* one's affects and to *control* one's passions. […] *Affects* and *passions* are essentially different from each other; the former belongs to feeling, insofar as these …make deliberation impossible or more difficult. These affects are therefore said to be *abrupt* … and reason says, through the concept of virtue, that one should *get a hold* of oneself… [Such a] storm [usually] blows over quickly. A tendency towards affect (e.g. *anger*) is therefore not as closely related to vice as passion is. *Passion* is in contrast the sensual desire that turns into a lingering inclination (e.g. *hatred* in contrast with anger)… Virtue thus, insofar as it is based on inner freedom, contains, for the human being, also a positive command, namely to bring all one's capacities and inclinations under one's rule (of reason), hence the mastery over oneself, which is added to the prohibition against allowing oneself to be ruled by one's feelings and inclinations (the duty of *apathy*); because only reason takes the reins of the ruler into its hands, then these [i.e. the affects and passions] will play the role of the master of the human being.[350]

Now in addition to this *duty of apathy*, which Kant thinks constitutes the cornerstone of the teaching of virtue, Langton also notes that:

> Something rather similar to apathy is [also] described in the *Critique of Practical Reason*, but this time it is not called apathy, but 'bliss' (*Seligkeit*). Bliss is the state of 'complete independence from inclinations and desires'. […] Bliss is 'the self-sufficiency which can be ascribed only to the Supreme Being'. The Supreme Being [unlike the human being] has no passions and inclinations. […] God is the being more apathetic than which none can be conceived.[351]

Kant claims, then, that what he calls "moral apathy" is an ideal for human beings, but that we associate an even higher state, "bliss", with the Supreme Being, a state of complete independence of which human beings are not capable. (It might be useful in this context to also recall that Kant thinks that our very idea of a Supreme Being in the first place

---

exhausted nearly all that can be said concerning virtue." (Footnote at R: 6:25) These are influences we need to keep in mind when we try to understand Kant's ethical theory.

[350] (MS: 6: 407-8)

[351] (Langton 1992: 497)

derives from our conception of morality, and that this idea is the idea of the highest conceivable moral perfection.) Is Kant right about this? Langton writes:

> What of Kant's moral patient? She is well beyond the virtue of apathy that goes with mastery of the inclinations. She has no inclinations left to master. She respects the moral law, and obeys it. But she needs not battle her passions to do so. She has no passions. She is empty – but for the clear vision of the moral law and the unshrinking obedience to it. She is well on the way to bliss, lucky woman, and, if Kant is right about bliss, well on the way to the Godhead. No wonder she feels that she – unlike her unnamed friend – does not quite 'fit the world'. She obeys the moral law in her day to day dealings with people from the motive of duty alone. She has no other motives. She is no heretic. She is a Kantian saint.[352]

Langton argues, then, that Kant describes an ideal for human beings ("apathy") that von Herbert easily achieves. And, in addition to that, Kant also describes an even higher ideal ("bliss") that only a Supreme Being is said to be capable of, but which nevertheless is another ideal, Langton claims, that von Herbert also lives up. But are these good ideals for a moral theory to be centered around? Is Kantian sainthood, as Langton understands this idea, an attractive state to strive for? Langton thinks not. As she herself puts it, commenting on von Herbert's state: "Oh brave new world, that has such moral saints in it." Von Herbert's life, Langton concludes, "constitutes a profound challenge" to Kant's moral philosophy.

Von Herbert is, Langton thus takes it, so obviously a Kantian saint that it is "hard to believe that Kant read her second letter." Kant himself, though, did not, as Langton also notes, recognize von Herbert as a Kantian saint. Ought he, out of consistency with his own theory, to have done so? Or is it rather that von Herbert, contrary to Langton's argument, does *not* actually constitute an example of the moral ideal for human beings around which Kant's moral theory is built?

[352] (Langton 1992: 497)

5. Human Flourishing and the Highest Good

Have we come this far only to find that the person who best approximates the moral ideal the Kantian theory puts forward is somebody who is so depressed that nothing no longer attracts her? The just-reviewed argument from Langton would, if successful, indicate that such is the case, and our story would have a rather unhappy and disappointing ending. Langton's argument is not, however, sound. Or so I shall argue in this section.

For Langton's argument against Kant's theory to work, it would need to be the case that (a) it involves a correct understanding of Kant's "duty of apathy" and the moral ideal for human beings that Kant's theory as a whole encompasses, and (b) that somebody like Maria von Herbert lives up to this ideal, constituting what would be an example of human moral flourishing from a Kantian point of view. Neither condition, however, holds. Langton understands "apathy" in the ordinary language sense of a state of total indifference, but that's not what Kant means by it. Langton also seems to overlook the moral ideal for human beings that Kant describes throughout his writings, because von Herbert does not at all match that ideal.

Start with the duty of apathy and why Langton thinks von Herbert is "well beyond the virtue of apathy." Commenting on von Herbert's inner state, as she herself describes it in her second letter to Kant, Langton writes:

> The passion, the turbulence, has vanished. Desolation has taken its place, a 'vast emptiness', a vision of the world and the self that is chilling in its clarity, chilling in its nihilism. Apathy reigns. Desire is dead. Nothing attracts.[353]

[353] (Langton 1992: 494)

Is *that* the type of apathy – i.e. apathy in the ordinary contemporary sense of the term – that Kant is talking about when he declares there to be a duty of apathy? No. As Kant himself writes (and this is interestingly enough even a passage Langton herself quotes):

> The word "apathy" has fallen into disrepute, as if it meant lack of feeling and so subjective indifference regarding objects of choice: it has been taken for weakness. We can prevent this misunderstanding by giving the name "*moral apathy*" to that freedom from agitation *which is to be distinguished from indifference*, for in it the feelings arising from sensuous impressions lose their influence on moral feeling only because respect for the law prevails over all such feelings.[354]

The kind of moral apathy Kant is talking about (and, again, his use of the word "apathy" of course comes from the old Stoic term ἀπάθεια (*apatheia*), which is why he talks about how the term "has fallen into disrepute") is *not* a state of indifference, but is rather "to be distinguished from indifference." And the kind of apathy that Langton thinks von Herbert suffers from is precisely a state of indifference. As she puts it, "Desire is dead." "Nothing attracts."

The kind of moral apathy that the ideally virtuous person enjoys is also different from the state of emotional apathy von Herbert suffers in yet another respect. Kant writes (and this is also from the section describing what he means by "moral apathy"):

> The true strength of virtue is the *peaceful mind*, with the deliberate and firm resolution to put its law into practice.[355] That is the state of *soundness* in moral life…[356]

[354] (MS: 6: 408-9, second emphasis added)
[355] Different Ancient schools had different ideas about how to achieve *ataraxia* (or freedom from disturbances), and Kant's suggestion is that steadfastness in the following in the laws of virtue

A mind like von Herbert's, which is "tormented by a boredom that makes life intolerable," a mind which is possessed of only one morbid desire – namely, "to shorten this pointless life" – is surely *not* a peaceful mind.[357]

But what about that sentence in the *Groundwork*, which Langton puts forward as evidence for her interpretation, and that goes like this: "The inclinations themselves, however, as sources of needs, possess so little absolute value on account of which one could have a wish for them, that to it must be a universal wish among reason-endowed beings to be completely free from them"? Doesn't the fact that Kant makes that claim suggest that Langton does after all understand the Kant's notion of apathy in the right way?

No. Notice that Kant chooses his words carefully here. He says that insofar as they are sources of needs, our inclinations have so little absolute value etc. that it must be

---

helps to achieve this kind of inner peace (in two senses of the term "peace", since he also thinks such an inner rule of law ends the war of all against all that otherwise rages within our minds).

[356] (MS: 6: 409)

[357] We can note here also that when Kant gives his advice von Herbert to focus on virtue and that consolation (i.e. the return of happiness) will ensue in time, this is also based on his overall theory, which, as we see in the *Metaphysics of Morals*, involves the idea that by practicing virtue, we thereby enable a love of human beings to grow within us: a love we are capable of by nature, but which we may need to cultivate. And the way to cultivate our internal love of human beings is to practice practical love, from which the feeling of love will be strengthened.

Langton mentions these ideas but sets them aside directly because she thinks they are "hard to reconcile" with Kant's claims about moral worth. And to make her case Langton brings up Kant's discussion following his claim that the moral worth of an action doesn't depend on the feeling that inspired it, but the principle of the will the agent is operating in accordance with, at the beginning of the *Groundwork*. What Langton, and so many others, miss about why Kant sets aside cases of people that happily do their duty, is that he wants to analyze the concept of acting from duty (since he thinks this will help him to analyze the concept of a good will), and that he thinks that it is easiest to do so if we focus on cases where people are clearly not acting for any other reason than a sense of duty, or, as he also puts it, out of respect for the law. This does not mean, as Langton seems to think, that he thinks that we should try to get rid of all emotions, and that our emotions are bad. As we have seen, he instead thinks that to do so would be "futile", "harmful", and "reprehensible." (R: 6: 58)

a universal "wish" to be completely free from them. Given that we are animals that by nature have certain inclinations it would, however, be a very bad idea, Kant thinks, to *try* to get rid of our natural inclinations.[358] Thus he writes (this time in the *Religion*):

> Natural inclinations are, *considered in themselves, good*, i.e. unobjectionable, and to have a will to exterminate them would not only be futile, but also harmful and reprehensible; one must rather only tame them, so that they will *not* wear each other out, but instead can be brought to harmonize into a whole, called happiness.[359]

In relation to *this* aspect of our human nature, then, human flourishing simply consists in happiness (where happiness is conceived of as the achievement of a state where we can be content with our lives since everything goes in accordance with our desires and wishes). This is why making the humanity in each person into an end of ours, on Kant's view, includes making the happiness of each person into an end of ours. And it is also why what Kant calls "the highest good" is *not* exhausted by universal virtuousness.[360]

Now if Maria von Herbert, as Langton claims, would exemplify the idea of a Kantian moral saint, and this is to serve as a basis of an objection to Kant's theory, then it would need to be true that von Herbert could be understood as flourishing completely from the point of view of what Kant calls the highest good: the best possible state of affairs imaginable from the point of view of the basic principles of practical reason. To

---

[358] One could rationally *wish* that one wasn't subject to all the needs one is subject to while at the same time realizing that actually trying to exterminate the inclinations upon which one's needs are founded would be futile, for which reason the wise way of relating to these desires instead is to try to tame one's inclinations and to bring them into harmonious whole. Only then can one enjoy the kind of inner peace that is associated with human virtue.

[359] (R: 6:58)

[360] With regard to our capacity for inner freedom – i.e. the *other* main aspect of our dual nature – human flourishing consists in the inner rule of law that we can only achieve by subjecting ourselves to basic guiding principles that are to fit to serve as universal laws for the preservation and full realization of our own nature.

explain why von Herbert did *not* flourish as a human being from the point of view of *the highest good*, as it is understood within Kant's theory, we can first explain what exactly Kant means by this phrase, and we will let Kant himself do so.

Here's how he puts things in the *Critique of Practical Reason*:

> The *good* or *evil* always refers to … a relation to the *will* insofar as it is determined by a law of reason to make something into its object[361]… The only objects of a practical reason are thus *the good* and *the evil*. Because by the former one understands a necessary object of the faculty of desire, by the second of the faculty aversion, both, however, in accordance with [or after[362]] a principle of reason…[363]
>
> The concept of the *highest* contains an ambiguity… The highest can mean the *supreme* (supremum) or also the *complete* (consummatum). The former is the sole condition that alone is unconditioned, i.e. subordinated to nothing else; the latter the whole, which is not a part of anything else of the same kind (perfectissimum). That *virtue* (as the worthiness to be happy) is the *supreme condition* of all that appears desirable to us [in accordance with principles of reason], hence including also our pursuit of happiness, and therefore is the *supreme good,* has been already proven… Therefore it is not yet, however, the whole and complete good, as the object of the faculty of desire of reason-endowed beings; because for that *happiness* is also required… […] Now insofar as virtue and happiness constitutes the highest good in a person, and happiness distributed in precise proportion to virtuousness (as the value of the person and her worthiness to be happy) thereby constitutes a possible world's *highest good*, it is also the whole, and complete good.[364]

---

[361] (KpV: 5: 60)

[362] I add the "after" here because the German word for "in accordance with" or "on the basis of" – i.e. "nach" – also means after, and thereby makes it clearer than its English translations that it is not the case that we first recognize something and good and then, on that basis, see that a principle of reason instructs us to make it the object of our will. It is rather that a principle of reason's instructing us to make something the object of our will or that something is an object of our will in accordance with a principle of reason that makes something good, or in which its being good (in this sense) consists in. In giving their reconstructions for Kant's argument for the humanity formula that we looked at in the third chapter, Korsgaard and Wood seem, as we saw, to overlook this. That is a mistake, and Kant makes it very clear that he means no such thing. Thus he writes, for example, that, "*the concepts of good and evil are not determined before the moral law ... but must instead be determined after and through it*." (KpV: 5: 62-3)Emphasis in the original.)

[363] (KpV: 5: 58)

[364] (KpV: 5: 110-1)

The highest good of a particular person (in the sense of the whole and complete good), then, is the whole and complete object of a self-directed will that is wholly determined by the laws of practical reason. Insofar as we think of ourselves as animals subject to various inclinations and needs, the object of a self-directed will governed by practical reason would, as we have seen above, be happiness, whereas insofar as we are beings capable of self-mastery the object of a self-directed will wholly governed by practical reason would be virtue: i.e. the inner rule of law through self-adopted basic guiding principles fit to serve as universal laws for all beings like us. Thus the highest (as in the whole and complete) good of a particular person is both happiness and virtue.[365]

A will directed at the world as a whole would, similarly, (if it were wholly determined by the constitutive laws of practical reason), have as its object that all the reason-endowed inhabitants of the world would get to enjoy all the happiness they desire that is compatible with their maximal virtuousness: that everyone would get to enjoy all the happiness they would also be worthy of if they were as virtuous as they can be. The highest conceivable good of a possible world as a whole would, therefore, be universal virtue coupled with each inhabitant's enjoying all the happiness she would desire and her virtue would make her worthy of.[366] Or so Kant argues.[367]

---

[365] … the combination of which is only possible, Kant thinks, if we subordinate our pursuit of happiness to the condition that we always act in according with and on the basis of self-adopted basic maxims that are fit to serve as universal laws in accordance with which all human beings can fully realize their nature as the particular kind of reason-endowed beings they are. This is why virtue, and not happiness, is the *supreme* good. But as Kant also writes in the beginning of the *Groundwork,* virtue (or a good will) is not the whole and complete human good on account of its being the supreme good: we only get the whole and complete good when we add happiness to virtue, the supreme good (in the sense explained by Kant in the above-quoted passage).

[366] Note that it would *not* be a matter of happiness aggregation of a utilitarian sort, whereby the total possible amount of aggregative happiness would be added to universal happiness. Nay, because in such a utility-maximizing world there might be any number of agents who would not get to enjoy the greatest possible amount of happiness she would desire and also be worthy of that's compatible with each agent's getting to enjoy all the happiness they desire and also are

Now what is most important for the purposes of evaluating Langton's "profound challenge" to Kant's theory on the basis of the example of the life of Maria von Herbert is, of course, that the highest good for an individual human being, in accordance with Kant's theory, is virtue *and* happiness. So if von Herbert truly were flourishing in accordance with the highest good that is possible for a human being as it is conceived of within Kant's theory, she would also be fortunate enough to enjoy happiness. But very sadly, that was not the case, since von Herbert was deeply unhappy, and so Maria von Herbert's very tragic life does *not* match up with the moral ideal of a human life that is at the heart of Kant's ethics.

Return next to the concept of *bliss* (Seligkeit) that Langton also discusses. Rather than functioning as a counterexample to Kant's claim that this state is not possible for human beings by being a human being that actually enjoys bliss, von Herbert (again, very

---

worthy of. The total amount of happiness (insofar as it all makes sense to think of happiness as something we can aggregate, which I doubt) could thus conceivably be greater in a world in which each agent didn't get all the happiness she desires and is worthy of than in a world in which this condition is fulfilled. Since virtue (and not happiness) is the supreme good, such a world could not possibly, on the Kantian view, be preferable to a reason-governed will directed at this world in comparison to an alternative possibility in which each agent did get to enjoy all the happiness she desires and is also worthy of that can possibly be combined with each other agent's also getting enjoy the maximal amount of happiness she desires and is worthy of. The idea of universal virtue and the greatest possible amount of happiness for each in proportion as they are worthy thereof, in other words, does not contain any utilitarian element (insofar as utilitarianism is understood as necessarily involving the aggregation of utility and happiness is considered as the sole measure of utility). For this reason it is a mistake to claim, as some writers have, that Kant's view can be understood as being a type of (or at least as containing a type of) utilitarianism.

But for contrary assessments, which I disagree with, see "Could Kant have been a Utilitarian?" in (Hare 1997); (Cummiskey 1996); and (Parfit 2011a). In addition to the fact that the Kantian view doesn't allow for utility-aggregation, Kant's view also doesn't involve the kind of supposedly objective non-moral values that Parfit's Kantian rule-consequentialism postulates. There are thus several reasons why a Kantian couldn't accept Parfit's theory.

[367] We don't have to concern ourselves here with whether this is a good argument. Our question is at the moment only whether Kant's theory implies that a person like Maria von Herbert exemplifies the ideal human life from the point of view of Kant's moral theory, and as we're seeing, that is not, contrary to what Langton argues, so.

sadly) instead serves as an example of how bliss is *not* possible for human beings. The type of apathy (in the normal sense of the term) that von Herbert suffered, which involves not being attracted to anything, leads a human being to a state in which she cannot function properly, and in which she may instead, as was also the case here, desire to kill herself. What Kant means by "bliss", in contrast, is a state of ultimate peace within oneself (the kind of freedom from all disturbances or complete *ataraxia* that members of some of the Ancient schools of philosophy tried to achieve[368]). Von Herbert was not enjoying such a state. In even supposing that von Herbert might have achieved holiness (the state of having a will that necessarily confirms to the moral law, since it has no contrary inclinations) Langton thus assumes that von Herbert would be capable of a kind of functioning that is not possible for human beings, as von Herbert herself sadly helped to illustrate.[369]

Can it not have been the case, however, that von Herbert exemplified a very high degree of virtuousness, and thereby accomplished the supreme good of a human being, namely, the possession of a good will? Could she not in that sense have been a Kantian moral saint? That is to say, a person who refuses to act other than on the basis of maxims

[368] Neither the German word Kant uses ("Seligkeit") nor the English translation Langton uses ("bliss") refers to anything even approximating the kind of mental state von Herbert describes herself as being in, so it is rather strange, just on the basis of the terminology Kant uses, why Langton would think that he would regard von Herbert as somebody enjoying bliss.

[369] Langton, of course, doesn't only get the vocabulary of moral sainthood from Susan Wolf, but also from Kant's own expression moral holiness. But, as Kant repeatedly points out, what he means by holiness is impossible in a being that, in addition to having a will (i.e. practical reason), also is subject to desires, needs, hopes, and fears, etc. etc. What such beings are capable of is virtue (in Kant's sense of governance by maxims of virtue), but not holiness (which can only be achieved by a being without any human needs).

What is confusing here, of course, is that Kant thinks that we should nevertheless "strive" to approximate the ideal of holiness, but he is careful to always point out that actually achieving holiness is beyond all human capacities. (R: 6:159) This is a reason why Kant thinks Christianity (as he interprets its ethical teaching) is an improvement on Stoicism, since the latter, but not the former doctrine, thinks that we can achieve holiness (in the sense Kant has in mind). Cf. (KpV: 5: 128); (R: 6:161)

that could serve as universal laws, and who, therefore, is worthy of happiness in Kant's intended sense. Here I think the answer could possibly be yes.

We don't quite know exactly how von Herbert understands the categorical imperative (other than that she thinks that it requires of her that she doesn't commit suicide), so we cannot be sure whether she was operating in accordance with maxims that could serve as universal laws in the sense Kant intends. But we can suppose that – although she was not able to achieve the kind of inner peace of mind Kant thought that virtue would lead to – von Herbert was indeed always and only acting on the basis of maxims that could serve as universal laws. Does this supposition turn her back into a counterexample to Kant's theory?

No. The apparent force of Langton's argument is based on how, as she describes Kant's theory, von Herbert allegedly is an example of a human being that flourishes completely in accordance with Kant's moral theory. The force of the argument does not, in other words, depend on whether it is possible that von Herbert might have only subjected herself to maxims that could serve as universal laws, which we are now also supposing that she did indeed do. What Langton tries to do is, in other words, to show that von Herbert's life has all the properties that would make it an ideal human life from the point of view of Kant's moral theory. And we have seen that von Herbert's life is very far from an ideal human life from the point of view of Kant's theory. So, we can conclude, von Herbert's life does not constitute a profound challenge to Kant's theory. Her life was instead, from the point of view of Kant's moral theory (as well as from any reasonable point of view), a highly tragic life since she got to enjoy so little (if any) of the

happiness of which she was worthy. Langton's objection based on the life of Maria von Herbert, therefore, fails.

6. Human Flourishing and the Highest Good, Continued

We can now briefly return to Langton's criticism of Kant himself, or, rather, to the interpretation of the humanity formula that Langton uses to criticize Kant. Is this a correct interpretation of Kant's formula? And can we use this reading to make sense of Kant's views about (a) what is involved in treating the humanity in each person as purpose in itself and (b) what the highest good is constituted by? I think not.

On Langton's reading, treating the humanity in each person as a purpose in itself consists in always interacting with those around us from the "interactive standpoint" we take up when we communicate and co-operate with others, thus treating all others as possible co-operators and possible conversation partners: as agents whose aims we may come to share or who may come to share our aims, where this co-operation must be decided upon based on mutual decisions. Treating humanity as a mere means, in contrast, consists, on this reading, in acting in ways that "block the agency" of the affected party, by not giving him or her a chance to decide for him or herself whether or not to contribute to the ends we ourselves are pursuing.[370]

Other influential writers also take very similar views of how to understand the humanity formula. Onora O'Neill is one example. Explaining her interpretation of the humanity formula, she writes:

[370] And the reason why Langton discusses the Kant-von Herbert correspondence is in part that she thinks, as we have seen, that it illustrates a shift, on Kant's part, from treating another as an end (Kant's initial interactive communication with von Herbert) to treating another as a mere means (Kant's using von Herbert's letters as an example of warning without asking for her consent, thereby blocking her opportunity to be a co-operator).

> To treat something as a mere means is to treat it in ways that are appropriate to things. Things, unlike persons, are neither free nor rational; they lack the capacities required for agency. They can only be props or implements, never sharers or collaborators, in any project. […] When we impose our wills one things we do not prevent, restrict or damage their agency – for they have none. […]
>
> By contrast, if we treat other agents as mere means, we do prevent, damage, or restrict their agency. We use them as props or implements in our own projects, in ways that preempt their willing and deny them the *possibility* of collaboration or consent – or dissent.[371]

Consider also Korsgaard's very similar reading of the humanity formula:

> The question whether another can assent to your way of acting can serve as a criterion for judging whether you are treating her as a mere means… [And] knowledge of what is going on and some power over the proceedings are the conditions of possible assent; without these, the concept of assent does not apply. [...] A similar analysis can be given of the possibility of holding the end of the very same action. In cases of violation of perfect duty, lying included, the other person is unable to hold the end of the very same action because the way you act prevents her from *choosing* whether to contribute to the realization of that end or not.[372]

On Korsgaard's reading of the mere means clause of the humanity formula, then, the way to avoid treating the humanity in others as a means only is to always make sure that they possess "knowledge of what is going on" and "some power over the proceedings". We treat the humanity in another as an end, in contrast, if we always give her a chance to decide for herself whether or not to contribute to our own ends. As Korsgaard also puts it, on her reading:

> Every rational being gets to reason out, for herself, what she is to think, choose, or do. So if you need someone's contribution to your end, you must put the facts before her and ask for her contribution. [...] Any attempt to control the actions and

[371] (O'Neill 1989: 138)
[372] (Korsgaard 1996b: 139)

> reactions of another by any means except an appeal to reason treats her as a mere means...[373]

These three very similar views of how to understand the humanity formula offered by Langton, O'Neill, and Korsgaard have in common that they all focus exclusively on agency by taking whether we are treating humanity as an end to solely be a matter of whether we acting so as to enable everyone affected by our actions to exercise agency. Though such readings capture much of the truth about how to understand Kant's formula, and Kant's theory's being a kind of constitutivism about the basic principles of morality makes it very tempting to understand his humanity formula in this way, these readings nevertheless only capture part of the truth. That we now have an understanding of the highest good, as Kant thinks of it, will help us to see why this is so. For if what the humanity formula told us to do was only to enable all to exercise agency, then we would expect the whole and complete good to be constituted by everyone's exercising agency. This would be the only object of a will completely determined by the constitutive laws of human practical reason (which, on the most general level, tell us to always treat the humanity in each person as an end, and never as a means only).

It is, however, not the case that the whole and complete good, as Kant thinks of it, is constituted by a world in which everyone has a capacity to exercise agency and also exercises this agency. The highest good is instead a world of virtue (which indeed involves the exercise of fully autonomous agency) in which everyone *also* gets all the happiness they desire and their virtue makes them worthy of. The wholly agency-focused reading of the humanity formula that writers such as Langton, O'Neill, and Korsgaard

[373] (Korsgaard 1996b: 142)

offer cannot, I believe, make sense of why it is *that* rather than a world of goal-directed agency, that constitutes the highest good, as Kant thinks of it.

Nor can defenders of the wholly agency-based reading make sense of Kant's corresponding claim that making the humanity in each person into an end involves making the happiness of each person into an end of ours.[374] A person's happiness, on Kant's view, does *not* consist in her exercising agency. It consists in a state where we can be content with how our lives are going since everything is going in accordance with our own desires and wishes, and all our needs, thus, are met. If treating humanity as an end only consisted in enabling people to exercise agency it would be a mystery why we should bother to try to make everyone happy, insofar as we are able to do anything about it.

The solution to the mystery is that the wholly agency-based reading only captures part of the truth about how to understand Kant's humanity formula. To make the humanity in each person into the most general aim or purpose around which all our actions are organized consists, as I have argued in these chapters, in making the preservation and full realization of the humanity in each person into our most general aim in life. And the humanity in our person, on Kant's view, consists not only in our capacity for agency, but also at the same time in how we are animals with desires, needs, hopes and fears and – insofar as these organized into a unified a whole – a general desire for

[374] This can appear to be a duty that we could not possibly discharge. How could I, as a single person, possibly promote the happiness of all? It would indeed be absurd to require of any single agent that she alone be *successful* in promoting the happiness of each other agent. But this objection overlooks the possibility of *group-agency*: it overlooks what several agents, jointly acting together as a single group-agent, can achieve through their shared efforts. Together we can strive to bring about the kind of egalitarian social conditions in which everyone has a chance of achieving happiness in accordance with her own conception of happiness. And on an individual level we can in addition be helpful, friendly, etc. towards our fellows.

happiness. Making the full realization of the humanity in each person into an end of ours involves, for Kant, making the flourishing of the human being as the particular type of being she is, with her own distinctive dual nature, into a positive end of ours. And this involves the flourishing of the whole human being, not only her flourishing with respect to her capacity for autonomous agency, but also with respect to her capacity as a being who is a seeker of happiness: the possible state where all our needs are met, all our desires are fulfilled, etc.[375]

One of the confusing things about Kant's theory as a whole is that he starts out with the conditions of "inner freedom" through autonomous agency – which we achieve by ourselves being the authors of the laws that govern us – but then moves to laws in accordance with which our particular human nature can be both preserved and fully realized or, in other words, in accordance with which we can flourish as the particular kind of beings that we are. This becomes less confusing, however, when we take note of Kant's suggestion that we let the laws of nature (as he understands them) serve as the "type" on which to model the maxims that are to serve as the laws we give to ourselves. The laws of nature, as Kant thinks of them, are, as we saw in chapter two, the most general principles in accordance with which things of particular kinds can exist and fully realize their nature. And in relation to the human being – whose nature does not only consist in the capacity for agency, but also is that of an animal with various desires and

[375] Allen Wood (2008) in effect implicitly offers an alternative suggestion as to why making the happiness of others into our end is part of treating the humanity in each person as an end. On his reading, this is because human beings are valuable in themselves and it, therefore, is good if their needs are met and they are happy. This is, I believe, too fast and loose to be what Kant has in mind. And we have already seen in chapter three that Wood is mistaken about the relation between what has absolute value (a good will) and what exists as a purpose in itself (the humanity within us). So we should reject Wood's explanation of why making humanity into an end involves making the happiness of each person into an end of ours.

needs – this means making the preservation and full realization of the whole human being into the object of the laws in question.[376]

With regard to Langton's criticism of Kant the person, then, I conclude the following. Langton's criticism of Kant's conduct on the given occasion might very well be right. But the *reading* the humanity formula Langton uses as the basis of her criticism – a reading that is very similar to those of influential writers such as O'Neill and Korsgaard – is not a reading that captures the whole of what Kant has in mind with his humanity formula. Treating the humanity in each person as a purpose in itself, as Kant thinks of it, not only involves making sure to enable people to function as agents, but also importantly involves having as our most general purpose that everyone be able to able to flourish in the dual nature they have as human beings, which involves achieving happiness and not just autonomous agency.

7. Concluding Remarks

This completes my attempt to replace what I believe to be some of the greatest misunderstandings about Kant's moral theory (and in particular the universal law and humanity formulas) within contemporary Anglophone moral philosophy with a more accurate reading of the basic ideas therein. I started, in chapter one, by claiming that Kant's moral theory isn't widely understood, and I hope to have shown that much of the

[376] Note that this does not amount to deriving the foundational principle of morality from contingent facts about human nature. Kant insists repeatedly that we mustn't do that. The basic principle of morality is to choose one's maxims on the basis of their fitness to serve as universal laws, and the argument for why that, on the most general level, is the foundational principle of morality is not based on facts about human nature in particular. The argument for this instead has to do with the preconditions of autonomous agency, and the analysis of the concept of the will of a reason-endowed being. It is in the choice of maxims that are to serve as universal laws that we move to using the laws of nature as our model and the preservation and full realization of the human nature as the particular things these maxims are to serve as laws of.

contemporary discussion both involves interpretative mistakes and is uncharitable to Kant. I also claimed that insofar as we want to follow Kant in thinking of moral requirements as categorical imperatives (rather than as mere hypothetical imperatives), we need an alternative to the kind of sensibility-relative methodological intuitionism that dominates so much of the contemporary debate of ethics within analytic philosophy. This was my second main motivation behind the discussion we've just completed.[377]

With regard to how to understand Kant's theory, we have seen that Korsgaard does indeed seem to be right in comparing Kant's theory to the theory that Plato puts forward in the *Republic*. Not only does Kant follow Plato in understanding virtue in terms of a kind of inner harmony among the different parts of our souls with our practical reason serving as the ruler that helps to accomplish this inner peace (on account of which we can constitute ourselves as responsible agents and the authors of our own actions). Kant also appears to follow Plato in taking a political analogy as the model on which he builds his theory of how to achieve this inner freedom.

---

[377] What I have in mind by this second motivation is that such intuition-based defenses of basic moral principles are all sensibility-relative; that our moral sensibilities differ; and that any such attempts to justify basic moral principles, therefore, only can enjoy justificatory force relative to the particular sensibilities appealed to by those offering these justifications (which are usually their own sensibilities). If basic principles could only be justified and defended relative to particular sensibilities, however, then they would ultimately be better regarded, I believe, as hypothetical imperatives that apply to us (if at all) insofar as we share those sensibilities, and we have a will to live in accordance with our own sensibilities.

The second main motivation I have had for investigating Kant's moral theory in this dissertation, then, is that it appears to constitute an alternative way of justifying basic moral principles, which is *sensibility-neutral*, and that our understanding of moral requirements as categorical imperatives should lead us to investigate all major alternatives to the methodological intuitionism that dominates so much of the contemporary debates. Kant's ethical theory is one of the major alternatives, and I hope that these chapters can contribute to a better understanding of this particular theory. I hope to have shown that Kant's theory is a counterexample to Scanlon's claim that all apparent alternatives to reflective equilibrium-seeking intuition-drive moral reasoning are "illusory."

But in Kant's case it is not Plato's ideal city that serves as the model, but instead Rousseau's ideal whereby the citizens of a Republic can be free at the same time as being subject to absolute laws by themselves being the legislators of those laws (in Rousseau's case as members of the body politic). In Kant's case the inner rule of law through which harmony of the soul is achieved, and the person thus becomes fully autonomous (i.e. self-governed through self-adopted laws), is achieved through our subjecting ourselves to maxims (i.e. basic guiding principles) that are fit to serve as universal laws for all reason-endowed beings with a will of their own. Making it very clear that he is using this kind of Platonic strategy, Kant also explicitly likens subjecting oneself to maxims that could serve as universal laws to moving from an ethical state of nature into an ethical state of virtue just as we can accomplish peace, as Hobbes argues, by moving from the juridical state of nature to a juridical state.

When it comes to the criteria by which maxims can qualify for such internal law-giving, Kant turns away from the political metaphor, and instead turns to the analogy of the laws of nature (as he understands them). And the key to understanding the universal law formula is, therefore, in part to understand what Kant means by "nature" and "laws of nature" within his philosophy as a whole (which is supposed to constitute one big philosophical system). We've seen that the laws of a particular *nature* (as in a particular kind of constitution) are the most general principles in accordance with which this type of nature can exist and be fully realized. Maxims that are fit to serve as universal laws of our particular nature are, therefore, basic guiding principles all of us could adopt and whose following would allow us to both preserve and fully realize our own distinctive nature.

To add content to this idea (and not have it turn out being the "empty formalism" that Hegel accused Kant of presenting) we need to add an idea of the human nature. Kant's suggestion is (in short) that it is distinctive of our nature that we have a dual nature whereby we are both (a) animals with inclinations and needs that, if unified into a harmonious whole, results in a general desire for happiness and (b) reason-endowed beings capable of autonomous agency. Maxims that can serve as laws of our nature are, therefore, basic guiding principles whose following would allow us to harmonize these two different main aspects of our nature. And on the most general level such maxims are guiding principles whereby our pursuit of happiness is subordinated to the condition of its compatibility with our retaining the capacity to always be able to conduct ourselves on the basis of self-adopted guiding principles.

This dual nature is what characterizes our distinct humanity, as Kant thinks of it. And since (a) it is only if we subject ourselves to basic guiding principles all could follow and thereby all preserve and fully realize our humanity, and (b) subjecting ourselves to such maxims would amount to making our own humanity in each person into the most general purpose around which all our actions are organized, it follows, Kant takes it, that from the point of view of the basic principles of autonomous human agency, our own humanity exists as a purpose in itself: as something that is an objective end in accordance with these basic laws of fully autonomous human agency.

Keeping in mind, in this way, that Kant's theory is a kind of constitutivism is thus, I have argued, the key to understanding the reasoning surrounding his initial statement of the humanity formula in the *Groundwork*. And the supposed equivalence between the universal law and humanity formulations of the categorical imperative,

which has seemed to so mysterious to so many writers, becomes obvious once we realize that following the universal law formula as it applies to human beings in particular itself simply amounts to treating the humanity in each person as the most general purpose around which our actions are organized. This is so because in following the universal law formula as it applies to us human beings, we would choose our basic guiding principles on the basis of their fitness to serve as the laws in accordance with which to preserve and fully realize our own humanity; and to do that is to have the preservation and full realization of the humanity within each person as the most general purpose around which we organize our actions.[378]

And so we have come to the end of our discussion. There are numerous issues we have not covered, such as various possible objections that might be raised against Kant's theory, as I have suggested that we ought to understand it. There are also other objections to Kant's theory in the literature, which we could offer new responses to using the interpretations of the universal law and humanity formulas put forward above. These, however, will have to be topics for another discussion, on some other occasion.

[378] There is a sense, then, in which the humanity formula itself is the most general maxim that could be will as a universal law. And that is why, I believe, Kant at one point makes the claim that while the universal law formula specifies the *form* all our maxims are to take, the humanity formula specifies their *matter*. This is another claim that, in light of the standard readings, has appeared mysterious to most commentators, but which I think we can make good sense of using the reading I have presented in this dissertation.

BIBLIOGRAPHY

Allison, Henry (2011): *Kant's Groundwork for the Metaphysics of Morals: A Commentary* (Oxford: Oxford University Press)

Aristotle (1999): *Nicomachean Ethics* Translated by T.H. Irwin (Indianapolis: Hackett Publishing Co.)

Benton, Robert J. (1982): "Political Expediency and Lying: Kant vs Benjamin Constant," *Journal of the History of Ideas*, Vol. 43, No. 1: 135-44

Cummiskey, David (1996): *Kantian Consequentialism* (Oxford: Oxford University Press)

Darwall, Stephen (editor) 2003: *Consequentialism* (Oxford: Blackwell)

Darwall, Stephen (2006): *The Second-person Standpoint: Morality, Respect, and Accountability* (Cambridge, Mass.: Harvard University Press)

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

Dean, Richard (2011): "Glasgow's Conception of Humanity", *Journal of the History of Philosophy* Vol. 46. No. 2: 307-14

Dietrichson, Paul (1969): "Kant's Criteria of Universalizability," in Robert Paul Wolff (ed.), *Kant: Foundations of the Metaphysics of Morals: Text and Critical Essays* (Indianapolis: Bobbs-Merrill)

Forst, Rainer (2011): *Kritik der Rechtfertigungsverhältnisse*, (Frankfurt am Main: Suhrkamp Verlag)

Greene, Joshua (2008): "The Secret Joke of Kant's Soul" in Walter Sinnot-Armstrong (ed.) *Moral Psychology, Vol. 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, (Cambridge, Mass.: The MIT Press): 35-79

Gregor, Mary (1963): *Laws of Freedom* (Oxford: Basil Blackwell)

Guyer, Paul (1998): "The Value of Reason and the Value of Freedom", *Ethics* 109: 22-35

Guyer, Paul (2006): *Kant* (New York: Routledge)

Habermas, Jürgen (1991): *Moral Consciousness and Communicative Action* (Cambridge, Mass.: The MIT Press)

Hare, R.M. (1997): *Sorting Out Ethics* (Oxford: Oxford University Press)

Hegel, G.W.F. (1822/1977): *Hegel's Philosophy of Right*, Translated by A.V. Miller (Oxford: Oxford University Press)

Herman, Barbara (1989): "Murder and Mayhem: Violence and Kantian Casuistry", *Monist* 72:3: 411-31

Herman, Barbara (1993): *The Practice of Moral Judgment* (Cambridge, Mass.: Harvard University Press)

Herman, Barbara (2011): "A Mismatch of Methods" in Samuel Sheffler (ed.), *On What Matters, Vol. Two*, (Oxford: Oxford University Press)

Hill, Thomas (1972): "The Kingdom of Ends" in L. White Beck (ed.) *Proceedings of the Third International Kant Congress*, (Dordrecht: D. Riedel): 307-15

Hill, Thomas (1980): "Humanity as an End in Itself", *Ethics*, Vol. 91: 84-99

Hill, Thomas (2002): *Human Welfare and Moral Worth* (New York: Oxford University Press)

Hobbes, Thomas (1651/1994): *Leviathan* Edited by E. Curley (Indianapolis: Hackett Publishing Co.)

Hume, David (1740/1968): *A Treatise of Human Nature* (Oxford: Oxford University Press)

Kane, Robert (2007): "Libertarianism" in John Martin Fischer (ed.), *Four Views on Free Will* (Oxford: Blackwell): 5-43

Kant, Immanuel (1781/1787): *Kritik der Reinen Vernunft,* in *Immanuel Kant: Werkausgabe, XII Bände* Edited by Wilhelm Weischedel (Frankfurt am Main: Suhrkamp, 1997)

Kant, Immanuel (1783): *Prolegomena zu jeden künftigen Metaphysik,* in *Immanuel Kant: Werkausgabe, XII Bände* Edited by Wilhelm Weischedel (Frankfurt am Main: Suhrkamp, 1997)

Kant, Immanuel (1785): *Grundlegung zur Metaphysik der Sitten,* in *Immanuel Kant:*

Werkausgabe, XII Bände Edited by Wilhelm Weischedel (Frankfurt am Main: Suhrkamp, 1997)

Kant, Immanuel (1786): *Metaphysische Anfangsgründe der Naturwissenschaft,* in *Immanuel Kant: Werkausgabe, XII Bände* Edited by Wilhelm Weischedel (Frankfurt am Main: Suhrkamp, 1997)

Kant, Immanuel (1788): *Kritik der Praktischen Vernunft,* in *Immanuel Kant: Werkausgabe, XII Bände* Edited by Wilhelm Weischedel (Frankfurt am Main: Suhrkamp, 1997)

Kant, Immanuel (1790): *Kritik der Urteilskraft,* in *Immanuel Kant: Werkausgabe, XII Bände* Edited by Wilhelm Weischedel (Frankfurt am Main: Suhrkamp, 1997)

Kant, Immanuel (1793): *Die Religion Innerhalb der Grenzen der Bloßen Vernunft,* in *Immanuel Kant: Werkausgabe, XII Bände* Edited by Wilhelm Weischedel (Frankfurt am Main: Suhrkamp, 1997)

Kant, Immanuel (1793): *Über den Gemeinspruch: Das mag in der Theorie richtig sein, taugt aber nicht für die Praxis,* in *Immanuel Kant: Werkausgabe, XII Bände* Edited by Wilhelm Weischedel (Frankfurt am Main: Suhrkamp, 1997)

Kant, Immanuel (1797): *Die Metaphysik der Sitten,* in *Immanuel Kant: Werkausgabe, XII Bände* Edited by Wilhelm Weischedel (Frankfurt am Main: Suhrkamp, 1997)

Kant, Immanuel (1797): *Über ein Vermeintes Recht aus Menschenliebe zu Lügen* In *Immanuel Kant: Werkausgabe, XII Bände* Edited by Wilhelm Weischedel (Frankfurt am Main: Suhrkamp, 1997)

Kant, Immanuel (1798): *Streit der Fakultäten,* in *Immanuel Kant: Werkausgabe, XII Bände* Edited by Wilhelm Weischedel (Frankfurt am Main: Suhrkamp, 1997)

Kant, Immanuel (1798): *Anthropologie in pragmatischer Hinsicht,* in *Immanuel Kant: Werkausgabe, XII Bände* Edited by Wilhelm Weischedel (Frankfurt am Main: Suhrkamp, 1997)

Kant, Immanuel (1967): *Philosophical Correspondence 1759-99* Translated and edited by Arnulf Zweig (Chicago: University of Chicago Press)

Kant, Immanuel (1800/1992): *Lectures on Logic* Translated and edited by J. Michael Young (Cambridge: Cambridge University Press)

Kant, Immanuel (1817/1996): *Lectures on Philosophical Doctrine of Religion* Translated by Allen Wood, in Allen W. Wood and George di Giovanni (editors), *Religion and Rational Theology*, (Cambridge: Cambridge University Press)

Kant, Immanuel (1997): *Lectures on Ethics* edited by Peter Heath and J.N. Schneewind (Cambridge: Cambridge University Press)

Kant, Immanuel (2005): *Notes and Fragments* Translated by Curtis Bowman, Paul Guyer, and Frederick Rauscher; edited by Paul Guyer

Kastafanas, Paul (2011): "Deriving Ethics from Action: A Nietzschean Version of Constitutivism", *Philosophy and Phenomenological Research* Vol. LXXXIII No. 3: 620-60

Kerstein, Samuel (2002): *Kant's Search for the Supreme Principle of Morality* (Cambridge: Cambridge University Press)

Kitcher, Philip (1986): "Projecting the Order of Nature" in R.E. Butts (ed) *Kant's Philosophy of Physical Science* (Dordrecht: Reidel)

Korsgaard, Christine (1986a): "Kant's Formula of Humanity", *Kant-Studien* 77:2: 183-202

Korsgaard, Christine (1986b): "Aristotle and Kant on the Source of Value", *Ethics* 96: 486-505

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

Korsgaard, Christine (1996b): *The Sources of Normativity* (Cambridge: Cambridge University Press)

Korsgaard, Christine (1999): "Self-Constitution in the Ethics of Plato and Kant", *The Journal of Ethics* 3: 1-29

Korsgaard, Christine (2009): *Self-Constitution* (Oxford: Oxford University Press)

Langton, Rae (1992): "Duty and Desolation", *Philosophy* 67: 481-505

Mackie, J.L. (1977): *Ethics: Inventing Right and Wrong* (New York: Penguin)

Markovits, Julia (2011): "Why be an Internalist about Reasons?", in Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics Vol. 6*, (Oxford: Oxford University Press, 2011): 255-79

Mill, John Stuart (1861/1963): *Utilitarianism* in J.M. Robson (ed.), *Collected Works of John Stuart Mill,* Vol. 10 (Toronto: University of Toronto Press): 203-59

O'Neill (Nell), Onora (1975): *Acting on Principle: An Essay on Kantian Ethics* (New York: Columbia University Press)

O'Neill, Onora (1989): *Constructions of Reason: Exploration of Kant's Practical Philosophy* (Cambridge: Cambridge University Press)

Parfit, Derek (1984): *Reasons and Persons* (Oxford: Clarendon Press)

Parfit, Derek (2011a): *On What Matters, Vol. One* (Oxford: Oxford University Press)

Parfit, Derek (2001b): *On What Matters, Vol. Two* (Oxford: Oxford University Press)

Persson, Ingmar (2005): *The Retreat of Reason: A Dilemma in the Philosophy of Life* (Oxford: Oxford University Press)

Pettit, Philip (1997): *Republicanism: A Theory of Freedom and Government* (Oxford: Clarendon Press)

Pettit, Philip (2003): "Consequentialism" in Stephen Darwall (ed.) *Consequentialism* (Oxford: Blackwell)

Plato (2004): *The Republic* Translated by C.D.C Reeve (Indianapolis: Hackett Publishing Co.)

Rabinowicz, Wlodek and Rønnnow-Rasmussen, Toni (2004): "The Strike of the Demon: on Fitting Pro-Attitudes and Value", in *Ethics* 114: 391-423

Railton, Peter (2000): "Normative Force and Normative Freedom: Hume and Kant, but not Hume *versus* Kant" in Jonathan Dancy (ed.), *Normativity* (Oxford: Blackwell)

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

Rawls, John (2000): *Lectures on the History of Moral Philosophy* (Cambridge, Mass.: Harvard University Press)

Rawls, John (2001): *Justice as Fairness – A Restatement* (Cambridge, Mass.: Belknap Press)

Regan, Donald H. (2002): "The Value of Rational Nature," *Ethics* 112: 267-91

Reich, Klaus (1939a): "Kant and Greek Ethics I", *Mind* Vol. 48, No. 191: 338-54

Reich, Klaus (1939b): "Kant and Greek Ethics II", *Mind* Vol. 48, No. 192: 446-63

Rousseau, Jean-Jacque (1762/2005): *Of The Social Contract, Or Principles of Political Right* Translated by G.D.H. Hole (Stilwell: Digireads)

Sayre-McCord, Geoffrey (2001): "Mill's "Proof" of the Principle of Utility: A More than

Half-Hearted Defense" *Social Philosophy and Policy* Vol. 18: No. 2: 330-60

Scanlon, T.M. (1998): *What We Owe to Each Other* (Cambridge, Mass.: Harvard University Press)

Scanlon, T.M. (2003): "Rawls on Justification", in Samuel Freedom (ed), *The Cambridge Companion to Rawls* (Cambridge: Cambridge University Press)

Scanlon, T.M. (2008): *Moral Dimensions: Permissibility, Meaning, Blame* (Cambridge, Mass.: Harvard University Press)

Schroeder, Mark (2007): *Slaves of the Passions* (Oxford: Oxford University Press)

Seneca (2010): *Seneca: Anger, Mercy, Revenge, The Complete Works of Lucius Annaeus Seneca* Translated by R. Kaster and M. Nussbaum and edited by Asmis, E., Bartsch, S., and Nussbaum, M. (Chicago: Chicago University Press)

Senson, Oliver (2009): "Kant's Conception of Human Dignity," *Kant-Studien* 100. Jahrg.: 309-31

Sheffler, Samuel (ed.) (2011): *On What Matters, Vol. Two* (Oxford: Oxford University Press)

Sidgwick, Henry (1907): *The Methods of Ethics, 7th Edition* (London: MacMillan)

Skinner, Quentin (1998): *Liberty before Liberalism* (Cambridge: Cambridge University Press)

Smith, Michael (1994): *The Moral Problem* (Oxford: Wiley-Blackwell)

Smith, Michael (1995): "Internal Reasons", *Philosophy and Phenomenological Research* Vol. LV, No. 1: 109-31

Smith, Michael (2011): "Deontological Moral Obligations and Non-Welfarist Agent-Relative Values", *Ratio (new series)* XXIV 4: 351-63

Smith, Michael (forthcoming): "A Constitutivist Theory of Reasons for Action", *Law, Ethics, and Philosophy*

Strawson, P.F. (1974): *Freedom and Resentment and Other Essays* (London: Methuen)

Street, Sharon (2006): "A Darwinian Dilemma for Moral Realism", *Philosophical Studies* Vol. 127, No. 1: 109-66

Timmermann, Jens (2006): "Value without Regress: Kant's 'Formula of Humanity' Revisited, *European Journal of Moral Philosophy*, 14:1: 69-93

Velleman, J. David (2000): *The Possibility of Practical Reason* (New York: Oxford University Press)

Wolf, Susan (1982): "Moral Saints", *Journal of Philosophy* Vol. LXXIX, No. 8: 419-39

Wolf, Susan (2011): "Hiking the Range" in Samuel Sheffler (ed) *On What Matters, Vol. Two* (Oxford: Oxford University Press)

Wood, Allen (1999): *Kant's Ethical Theory* (Cambridge: Cambridge University Press)

Wood, Allen (2002): Translation of, and introduction to, *Groundwork for the Metaphysics of Morals* (New Haven: Yale University Press)

Wood, Allen (2006): "The Supreme Principle of Morality" in Paul Guyer (ed.) *The Cambridge Companion to Kant and Modern Philosophy* (Cambridge: Cambridge University Press): 342-81

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

Wood, Allen (2011): "Humanity as an End in Itself" in Samuel Sheffler (ed) *On What Matters, Vol. Two* (Oxford: Oxford University Press)