The Moral Point of View in Hume, Kant and Mill

Margaret Marie Chiovoloni

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Philosophy.

Chapel Hill

2011

Approved by

Thomas E. Hill, Jr.

Susan Wolf

Alan Nelson

Douglas MacLean

Bernard Boxill

Abstract

Margaret Marie Chiovoloni

The Moral Point of View in Hume, Kant and Mill

(Under the direction of Thomas E. Hill, Jr.)

Hume, Kant and Mill each approach morality with distinctly different frameworks and methodologies, but it is important to acknowledge that they all share the crucial thought that morality involves an impartial point of view. Hume and Kant both recognize that morality is universal: what is right for one must be right for all. They both use this recognition as the starting point for their investigations. Furthermore, we can interpret Mill's proof so that it moves from a first-personal point of view to an impartial point of view. By addressing the most serious objections to the role of universality in each of these philosophical systems, I make room for acknowledging the shared ground from which each of these philosophers begin their investigations.

Critics of Hume have worried that emphasizing the general point of view in his system will lead to conflicts with those passages in which he denies the role of reason in morality. I argue that these conflicts can be resolved by attending to a distinction between moral reactions and moral judgments.

Critics of Kant object to his argument that the universal law formulation is a formulation of the Categorical Imperative by claiming that this general point of view does not accurately pick out worthy maxims. I argue that Kant does not expect the

general point of view, as expressed in FUL, alone to do this. Instead, we must also rely on a universal end—humanity—and this insight is expressed in the humanity formulation.

Critics of Mill have objected to his proof by objecting to two central premises: that what is desired is desirable, and that if each person's happiness is desirable, the aggregate happiness is desirable. I argue that it is essential to Mill's proof that the former is said from the first-person point of view, whereas the latter is said from the moral (general) point of view. The former premise means that the fact that I desire something is evidence to me that it is desirable for me. When we move through the argument to the latter premise, we move to a more general point of view.

Acknowledgements

I am grateful to my husband, David Landy, for his untiring support.

Table of Contents

Introduction

Hume, Kant and Mill each approach morality with distinctly different frameworks and methodologies, but it is important to acknowledge that they all share the crucial thought that morality involves an impartial point of view. Of course, to say that morality *involves* an impartial point of view is to say something very vague—and the way in which morality involves an impartial point of view will be importantly different for all three thinkers. Hume and Kant both recognize that morality is *universal*, that it applies to all of us[1], and so what is right for one of us must be right for all of us. Hume and Kant both use this recognition as the starting point for their investigation, and we will turn in a moment to discuss the relationship for each between an impartial point of view and universality. For Mill, this recognition plays a more subtle role, but I will argue that we can rescue the proof of utilitarianism in Chapter 4 from the traditional objections with an interpretation according to which Mill moves with us *from* a first-personal point of view *to* a general or impartial point of view.

As I am using the term, an impartial point of view differs from a first-person point of view in that the unique concerns and commitments inherent in a first-person point of view are not as forceful when considered from the impartial point of view. From my own point of view, i.e. from my "first-person" point of view, I have a set of interests, goals,

[1]For Hume, "all of us" means all *humans*, whereas for Kant, it means all rational beings (which presumably includes most humans).

commitments, feelings, etc., that is unique to me, and not shared by others.[2] When I make a decision, I will tend to put that set of concerns and commitments front and center. They will be the most important considerations in my decision-making process. From an impartial point of view, however, I must take this set of concerns and commitments and "put it in perspective"—i.e., recognize that it is merely one set among many sets of concerns and commitments. This is my set of concerns and commitments, but others have such sets as well. The impersonal point of view is one from which all of these sets are viewed from the outside, so to speak. From this perspective, the *grip* that my concerns and commitments have on me *as their subject* is loosened, and I am able to weigh the concerns and commitments of others more impartially. On one plausible line, taking up this kind of impartial point of view is precisely how it is that *universality* is introduced into moral thinking as well. The impartial point of view will be a universal point of view—that is, it is a point of view that arguably will be the same for all of us. There will be contingencies and differences in how each philosopher makes this connection, though, so let us now take a closer look at the impartial point of view in each moral theory.

Hume is appropriately sensitive to the universality issue, as can be seen in the *Treatise*, where he addresses an objection to his attempt to ground morality in *sympathy*. According to the objection, an account that takes sympathy as its central concept will be too variable and insufficiently universal to represent what we expect from an account of morality. Hume acknowledges the concern, agreeing that morality requires universality. He attempts to appease the objector by introducing a *general point of view*. The modification is an attempt to capture *both* sentiment (sympathy) *and* universality on his

[2]Of course, *some* of my concerns and commitments may be shared by some other people. It may even be conceivable that there would be two people who shared exactly the same concerns and commitments. The important point, though, is that there are differences in interests between agents. Some people care about some things, and other people care about other different things.

view. When we adopt the general point of view, he claims, we ignore the differences that distance and time can make to our non-moral reactions. From the general point of view, it does not matter how close to or far from someone we are. We adopt a point of view where we ignore the features particular to us, in our particular circumstances.

> In general, all sentiments of blame or praise are variable, according to our situation of nearness or remoteness, with regard to the person blam'd or prais'd, and according to the present disposition of our mind. But these variations we regard not, in our general decisions, but still apply the terms expressive of our liking or dislike, in the same manner, as if we remain'd in one point of view.

Hume's view, then, is that *moral* reactions are reactions felt from the general point of view, and *moral* evaluations are evaluations that only draw on *moral* reactions (both our own and those of others). True moral judgments will not rely on non-moral reactions as evidence—i.e. true moral judgments will only rely for evidence on reactions that are had *from the general point of view*. We will see that there are complications, but the essential point here is to observe the centrality of a general point of view in Hume's moral system.

Critics have worried that an interpretation which emphasizes the general point of view in Hume's system will lead to conflicts with passages in which he denies the role of reason in morality. I will argue that these conflicts can be resolved by attending to a distinction between moral *reactions* and moral *judgments*. Reason has a proper place in moral judgment, but it does not play a role in determining virtue and vice, which are instead determined by the reactions we have from the moral point of view (hypothetical or actual). This distinction allows us to respect the importance of the general point of view in Hume's ethics while resolving any worries about purported conflicting passages.

Kant is less explicit regarding the general point of view, but he is explicit about *universality*. Kant thinks that the content in the first formulation, the universal law formulation, is derived from the very concept of a categorical imperative, from the very

concept of a moral law, and this concept essentially involves universality. The categorical imperative is a universal law, and as such it applies to all of us. Kant writes,

> But if I think of a *categorical imperative*, I know right away what it contains. For since this imperative contains, besides the law, only the necessity that the maxim conform to this law, while the law, as we have seen, contains no condition limiting it, there is nothing left over to which the maxim of action should conform except the universality of a law as such; and it is only this conformity that the imperative asserts to be necessary (4:222).

If the categorical imperative applies to all of us, then we must only act in accordance with maxims that can be willed to be universal law (because if the maxim cannot be willed to be universal law, then it does not apply to all of us and therefore cannot be a principle of morality). But this is the universal law formulation: 'Act only on that maxim by which you can at the same time will that it should become a universal law' (4:222).

So far, then, we have seen that the very concept of a categorical imperative is the concept of a *universal* law. But I also think that the universal law formulation is Kant's way of capturing the general point of view. We test our maxims by making sure that they could be willed to be universal law—that is, by making sure that they can apply to *everyone*, and not just us. But this test involves adopting a general point of view, in a way. It involves ignoring my own aims, inclinations, and desires, and instead evaluating the maxim of my action from a general perspective. Worthy maxims will be those that I choose, not from my own perspective (keeping in mind my own goals and inclinations), but from a more general perspective, one that involves evaluating them for adoption *universally*.

Critics object to Kant by claiming that this general point of view does not accurately pick out all and only worthy maxims. I will argue that Kant does not expect the general point of view, as expressed in the universal law formulation, alone to do this. Instead, we must also rely on a universal end—humanity—and this insight is expressed in

the humanity formulation. As always, there will be complications, but again, the important thing is to recognize that for Kant as for Hume, the general point of view (the universal law formulation) is a way of capturing universality, which is essential for any proper system of morals.

It is far less clear that Mill's starting point is the same as that of Hume and Kant, the universality of morality and the importance of the general point of view. However, I will argue that a charitable interpretation of his proof is one that pays careful attention to the transition from the first-person point of view into the general point of view that his conclusion requires. We will see that the relevant part of his proof can be reconstructed as follows:

P1. What is desired is desirable. (Or, a desire for X is evidence that X is desirable.)
P2. Happiness is *desired*.
P3. Happiness is *desirable* (or at least, we have evidence that happiness is desirable, from P1 and P2).
P4. Since each person's happiness is desirable, the aggregate happiness is desirable.

Mill begins his proof with the claim that the fact that something (in this case, happiness, but the principle is broad enough to apply to anything) is desir*ed* is evidence that it is desir*able*—P1. Later, in P4, he goes on to move *from* the fact that individual happiness is desirable (to the individual) *to* the fact that the aggregate happiness is desirable (to the aggregate). Critics have tended to object to P1 on the grounds that the fact that I desire something is no evidence that it is desir*able* (and they understand desirable as being worthy of desire very generally). Critics have further tended to object to P4 on the grounds that *I* have no reason to pursue something that is desirable to the aggregate.

I will argue that it is essential to Mill's proof that P1 is said from the *first-person* point of view, whereas P4 is said from the *moral* (general) point of view. When Mill claims that the fact that I desire x implies that x is desirable, the implication is claimed

from the first-person point of view. That *I* desire something is evidence *to me* that it is desirable *for me*. But when we move through the argument to P4, we move to a more general, *moral* point of view. The move *from* my happiness being good to me *to* the aggregate happiness being a good to the aggregate is legitimated by taking up a general, moral point of view. I will argue that interpreting P4 from the *general, moral* point of view will gain traction when we recognize that since Mill addressed moral motivation in chapter 3, this proof is more plausibly interpreted as addressing the question of what makes something moral at all, and not as addressing the question of why I should *be* moral. As I will show in my paper, we should instead understand P4 as a claim about what is desirable *from the moral point of view.*

Hume and the General Point of View

Some commentators on Hume have found it difficult to reconcile his introduction of the "general point of view" with the rest of his moral theory. In particular, there seems to be a tension between his frequent proclamations against any role of reason in morality, on the one hand, and what appears to be the required place of reason in the general point of view, on the other. If it is true that "in order…to arrive at a more *stable* judgment of things, we fix on some *steady* and *general* points of view; **and always, in our thoughts, place ourselves in them, whatever may be our present situation"** (371-2)[3], then, commentators have argued, the view implies that we must exercise our faculty of reason in order to evaluate certain moral situations. Since the general point of view is explicitly a point of view different than our own, these commentators argue that we must use our faculty of reason to figure out what we *would* feel if we had actually adopted that point of view. However, if we must exercise our faculty of reason in order to adopt the general point of view, then it becomes difficult to make sense of Hume's proclamations *against* the role of reason in moral evaluation. To put it another way, some passages in Hume (namely, the passages in which he argues that we adopt a general point of view when making moral judgments) appear to be incompatible with what is standardly taken to be the "Humean" position, i.e. a form of antirationalism or noncognitivism in which there is no role for reason in moral evaluation.

[3] David Hume, *A Treatise of Human Nature*, ed. David Fate Norton and Mary J. Norton. Oxford University Press, 2000.

I argue that making a distinction between primary moral *reactions* (which are not the result of reasoning) and secondary moral *evaluations/judgments* (which can be the result of reasoning) will resolve this prime facie tension in Hume's moral system. On this interpretation, our faculty of reason has no role in our primary moral reactions. However, we do use our faculty of reason when we make secondary moral evaluations/judgments. I also argue that, for Hume, we first adopt the general point of view, and we then *feel* sentiments from that point of view. While we can, in theory, deduce what someone in the general point of view would feel by employing our faculty of reason, we can also actually adopt this point of view, and thereby *feel* the sentiments, instead of using our faculty of reason to deduce these sentiments. Ultimately, I argue that Hume is *not* an antirationalist or noncognitivist, even though on his view morality is not constituted by reason. Recognizing a distinction in his theory between moral reactions and moral judgments is what makes it possible to understand Hume in this way.

In the first section of this paper, I lay out my interpretation of the role of the general point of view, paying particular attention to how we should understand it in the context of Hume's overall moral theory. On this subject, we will consider first the Treatise, and then the Enquiry. In the second section, I defend my interpretation of the general point of view against an objection given by Cohon.[4] I will also consider one of Hume's central arguments, which I will argue supports my interpretation of the general point of view against Cohon's objection. In the third section, I extend this defense to Hume's claims about the motivational aspect of morality.

[4]Rachel Cohon, "The Common Point of View in Hume's Ethics," *Philosophy and Phenomenological Research*, v57:4, pp.827-850.

Section 1: The general point of view

Before we discuss why Hume introduces the general point of view, let us first take some time to understand the mechanism of sympathy. A deeper understanding of its mechanism will aid us well, given the role that sympathy plays in his moral system. The mechanism works as follows: We see another person, who feels some sentiment. We are aware of what sentiment the creature feels from outward signs (things he says, faces he makes), and therefore have an idea of the sentiment he feels.

> When any affection is infus'd by sympathy, it is as first known only by its effects, and by those external signs in the countenance and conversation, which convey an idea of it (206).

For example, suppose that you see a friend who has tears streaming down his face, shoulders slumped, etc. You then have an idea of sadness. The impression of his face causes you to have an idea of sadness, via the association of resembling ideas—his face resembles other faces when they are sad, which brings to mind the idea of sadness.

Then, we feel an impression of that very same sentiment. To continue the example, you yourself would feel sadness. The idea you have of your friend's sadness causes you to feel sadness, to have an impression of sadness. Then, the impression of that sentiment serves to enliven the idea we felt of his sentiment.

> This idea is presently converted into an impression, and acquires such a degree of force and vivacity, as to become the very passion itself, and produce an equal emotion, as any original affection" (206).

Your impression of your own sadness causes your idea of your friend's sadness to become enlivened, and hence to become an impression itself. Hume writes, on the phenomenon of enlivening ideas so that they become impressions,

> It has been remark'd in the beginning of this treatise, that all ideas are borrow'd from impressions, and that these two kinds of perceptions differ

> only in the degrees of force and vivacity, with which they strike upon the soul….The different degrees of their force and vivacity are, therefore, the only particulars, that distinguish them: And as this difference may be remov'd, in some measure, by a relation betwixt the impressions and ideas, 'tis no wonder an idea of a sentiment or passion, may by this means be so enliven'd as to become the very sentiment or passion." (207)

The mechanism by which an idea becomes an impression by becoming enlivened operates mainly by resemblance[5].

> There is a very remarkable resemblance [among human beings], which preserves itself amidst all their variety; and this resemblance must very much contribute to make us enter into the sentiments of others, and embrace them with facility and pleasure." (207).

However, it is also affected a great deal by certain other factors, such as contiguity in space and time (how close you are to the person in space and time), and by causation (blood relations, for example)[6].

> Accordingly we find, that where, beside the general resemblance of our natures, there is any peculiar similarity in our manners, or character, or country, or language, it facilitates the sympathy. **The stronger the relation is betwixt ourselves and any object, the more easily does the imagination make the transition, and convey to the related idea the vivacity of conception, with which we always form the idea of our own person.** (207, bold emphasis mine.)

Your feelings of sympathy for your neighbor are more pronounced than your feelings of sympathy for someone in a distant country whom you have never met (contiguity).[7] Hume writes,

---

[5]"The stronger the relation is betwixt ourselves and any object, the more easily does the imagination make the transition, and convey to the related idea the vivacity of conception, with which we always form the idea of our own person" (207).

[6]"The relations of blood, being a species of causation, may sometimes contribute to the same effect…" (207).

[7]"The sentiments of others have little influence, when far remov'd from us, and require the relation of contiguity, to make them communicate themselves entirely" (207).

> When we reflect, therefore, on any object distant from ourselves, we are oblig'd not only to reach it at first by passing thro' all the intermediate space betwixt ourselves and the object, but also to renew our progress every moment; being every moment recall'd to the consideration of ourselves and our present situation. 'Tis easily conceiv'd, that this interruption must weaken the idea by breaking the action of the mind, and hindering the conception from being so intense and continu'd, as when we reflect on a nearer object. The *fewer* steps we make to arrive at the object, and the *smoother* the road is, this diminution of vivacity is less sensibly felt, but still may be observ'd more or less in proportion to the degrees of distance and difficulty. (274)

Similarly, according to Hume, your feelings of sympathy for your relatives will be more pronounced than your feelings of sympathy for those to whom you are not related (cause and effect). Your mind is more easily carried from yourself to your relatives than it is from yourself to strangers.

With a basic understanding of Hume's account of sympathy before us, we are now in a position to consider the context in which Hume introduces the general point of view in the *Treatise*. The general point of view is Hume's attempt to answer an objection according to which moral feelings cannot be the result of sympathy. According to the objection, we can experience variations in our sympathy even when there are no corresponding variations in our moral judgments. If this is true, according to the objector, morality cannot be a function of our sympathy. Hume, as the objector, writes,

> When any quality, or character, has a tendency to the good of mankind, we are pleas'd with it, and approve of it; because it presents the lively idea of pleasure; which idea affects us by sympathy, and is itself a kind of pleasure. But as this sympathy is very variable, it may be thought, that our sentiments of morals must admit of all the same variations…. But notwithstanding this variation of our sympathy, we give the same approbation to the same moral qualities in *China* as in *England*. They appear equally virtuous, and recommend themselves equally to the esteem of a judicious spectator. The sympathy varies without a variation in our esteem. Our esteem, therefore, proceeds not from sympathy. (371).

Since the mechanism of sympathy depends upon the relation of ideas to enliven one's idea of someone else's feelings into an impression, we feel sympathy more strongly when the action is near us physically or socially. Resemblance, causation, and continuity in space and time each will have a greater enlivening effect when we are closer to the action in question (closer either physically or socially). The objector exploits this variation in our sympathy by contrasting it with our steadier moral judgments. Whereas we might be more pleased by benevolence in someone near to us (physically or socially) than by the same degree of benevolence in someone who is far away, we feel the same amount of moral approbation for the two. The objector concludes that, since our sympathy varies but our moral approbation does not, our moral approbation cannot be a function of our sympathy.

The objector attributes a certain kind of universality to moral approval and disapproval,[8] and Hume does not dispute this universality. The force of the objection comes from the perception that Hume's theory (centered, as it is, on our sentiments of sympathy) cannot explain the universality of moral reactions. True moral reactions are universal, in the sense that they are consistent *within* reactor as well as *between* reactors.

We can distinguish two kinds of universality that are implied (if only loosely) in the objection. First, we feel the same approval or disapproval of *everyone*, no matter how near or far. Hume writes, "…notwithstanding this variation of our sympathy, we give the same approbation to the same moral qualities in *China* as in *England*." In other words, our moral reactions remain constant no matter who is the object of our reaction, no matter

[8]I will refer to moral approval and moral disapproval as "moral reactions" for short. This term may not be completely accurate, but it is more accurate than "moral judgments", since moral approval and moral disapproval are feelings. "Judgment" would seem to imply that these feelings are the result of an exercise of our faculty of reason, which they are not. I will return to this issue in much more detail in section 2.

what is the relationship we have with the object of our reaction. For example, if we approve of benevolence in our friend, we will approve of the same amount of benevolence in someone we have never met, in a faraway country. The assumption of the consistency of moral reactions (within a reactor) by the objector is explicit in the objection, and in fact is at the heart of the objection.

A second kind of universality is that we *all* have these feelings and agree in our moral reactions of approval and disapproval. If I feel *moral* approval for benevolence in my friend, you will feel moral approval as well (upon considering his benevolence), even if he isn't your friend—even if he is in a faraway country.[9] This kind of universality is more loosely implied. The language chosen to express the objection, however, certainly suggests that we all agree in our feelings of moral approval and disapproval. He writes, "**we** are pleas'd with it, and approve of it," "**we** give the same approbation to the same moral qualities in *China* as in *England*," and "**our** esteem, therefore, proceeds not from sympathy" (bold emphases mine). His use of the first-person *plural* rather than singular suggests moral agreement.

Hume, we ought to note, seems to agree with the objector about the universality of moral reactions. Rather than dispute the underpinnings of this objection (the universality of moral reactions), he introduces the general point of view as an answer to the objection.

> Besides, every particular man has a peculiar position with regard to others; and 'tis impossible we cou'd ever converse together on any reasonable terms, were each of us to consider characters and persons, only as they appear from his peculiar point of view.

[9]As we will see later, this approval is only guaranteed under good conditions—that both of us, for example, have *successfully* adopted the general point of view. But for now, moral agreement is taken for granted.

He continues,

> In order, therefore, to prevent those continual *contradictions*, and arrive at a more *stable* judgment of things, we fix on some *steady* and *general* points of view; and always, in our thoughts, place ourselves in them, whatever may be our present situation. (371-2).

The introduction of the General Point of View captures the first kind of universality that Hume and the objector agree morality has, and makes the second kind of universality possible.[10] When we adopt the general point of view, we ignore the differences that distance and time can make to our non-moral reactions. From the general point of view, it does not matter how close to or far from someone we are. We adopt a point of view where we ignore the features particular to us, in our particular circumstances (in fact, where we ignore, or at least counterbalance, features particular to any one individual).

> In general, all sentiments of blame or praise are variable, according to our situation of nearness or remoteness, with regard to the person blam'd or prais'd, and according to the present disposition of our mind. But these variations we regard not, in our general decisions, but still apply the terms expressive of our liking or dislike, in the same manner, as if we remain'd in one point of view.

The point of view must be one where we ignore or counterbalance those particular features, because that is how it becomes *general*.

Hume's theory of abstract ideas can help us understand the details of how we adopt a *general* point of view. The groundwork that makes it possible is that we notice similarities in things and use the same word for all these objects.

> When we have found a resemblance among several objects, that often occur to us, we apply the same name to all of them, whatever differences we may observe in the degrees of their quantity and quality, and whatever other differences may appear among them (18-19).

[10]The second kind of universality is captured by the fact that we have feelings of *sympathy* from this general perspective, and all humans have sympathy. Therefore, we all have these feelings.

Once we have this custom, using the word for one thing leaves us disposed to call to mind some of the other things to which the word applies.

> But as the same word is suppos'd to have been frequently apply'd to other individuals, that are different in many respects from that idea, which is immediately present to the mind; the word not being able to revive the idea of all these individuals, only touches the soul, if I may be allow'd so to speak, and revives that custom, which we have acquired by surveying them (18-19).

However, we will not actually call these other objects to mind unless we have some need to do so or are prompted.

> They are not really and in fact present to the mind, but only in power; nor do we draw them all out distinctly in the imagination, but keep ourselves in a readiness to survey any of them, as we may be prompted by a present design or necessity. The word raises up an individual idea, along with a certain custom; and that custom produces any other individual one, for which we may have occasion. (19)

There are at least two ways, consistent with this theory of abstract ideas, that we could make our point of view general. We can abstract from the *features* of the situation, or we can abstract from the point of view itself. To elaborate: on the one hand, when we are considering a situation where there is a feature that is particular to our situation, we could substitute that feature for another comparable feature that would remove the bias. For example, if I discover that my friend has been laundering money, the fact that the perpetrator is my friend may be biasing me in my appraisal of the situation. To counteract this bias, I may imagine that someone who wasn't my friend was the perpetrator, and see how that would affect my appraisal. I will probably need to imagine several other people in the situation (both as perpetrators and as victims) to come to a truly general point of view. Substituting the other people for both perpetrators and victims in the situation will help me locate bias and counteract it, making my point of

view more general and less specifically mine. On the other hand, I could instead imagine the *points of view* of several other people who are not his friend. I could imagine the point of view from his acquaintances, his family, his enemies, and people who never met him or heard of him. I could further imagine the point of view from the friends, enemies, and acquaintances of the victim, as well as people who never met or heard of the victim. Presumably, this also is a way of counteracting bias and making the point of view more general and less specifically mine.

Let us return to the objection that sympathy cannot play the central role in Hume's system, since sympathy is variable and morality is not. The key to answering the objection that morality cannot be a function of sympathy is noticing that once we have adopted a more general perspective, we will have the same feelings regardless of distance. The feelings of sympathy *that are relevant to our moral reactions* (moral approval and moral disapproval) are not variable in the way the objector thought they were.

> Such corrections are common with regard to all the senses; and indeed 'twere impossible we cou'd ever make use of language, or communicate our sentiments to one another, did we not correct the momentary appearances of things, and overlook our present situation. (372)

The feelings of sympathy that are relevant to our moral reactions are those feelings that we have *after* we have taken up the general point of view, and those feelings are consistent across differences in distance and time.

> 'Tis therefore from the influence of characters and qualities, upon those who have an intercourse with any person, that we blame or praise him. We consider not whether the persons, affected by the qualities, be our acquaintance or strangers, countrymen or foreigners. Nay, we over-look our own interest in those general judgments; and blame not a man for opposing us in any of our pretensions, when his own interest is particularly concern'd. We make allowance for a certain degree of

> selfishness in men; because we know it to be inseparable from human nature, and inherent in our frame and constitution. By this reflection we correct those sentiments of blame, which so naturally arise upon any opposition. (372).

Hume introduces the general point of view because he recognizes that in order for his theory to be a theory about *morality*, our *moral* reactions must meet certain criteria. The general point of view is supposed to help us experience moral reactions that meet these criteria. Hume introduces the general point of view because he and the objector share as common ground the belief that a theory about *morality* must include and explain the two kinds of universality described above.

The universality of morality depends on the universality (which here takes the form of consistency) of our sympathetic reactions and the effects of the associations of ideas. Without the universality of sympathetic reactions (combined with the stable effects of the association of ideas), there would be no fact of the matter about what someone who had adopted a general point of view would feel. I may feel one reaction when I adopt the general point of view and consider a crime, and you may feel a different reaction when you adopt the general point of view and consider the same crime. Only the universality of sympathetic reactions can ensure that we feel the *same* reaction when we consider the same crime from the general point of view.

Let us now turn to the *Enquiry*. In the *Enquiry*, Hume refines and elaborates on the idea of the general point of view and the presupposition of universality in the moral domain. He turns his attention to the domain of concepts and language, and he explores the commitments that we adopt when we use the concepts and speak the language of morality. He writes,

> The notion of morals implies some sentiment common to all mankind, which recommends the same object to general approbation, and makes every man, or most men, agree in the same opinion or decision concerning it. It also implies some sentiment so universal and comprehensive as to extend to all mankind and render the actions and conduct, even of the persons the most remote, an object of applause or censure, according as they agree or disagree with that rule of right which is established. (74-5).[11]

Hume begins by claiming that *the very idea of morality* implies some universal sentiment. He then analyzes the implied universal sentiment as being universal in two ways. The sentiment is universal in the same two ways we saw earlier in the *Treatise* (although presented in reverse order here). The sentiment is universal in the sense that we all have feelings of moral approbation and disapprobation; furthermore, we have interpersonal *agreement* in our moral reactions.[12] The sentiment is also universal in the sense that our moral reactions are consistent through changes in the object of our reactions, despite variations that would affect our feelings of sympathy, such as variations in distance and time.[13] While you may *hate* the villain who robbed you more than I hate him, our moral disapproval is *equal*.

His comments in the Enquiry are interestingly ambiguous between expressing reactions and expressing judgments. When I call someone "virtuous", I could be expressing moral approval (approval from the general point of view), or I could be making a judgment that the person meets the criteria for having a virtue (on Hume's

---

[11] David Hume, *An Enquiry Concerning the Principles of Morals*, ed. J. B. Schneewind. Hackett Publishing Company, Indianapolis, Indiana, 1983.

[12]This is what he means when he writes that the sentiment "recommends the same object to general approbation, and makes every man, or most men, agree in the same opinion or decision concerning it." This is the universality that was only *loosely* implied in the language of the objection in the Treatise.

[13]This is what he means when he writes, "It also implies some sentiment so universal and comprehensive as to extend to all mankind and render the actions and conduct, even of the persons the most remote, and object of applause or censure, according as they agree or disagree with that rule of right which is established." This is the universality that was at the heart of the objection in the Treatise.

theory, that would be that he has a quality that is useful or beneficial, or immediately agreeable, to himself or others). On the first understanding, I would be *expressing* certain feelings, whereas on the second I would be performing (and reporting) a certain cognitive act. For simplicity's sake, I will refer to these quite different acts under the joint term "evaluations" (and, therefore, intend to use "evaluation" in a way that does not privilege its use to refer to judgments as opposed to sentiments).

It would be a mistake to think that the sentiment of humanity that Hume employs in his explanation of morality in the *Enquiry* is just a different word for the sentiment of sympathy as used in the *Treatise*. Rather, the feelings of humanity are a subset of the sympathetic sentiments he wrote about in the *Treatise*. The feelings of humanity are those sympathetic sentiments that are *universal*. He writes, about sentiments of humanity,

> But when he bestows on any man the epithets of *vicious* or *odious* or *depraved*, he then speaks another language, and expresses sentiments, in which, he expects, all his audience are to concur with him…. He must move some universal principle of the human frame, and touch a string, to which all mankind have an accord and sympathy. (75)

The feelings of humanity he describes in the *Enquiry* are not sympathetic responses to the plight (or good fortune) of another person. Sympathetic responses, as we learned, can vary according to factors such as proximity to another person in distance and time. Rather, these feelings are responses to someone's *character* and *behavior*. Our response is approval, if the character or behavior tends to be useful or immediately agreeable to one's self or to others; disapproval, if the character or behavior tends to the opposite (and, of course, there is a third possibility, a neutral response, when the character or behavior is neither positive nor negative on these dimensions). *These* feelings do not vary according to proximity in the way our sympathetic responses do. He says,

> But the sentiments, which arise from humanity, are not only the same in all human creatures, and produce the same approbation or censure; but they also comprehend all human creatures; nor is there any one whose conduct or character is not, by their means, an object, to every one, of censure or approbation. (75)

As we saw in the *Treatise*, feelings of sympathy do not need to be universal. In fact, it was just that feature of sympathetic sentiments that led to the introduction of the general point of view in the first place. However, in the *Enquiry*, Hume writes that the distinction between the sentiments of humanity and other sentiments is that the sentiments of humanity are *universal*, whereas all other sentiments are not[14].

> Whatever conduct gains my approbation, by touching my humanity, procures also the applause of all mankind, by affecting the same principle in them: But what serves my avarice and ambition pleases these passions in me alone, and affects not the avarice and ambition of the rest of mankind. (76)

Here, again, Hume shapes his moral theory to accommodate the kinds of universality that must be present in any moral theory. However, rather than taking our moral reactions as a starting point, as he seems to in the *Treatise,* here he begins with our moral language. He begins by exploring moral language, and finds that our moral language implies universality. Moral language and moral concepts themselves imply something universal.[15]

Hume thinks that morality implies a *universal* sentiment because when we speak using moral language, we speak having adopted the general point of view, and we expect that others have adopted this point of view as well. As such, when we use the language

[14]The *capacity* to have feelings of sympathy is universal, even in the Treatise, but any given sympathetic feeling could fail to be universal, unless the individual in question had already adopted the general point of view.

[15]Of course, for Hume, a universal *sentiment* is implied. However, for now, let us focus primarily on the universality.

of morality to express our moral reactions, and even our moral judgments, our language implies an expectation that everyone will agree and a belief of ourselves that we are not influenced by our own particular circumstances.[16]

He writes, describing our language when making evaluations outside of the domain of morality,

> When a man denominates another his *enemy*, his *rival*, his *antagonist*, his *adversary*, he is understood to speak the language of self-love, and to express sentiments, peculiar to himself, and arising from his particular circumstances and situation (75).

Hume continues immediately to describe the language involved in *moral* evaluations,

> But when he bestows on any man the epithets of *vicious* or *odious* or *depraved*, he then speaks another language, and expresses sentiments, in which, he expects, all his audience are to concur with him (75).

For many *non-moral* evaluations, such as evaluations of what one most wants or evaluations about what would be best for one (prudential evaluations) the point of view adopted is simply that of the speaker. For evaluations about what one wants, whatever feelings strike the speaker, at that moment, in that particular mood, in that particular context, count as relevant to his evaluation. For prudential evaluations, we are more strict about what counts as relevant, although the point of view is still that of the speaker—but considerations about long-term ends, and the relative strengths of various

[16]Jensen offers a similar interpretation. He writes that Hume "…provides clear outlines of a theory of moral language. According to this theory, to employ moral language is to be committed to setting restrictions on what we are prepared to regard as a moral term or a moral judgment. Moral language may not be used to express purely private sentiments occasioned by some particular circumstance at some particular time. Instead, moral judgments must be universalizable, they must concern *everyone* and must concern a *kind* of act" (502). Jensen, however, also claims that we cannot get substantive moral judgments from the requirements of moral language, and I disagree. I argue that moral judgments are judgments about what feelings we would have from the general point of view. More on this later. Henning Jensen, "Hume on Moral Agreement", *Mind* New Series, v.86:344, pp. 497-513.

desires will play more of a role in this kind of evaluation. However, again, the point of view is only the speaker's point of view.[17]

For *moral* evaluations, however, the point of view adopted must be the *general* point of view, a point of view where we ignore (or at least counterbalance) the features particular to that speaker (namely, ourselves), in our particular circumstance—in fact, we ignore the features particular to any one individual. The point of view must be one where we ignore those particular features, because that is how it becomes *general*. Once a person speaks from the general point of view, he can engage the moral sentiments of others, who can also look upon the situation without bias from features particular to themselves.

We have seen that Hume's answer to the objection in the Treatise is to argue that moral reactions and judgments imply two kinds of universality, and that therefore when we make such judgments or experience such reactions, we make them having already adopted the general point of view. I have argued that Hume identifies two kinds of universality that are implied by moral language and moral evaluations. I have further argued that the general point of view is Hume's way of accommodating the universality of morality.

Section 2: Cohon's objection

Let us turn to an issue that some commentators have seen as a tension in Hume's view. These commentators interpret the general point of view as a hypothetical construct, and moral evaluations are evaluations about what we would feel from the

[17]We can make judgments about what another most wants, or what is most prudential for another person. In this case, the point of view adopted will be that of the subject of our consideration. So, if I make a judgment about what is most prudential for Jim, then in making that judgment, I attempt to adopt his point of view, and consider his long-term ends and the relative strengths of his desires.

general point of view. If the general point of view is a hypothetical construct, however, then there seems to be a tension between those passages in which Hume claims that reason has no role in moral evaluations and those passages in which Hume says that moral evaluations must be made from the general point of view. Since, *prima facie*, judgments about what we would feel from a point of view that is explicitly not our own seem to require the use of our faculty of reason, reason on this interpretation would seem to have quite a large role in moral evaluations. Furthermore, if Hume is understood as claiming that reason has no role in moral evaluations, we are left with a question about how to understand the role of our faculty of reason in the context of the general point of view. This tension raises questions about how much of Hume's seeming non-cognitivism or antirationalism[18] must be abandoned in light of his comments about the general point of view. If morality "consists not in any *matter of fact*, which can be discover'd by the understanding" (301), then how can we maintain that moral judgments are judgments about how we would feel (which judgments we presumably could not make without our faculty of reason)? Hume says, "morality…is more properly felt than judg'd of," (p.302). If this is true, how could judgments about morality be possible at all?

Cohon articulates this problem clearly and compellingly:

> Moral evaluations become inductive empirical beliefs about what we would feel if we really occupied the imagined common point of view, and hence the deliverances of causal reason; this contradicts Hume's claim that the making of a moral evaluation is not an activity of reason but of sentiment.

[18]These are two distinct interpretations. If we read Hume as a traditional non-cognitivist, then when he writes that "virtue and vice are not matters of fact" (301), we take that to mean that all moral discourse is just an expression of a sentiment, and not referring to external moral facts. So, if I say that some action is virtuous, and you say that same action is not virtuous, we do not disagree, since it is entirely consistent for me to have a sentiment of approval and for you to have a sentiment of disapproval. If we read Hume as an anti-rationalist, then the thesis is somewhat weaker—there can be moral facts, but we cannot learn of them via our faculty of reason.

Any explanation or interpretation of Hume's use of the general point of view must respect certain central passages in which Hume claims that morality is the product of sentiment and not reason. In addition to the preceding passages, consider:

> But can there be any difficulty in proving, that vice and virtue are not matters of fact, whose existence we can infer by reason? (301).
>
> To have the sense of virtue, is nothing but to *feel* a satisfaction of a particular kind from the contemplation of a character. The very *feeling* constitutes our praise or admiration…. Our approbation is imply'd in the immediate pleasure they convey to us (303).

If virtue and vice are determined by feelings and sentiments that can only be felt from a particular point of view (a point of view, it might seem, that is explicitly *not* our own), then it would seem that our moral judgments are the result of reason. Cohon writes,

> The resulting account seems to treat moral judgments as cognitions, specifically, beliefs (frequently counterfactual ones) about what someone or anyone would feel if she occupied a point of view close to the person being evaluated. This would make moral evaluations into inductive, empirical beliefs, presumably based on past experience of the effects of people's character traits on themselves and their closest associates. Such beliefs are, for Hume, the deliverances of (causal) reason (833).

In these passages, Hume is quite explicit that our faculty of reason cannot be the source of any "knowledge" of morals. It *seems*, therefore, that if Hume introduces the general point of view as a tool for evaluating traits of character and actions, used by our faculty of reason, then he will have gone back on his earlier claims that reason has no role in moral evaluation. The interpretive task at hand, then, is to explain the role of the general point of view within Hume's system without contradicting any of his earlier claims.

We saw in the preceding section that, for Hume, moral evaluations concern our feelings from the general point of view. Moral approval is the approval we feel toward

certain characters or behaviors after we have taken up the general point of view, and moral disapproval is the disapproval we feel when regarding other characters or behaviors from the general point of view. It is that feature, in part, that differentiates *moral* evaluations from, for example, prudential evaluations.

If we understand the general point of view not as a hypothetical construct, but instead as an actual point of view that we can and do adopt, we do not need to view morally relevant feelings as the "deliverances of causal reason," as Cohon says. Rather, we should regard them as actual feelings we have when we look at a situation from a point of view that is not biased in the way our own point of view usually is. As we saw earlier, this is not a foreign experience—we do this in our daily lives. I can and do make prudential evaluations for other people by looking at something from their perspective. We plead to others, "but look at it from my point of view." There is no reason to think we cannot similarly adopt a "general" point of view for moral evaluations.

However, the tension Cohon and other commentators discuss raises another important issue which should not be quickly dismissed. As Cohon pointed out, there are places in the *Treatise* where Hume seems to say that reason plays no role in morality. For our account to be complete, we must account for these passages. Later in this section, I will use one of Hume's central arguments to illustrate why he does not actually mean that reason plays *no* role in morality. It is true that Hume does clearly rule out reason in *some* facets of moral evaluation. I will argue, however, that for Hume, we do use our faculty of reason to make some moral evaluations. On my interpretation of Hume's view, we use our faculty of reason to make moral judgments.

Let us first recognize that there are two quite different features of moral experience in Hume's account. On the one hand, there are moral *reactions*—these are the sentiments of pleasure, or approval, and displeasure, or disapproval, one feels when reflecting on another person's character or behavior. We feel these sentiments of pleasure or displeasure, approval or disapproval, derived from sympathy, upon the survey of some character or action. These sentiments are, for Hume, *primary* moral experiences. On the other hand, there are moral *judgments*: judgments that an action is right or wrong or that a quality of character is virtuous or vicious.[19] These are our judgments that something is wrong because it *would* meet with disapproval (or our judgment that something is right because it *would* meet with approval) from the general point of view. This sort of judgment is secondary in a way, because the evidence in favor of such a judgment can only be, for Hume, our (or others') primary moral experiences.[20] Our moral reactions form the basis of all of our moral judgments and experiences. That is, our moral judgments are *about* our (actual and hypothetical) moral reactions, and the only evidence we can have in favor or against moral judgments are moral reactions, both our

[19]On Hume's account, the judgment that an action is wrong or that a character trait is a vice is the judgment that it (action or trait) would meet with disapproval from the general point of view—the judgment that it is right or a virtue is the judgment that it would meet with approval from the general point of view. Of course, someone other than Hume could judge that something is wrong using different criteria than the ones Hume himself presents, while still maintaining the central role of the general point of view in moral evaluation.

[20]We should here distinguish between moral reactions and what I will call psuedo-moral reactions, even though Hume does not explicitly make this distinction. Moral reactions are reactions that are had under *appropriate* circumstances—that is, from a general point of view. Psuedo-moral reactions are reactions that we take to be moral reactions, but, unbeknownst to us, they are not actually had under appropriate circumstances. For example, if I feel I have been slighted, I may believe that I have successfully adopted the general point of view, but I may still be biased in my indignation. In this case, judgments that are made when I believe I have adopted the general point of view but have not are *psuedo-moral*. Both moral reactions and psuedo-moral reactions are taken as evidence for our moral judgments, but only moral reactions *determine* virtue and vice, on Hume's view.

own and that of others.[21] It is in this sense that our moral reactions are *primary* moral evidence.

Resolving the textual tension involves seeing that when Hume argues that morality is the domain of sentiment and not reason, he is claiming that our *moral reactions* are not the product of reason. When he writes that morality is more properly felt than judged, he means that sentiments (our moral reactions) are our primary moral evidence. Consider the following quotations:

> Thus the course of the argument leads us to conclude…[that] it must be by means of some impression or sentiment [vice and virtue] they occasion, that we are able to mark the difference between them…. Morality, therefore, is more properly felt than judg'd of… (302).
>
> An action or sentiment, or character is virtuous or vicious; why? Because its view causes a pleasure or uneasiness of a particular kind. In giving a reason, therefore, for the pleasure or uneasiness, we sufficiently explain the vice or virtue…. To have the sense of virtue, is nothing but to *feel* a satisfaction of a particular kind from the contemplation of a character (303).
>
> We do not infer a character to be virtuous, because it pleases: But in feeling that it pleases after such a particular manner, we in effect feel that it is virtuous (303).

In each of these passages, Hume is describing our primary moral experiences, our moral reactions.[22] He is *only* concerned with arguing that our moral reactions are not the product of reason, and not at all concerned with arguing that our faculty of reason cannot make judgments *about* moral reactions.

[21]It's important to keep in mind that anyone's moral reactions can be evidence for my moral judgments. While my own moral reactions are the most easily accessible evidence for me, the moral reactions of others are also relevant to my judgment that something would or would not be approved of from a general point of view.

[22]I will consider one of his *arguments* in detail later.

In fact, it is difficult to see what reason there is to doubt that our faculty of reason could make judgments *about* moral reactions. Cohon and other commentators go wrong by failing to appreciate fully the difference between reactions, on the one hand, and evaluations in the sense of judgments, on the other, and the significance of the distinction for Hume's theory. Hume does think that morality is not *constituted* by reason, but that is not enough to make him a noncognitivist. He leaves room for us to make judgments about moral *reactions*—just so long as our moral reactions are not perceived as themselves being the result of a process of reasoning.

Hume goes on to suggest that we *can* form judgments on the *basis* of our moral reactions. If it is true that we can form judgments on the basis of our moral reactions, then there is not actually any tension between the role of the general point of view and Hume's remarks about the impossibility of our knowledge of morals being the product of our faculty of reason. The judgments we make about how we would feel from this general perspective are secondary moral *judgments*, and do not clash at all with his claims that morality (i.e. our moral reactions) cannot be the product of our faculty of reason.[23] For example, he writes,

> We blame equally a bad action, which we read of in history, with one perform'd in our neighborhood the other day: **The meaning of which is**, that we know from reflection, that the former action wou'd excite as strong sentiments of disapprobation as the latter, were it plac'd in the same position (373; bold emphasis mine).

[23]There are also the feelings we have after having adopted the general point of view. These are *moral* sentiments, our moral reactions, and are not judgments about how we would feel if we did adopt the general point of view. These are very important pieces of primary moral evidence. My task here is just to defend the possibility of judgments about how we would feel from the general point of view in cases where we have not actually adopted the general point of view.

When we judge that the historical action is equally as bad as the recent action in our neighborhood, the *content* of that judgment essentially includes that we and others do and would experience equally strong sentiments (moral reactions) *from* the general point of view. Sometimes the general point of view is actually adopted (this is one way to get evidence in favor of the judgment); at other times, we deduce what someone in the general point of view *would* feel. Either way, the evaluation/judgment relies on our primary moral reactions. Again, while he is explicit that our moral reactions cannot be the result of a process of reasoning, he has not taken the extra step that would make him a noncognitivist—namely, claiming that we cannot make *judgments* about morality using our faculty of reason.

To make it even more clear that this is what Hume means, we can turn our attention to one of the arguments where he purportedly argues that reason has no role in morality. Take the argument in the following passage, which he claims "proves *directly* that actions do not derive their merit from conformity to reason, nor their blame from a contrariety to it" (295). Examining this argument will help support our earlier claim, which is that Hume does think there is a role for reason in moral evaluation, and that Hume is not a noncognitivist, as Cohon and others suppose. The examination of this argument will show on Hume's view exactly what role reason can play in moral judgment. He writes,

> Reason is the discovery of truth or falsehood. Truth or falsehood consists in an agreement or disagreement either to the *real* relations of ideas, or to *real* existence and matter of fact. Whatever, therefore, is not susceptible of this agreement or disagreement, is incapable of being true or false, and can never be an object of our reason. Now 'tis evident our passions, volitions, and actions, are not susceptible of any such agreement or disagreement; being original facts and realities, compleat in themselves, and implying no reference to other passions, volitions, and actions. 'Tis

> impossible, therefore, they can be pronounc'd either true or false, and be either contrary or conformable to reason (295).

We should take care to distinguish two separate claims here, both of which Hume seems to argue for, although it is not entirely clear which he has in mind at any given time. On the one hand, he argues that *our faculty of reason* cannot be the source of our knowledge about morals. On the other hand, he argues that r*eason* (but not our faculty of reason) does not determine what is virtuous and what is vicious (i.e. there is not a real relation of ideas that settles the virtue/vice question).

Let me say a bit more about what I take the distinction between *reason* and the *faculty* of reason to be. Reason seems to provide the standard at which our *faculty* of reason aims—so that our faculty of reason is operating properly when it reasons along the lines that reason demands. Reason determines the *real* relations of ideas, whereas our faculty of reason attempts to conform to reason, such as by identifying those real relations and discovering real existence and matters of fact.[24] Reason can be identified as what makes a line of reasoning successful, and it is our faculty of reason that carries out this line of reasoning.[25]

Let us return to Hume's argument. The structure of this argument is difficult to follow, but I think it goes something like this: Hume's strategy is to emphasize that we can know what can be conformable to reason by learning what can be an object of our faculty of reason. If something is conformable to reason, then it is something that can be

[24]For example, if I make a bad inference (e.g. from "P->Q" and "Q", I infer "P"), the faculty of reason is operating improperly—because reason would not license the inference.

[25]Contrast "reason" with "right reason" or "reason when functioning successfully. The latter two cases describe a situation where the faculty of reason is operating successfully—and we can say it is operating successfully because of a conformity to reason.

an object of our faculty of reason. If it is not conformable to reason, then it cannot be an object of our faculty of reason.

> Whatever, therefore, is not susceptible of this agreement [to *real* relations of ideas, or to *real* existence and matter of fact], is incapable of being true or false, and **can never be an object of our reason** (295, bold emphasis mine).

1. Passions, volitions and actions are in principle not conformable to reason (because they are not truth-apt).
2. If passions, volitions and actions are in principle not conformable to reason, and if therefore there is no relation of ideas that serves to license or forbid inferences about vice and virtue, then reason cannot be what determines what is virtuous and vicious. For Hume, because there is no relation of ideas that serves to license or forbid these inferences, passions, volitions and actions cannot derive their merit from conformity to reason.
3. Therefore, reason does not determine what is virtuous and what is vicious.

> 'Tis impossible, therefore, they can be pronounc'd either true or false, and be either contrary or conformable to reason.

4. Therefore, our faculty of reason cannot give us any knowledge of virtue and vice.

Hume begins this argument by looking at what can (and cannot) conform to reason. Since the objects of moral evaluation (passions, volitions and actions) cannot conform to reason, their "merit" cannot derive from such conformity. Therefore, we cannot learn *of their merit* by using our faculty of reason. While we can learn *something* about morality using causal reasoning, (and Hume's moral theory is an example of using causal reasoning to learn something about morality!), we cannot learn the *merit* of anything—*that* can only be felt.

Hume's clear primary aim in this argument is to show that reason does not license or forbid inferences about virtue and vice (at least, inferences where the conclusion of the argument is that someone is virtuous or something is vicious). To see this, consider his summary of the argument's accomplishments as follows:

> [This argument] proves *directly* that actions do not derive their merit from a conformity to reason, nor their blame from a contrariety to it; and it proves the same truth more *indirectly*, by showing us, that as reason can never immediately prevent or produce an action by contradicting or approving of it, it cannot be the source of the distinction betwixt moral good and evil, which are found to have that influence (295).

Hume concludes, therefore, that passions, volitions, and actions do not *derive their merit* from conformity to reason.

Hume thinks that by proving that reason does not determine virtue and vice, he will prove that our faculty of reason cannot be the source of our knowledge about virtue and vice. Consider the following two claims:

1. Our faculty of reason is the source of our knowledge about virtue and vice
2. Reason determines what is virtuous and what is vicious (i.e. that there are relations of ideas that license or forbid judgments about what is virtuous and what is vicious).

I have argued that, for Hume, the first claim would derive from and depend on the second claim. For Hume, showing the second claim to be false will itself show the first claim to be false. In this argument, the *only reason* we have been given for thinking that the first claim is false is that reason does not determine virtue and vice. Hume has not given us an independent reason for thinking that our knowledge of virtue and vice does not come from our faculty of reason.

Once we see that the only reason we have to think that our faculty of reason does not give us knowledge about virtue and vice is that reason does not determine virtue and vice, we can see that Hume's conclusion is only that our faculty of reason will not provide access to *primary moral evidence.* However, this argument does not at all demonstrate that our faculty of reason cannot aid us in making *secondary moral judgments* (for which the primary moral evidence would be our sentiments, our moral reactions). Commentators, such as Cohon, who read Hume as being a noncognitivist, seem to mistake him to be making the latter claim in addition—that our faculty of reason cannot aid us in making secondary moral judgments. However, as we have seen, the text seems to suggest otherwise—that Hume's only concern is to prove that our faculty of reason cannot provide us with primary moral evidence, *not* that our faculty of reason cannot aid us in making secondary moral judgments.

Hume argues that our *moral reactions* cannot be discovered by using our faculty of reason. He does not claim that we cannot use our moral reactions as evidence in moral judgments. Consider the following passage:

> Nor does this reasoning only prove, that morality consists not in any relations, that are the objects of science; but if examin'd, will prove with equal certainty, that it consists not in any *matter of fact*, which can be discover'd by the understanding. This is the *second* part of our argument; and if it can be made evident, we may conclude, that morality is not an object of reason.

Hume here (similarly to the argument discussed above) argues *from* his claim that morality is not determined by reason, *to* the claim that we cannot use our faculty of reason to discover moral facts. But, as we have just seen, this sort of argument is best understood as an argument that our moral reactions cannot be discovered via our faculty

of reason, and *not* as an argument that our faculty of reason has no place in morality (say, in making secondary moral judgments).[26]

I have argued that the supposed tension between certain of Hume's claims (such as when he claims that morality is felt, not judged) and the introduction of a general point of view can be resolved by making a distinction between primary moral evidence and secondary moral judgments. The passages that Cohon cites seem, on a first pass, to be in tension with the cognitivist implications of the general point of view, but the distinction I articulated helps to relieve this tension. Specifically, the passages she cites in which Hume argues that our knowledge of morals cannot be the product of our faculty of reason are best understood as passages in which Hume is concerned with arguing that our *moral reactions* (which are our primary moral evidence) cannot be known via our faculty of reason. Hume seems to be quite comfortable with the possibility that we sometimes reason *about* our moral reactions, and he certainly at times suggests that we can reason about how we would feel from the general point of view.

This resolution to the problem can gain traction when we notice that Hume compares morality to things like colors, which are plausibly understood dispositionally.

> If, in the sound state of the organ, there be an entire or considerable uniformity of sentiment among men, we may thence derive an idea of the

[26]Cohon herself offers an interpretation of Hume that bears similarity to the interpretation I offer above. She argues that Hume offers what she calls a "moral sensing" view, on which there are two types of moral perception. She distinguishes between moral impressions and moral ideas. Moral impressions (similar to what I call moral reactions) are *felt*, whereas moral ideas (similar to what I call moral judgments), which have cognitive content, are not felt. She writes, "But one crucial passage shows him to believe that moral properties are essentially reaction-dependent properties: they depend for their existence on the emotional responses of sensitive beings. This makes him an ethical anti-realist in our sense….But his anti-realist view is perfectly compatible with truth-cognitivism" (7). We disagree, however, in our interpretation of the *content* of what she calls moral ideas and what I call moral judgment. She argues that moral ideas cannot be *causal* beliefs, i.e. beliefs that a property would cause a certain reaction in a person, for the textual reasons we have already discussed. However, I think that in the relevant passages, Hume is talking about *moral impressions* or *moral reactions*, and so they are not evidence that moral *judgments* or *ideas* cannot have causal content.

> perfect beauty; in like manner as the appearance of objects in daylight, to the eye of a man in health, is denominated their true and real colour, even while colour is allowed to be merely a phantasm of the senses. (Standard of taste.)
>
> Vice and virtue, therefore, may be compar'd to sounds, colours, heat and cold, which, according to modern philosophy, are not qualities in objects, but perceptions in the mind… (301).

If our knowledge of whether a given object is red or not depends on our knowledge that it appears red in good conditions to competent observers, then it seems clear that we can make secondary judgments about colors. Our primary evidence for colors, of course, would be our own experiences and sensations of color, but we could also make judgments about whether the object appears such-ly in such circumstances. In that case, such judgments would be the product of our faculty of reason.

If I am right that Hume is discussing our moral reactions in the passages that Cohon finds troubling, then these passages are not any evidence at all that an interpretation that emphasizes the role of the general point of view is in conflict with key passages in the *Treatise*. I have argued that we should interpret Hume as distinguishing between moral reactions and moral judgments. While Hume certainly argues that our moral reactions cannot be the result of any process of reasoning, there is no reason to suppose that we cannot *use* our reactions in a process of reasoning to make moral judgments. My interpretation helps show that Hume is not a noncognitivist, and therefore that there is no tension between the general point of view and the passages where he argues that morality is not an object of reason.

Section 3: Moral Motivation

Cohon's worry that the introduction of the general point of view is incompatible with Hume's claims that our moral knowledge is not the product of our faculty of reason

can naturally be extended to his arguments that morality must be motivating. If Hume's general point of view turns moral knowledge into a product of reasoning, then moral knowledge cannot be motivating. Yet, Hume is explicit that morality *must* be motivating—that is his reason for claiming that moral knowledge is *not* a product of reasoning. He argues that Reason and our faculty of reason are not the kind of thing that can motivate us, and yet since morality does motivate us, morality cannot depend on Reason (and our faculty of reason). He writes,

> Since morals, therefore, have an influence on the actions and affections, it follows, that they cannot be deriv'd from reason; and that because reason alone, as we have already prov'd, can never have any such influence. Morals excite passions, and produce or prevent actions. Reason of itself is utterly impotent in this particular. The rules of morality, therefore, are not conclusions of our reason (294).

He also writes, later,

> As long as it is allow'd, that reason has no influence on our passions and actions, 'tis in vain to pretend that morality is discover'd only by a deduction of reason.

Any interpretation of Hume's use of the general point of view needs to respect this argument and these passages. Any interpretation must explain the fact that morality must *motivate* us, and reason cannot do so.

The distinction between moral reactions and moral judgments can help us make sense of these passages. Our moral reactions, our moral sentiments, can motivate us. Our moral reactions, remember, are our primary moral evidence. On the basis of our primary moral evidence, we can, from time to time, make (secondary) moral judgments, judgments about what moral reactions we would have, judgments about what sentiments we would have from the general point of view. These judgments would not serve to motivate us, except insofar as we actually *feel* these sentiments, these moral reactions.

The introduction of the general point of view, then, does not require supposing that verdicts of reason will motivate us. It is our own primary moral reactions (sympathetic sentiments) that actually motivate us.

This reading of Hume's arguments can help us make sense of some passages that otherwise would remain puzzling. In some of these passages, Hume acknowledges that our sentiments and our reason can conflict. He acknowledges that sometimes our sentiments do not agree with our moral judgments. For example, he writes,

> Experience soon teaches us this method of correcting our sentiments, or at least, of correcting our language, where the sentiments are more stubborn and inalterable (372).

Similarly, he also writes,

> But however the general principle of our blame or praise may be corrected by those other principles, 'tis certain, they are not altogether efficacious, nor do our passions often correspond entirely to the present theory (372).

These passages seem to be in tension with his arguments that morality cannot be derived from reason, since reason cannot motivate. However, once we distinguish between moral (and pseudo-moral) reactions, on the one hand, and moral judgments, on the other hand, the tension disappears. Hume explains that our moral judgments cannot always bring our sentiments under control. By way of getting us to take up the general point of view, they can influence our sentiments to some extent, but not in ways strong enough to overpower our sentiments from our first-person points of view. Clearly, on his view, our moral judgments cannot motivate us, since they cannot control our sentiments. However, we can still make them—they are the product of our faculty of reason—and the possibility of our making them does not contradict his claim that morality must motivate us. Consider also the following passage:

> Here we are contented with saying, that reason requires such an impartial conduct, but that 'tis seldom we can bring ourselves to it, and that our passions do not readily follow the determination of our judgment.

Here, he seems to suggest that it is *best* when our passions and sentiments conform to our moral judgment (that is, when things are going as they are supposed to), but that sometimes they don't conform. Sometimes we continue to feel rogue pseudo-moral sentiments, sentiments that are not had from the general point of view. We can continue to feel them, even when our considered moral judgment conflicts with them.

These kinds of passages make clear that we must be careful in how we interpret his arguments that, since reason cannot motivate, morality cannot be derived from reason. In these passages, he clearly suggests that sometimes our moral judgments are correct even in the face of contradictory sentiments and passions. The way to resolve the tension between these passages and earlier passages about the motivational power of morality lies in the distinction between moral reactions and moral evaluations. Our moral (and pseudo-moral) reactions have motivational power, and our moral (not pseudo-moral) reactions determine virtue and vice. However, we can still make (secondary) moral judgments about what is virtuous and what is vicious, using our moral reactions as evidence. In these cases, these moral judgments will not have motivational power.

What motivation do we have to adopt the general point of view? We don't have *motivation*, per se. Rather, according to his theory of abstract ideas, we are compelled to adopt it whenever we make a judgment that potentially conflicts with the other ideas that comprise the general idea at issue. He writes,

> …after the mind has produc'd an individual idea, upon which we reason, the attendant custom, reviv'd by the general or abstract term, readily suggests any other individual, if by chance we form any reasoning, that agrees not with it. Thus shou'd we mention the word, *triangle*, and form

> the idea of a particular equilateral one to correspond to it, and shou'd we afterwards assert, *that the three angles of a triangle are equal to each other*, the other individuals of a scalenum and isosceles, which we over-look'd at first, immediately crowd in upon us, and make us perceive the falshood of this proposition, tho' it be true with relation to that idea, which we had form'd. (19)

The process is mostly automated, unless our faculties are not operating properly:

> If the mind suggests not always these ideas upon occasion, it proceeds from some imperfection in its faculties; and such-a-one as is often the source of false reasoning and sophistry.

In the moral case, the process will proceed similarly. For example, if I claim that the person who stole a bagel from my son is as evil as the person who stole valuable medicine from an elderly person in need, my mind will naturally think of other people who have stolen similar items (other baked goods, from retailers other than my son), and I will realize that I am unwilling to claim that the robbers in these other cases are as villainous as the person who stole the medicine from an elderly person. Our mind naturally engages in a sort of "challenge-response" whenever we react to a situation and make moral judgments about it. The "challenge" leads to the "response", which just is taking up the general point of view.[27,28]

[27]The connection between the reaction and the judgment, for Hume, is less clear, and we need the judgment in order to be compelled to enter the general point of view (for it is the judgment that propels us into the abstraction of the general point of view). My suspicion is that, for Hume, the answer will come down (as it often does) to the associations of ideas. For example, if I observe someone stealing from my friend, and I have the sympathetic reaction of anger, I will therefore be caused to judge that the thief is bad. The association of ideas will connect anger with bad people, since by and large only bad things tend to cause anger.

[28]Cohon offers her own explanation of our motivation to adopt the general point of view. Her view is not necessarily in conflict with this view, and in fact the two may complement each other. On her view, we adopt the general point of view because, since our moral judgments form the basis of other, important, causal and predictive judgments, we need our moral judgments to form a stable basis so that our predictions will be more likely to be true. "Whenever we make moral evaluations we also make objective, causal judgments about the love and hatred, pride and humility that the trait in question will produce. Because of the way we depend on this information, it is very inconvenient to have false judgments of this kind, and for different people to make mutually contradictory ones. So it is important to us to insure that they be true and consistent. If we make our causal judgments from the common point of view, they tend to be true, and tend

Section 4: Conclusion

I will conclude by highlighting the main points I have argued in this paper. The first, and most controversial, position I have taken is to argue that Hume is a cognitivist, not a noncognitivist, about ethics. The passages that are generally taken as demonstrating that Hume is a noncognitivist are passages in which he allegedly argues that reason has no place in morality. However, I make a distinction between moral reactions and moral judgments, and further I argue that this distinction is important in Hume's theory. If this distinction is granted, I have argued, the purported noncognitivist passages can instead be read as passages in which Hume denies that morality is *determined* or *constituted* by reason. This leaves plenty of room in his theory for making judgments *about* morality, which he elsewhere seems to endorse. Furthermore, since there is no inconsistency between a cognitivist moral theory and a moral theory in which morality is not determined by reason (for example, utilitarianism is, in many of its instances, such a theory), there should be perceived no inconsistency here.

Relatedly, I have carefully explored the role of reason and our faculty of reason in Hume's moral theory. I have argued that for Hume, at the most basic level, there is no role for reason. Morality is constituted by (a subset of) our sentiments, our moral reactions, and not at all by the dictates of reason. However, reason and our faculty of reason still have a role to play in morality—namely, the role of making moral judgments.

---

to be consistent over time and between persons, whereas if we make them without adopting the common point of view they tend to be sometimes false and often mutually contradictory" (846). The only hesitation I have about this sort of explanation is whether it is putting the cart before the horse. Since we appear to need a judgment *before* we can enter the general point of view (since it is the *challenge* to the judgment that allows us to abstract and therefore enter the general point of view), it is unclear whether we can make sense of Cohon's claim that we enter the general point of view *in order to make* judgments.

Furthermore, I have offered an interpretation of the general point of view according to which Hume introduces it because of his commitment to the importance of universality in morality. The role of the general point of view itself in his moral theory depends on the universality of human feelings. Without the universality of human feelings, there would be no consistency in moral reactions made from the general point of view. The interpretation I have offered draws on his theory of abstract ideas. According to this interpretation, we can actually adopt the general point of view and the sentiments we feel having done so are our moral reactions. When we fail to adopt the general point of view, we can instead deduce what we would feel from that point of view, if we so choose.

Finally, I have argued that my interpretation, an interpretation according to which Hume is a cognitivist, an interpretation according to which the distinction between moral reactions and moral judgments plays a significant role, dovetails nicely with Hume's remarks on moral motivation. Hume is explicit that morality must be motivating, and any successful moral theory must meet that requirement. Moral reactions, which are the sentiments we feel from the general point of view, are motivating, on my interpretation of Hume, thereby meeting his requirement. The distinction between moral reactions and moral judgments allows Hume's moral theory to have one foot on the reason side and one foot on the motivating/sentimental side.

Kant and the Universal Law formulation of the Categorical Imperative

Kant is less explicit than Hume regarding the general point of view, but he is explicit about *universality*. Kant thinks that the content in the first formulation, the universal law formulation, is derived from the very concept of a categorical imperative, from the very concept of a moral law. Furthermore, he thinks, the concept of a moral law essentially involves universality. The categorical imperative is a universal law, and as such it applies to all of us. If the categorical imperative applies to all of us, then we must only act in accordance with maxims that can be willed to be universal law, for if the maxim cannot be willed to be universal law, then it does not apply to all of us and therefore cannot be a principle of morality. But notice: this is the universal law formulation: 'Act only on that maxim by which you can at the same time will that it should become a universal law' (4:222).[1]

I will argue that, for Kant, the very concept of a categorical imperative is the concept of a *universal* law. But I also think that the universal law formulation is Kant's way of capturing the general point of view. We test our maxims by making sure that they could be willed to be universal law—that is, by making sure that they can apply to *everyone*, and not just us. But this test involves adopting a general or impartial point of view, in a way. It involves ignoring my own aims, inclinations, and desires, and instead evaluating the maxim of my action from a general perspective. Worthy maxims will be

[1] Immanuel Kant, *Groundwork for the Metaphysics of Morals*, trans. Arnulf Zweig, ed. Thomas E. Hill Jr., Arnulf Zweig. Oxford University Press, 2002.

those that I choose, not from my own perspective (keeping in mind my own goals and inclinations), but from a more general perspective, one that involves evaluating them for adoption *universally*.

Some critics have objected that Kant's argument that the Universal Law Formulation (FUL) is a formulation of the Categorical Imperative relies on an invalid inference.[1,2] The objection tends to hinge on the accusation that Kant confuses, on the one hand, the question of what is rational for an agent to will for *himself,* and on the other hand, the question of what is rational for an agent to will for *everyone*.

I believe, however, that by distinguishing carefully between what Kant wants to accomplish in the universal law formulation and what he wants to accomplish in the Humanity Formulation (FH), we can interpret Kant in such a way that he does not make the mistake these commentators attribute to him. I will argue that Kant does not expect the general point of view, as expressed in the universal law formulation, alone to pick out all and only worthy maxims. Instead, we must also rely on a universal end—humanity—and this insight is expressed in Kant's humanity formulation. In the end, I hope we will be able to say that for Kant as for Hume, the general point of view (as expressed in the universal law formulation) is a way of capturing universality, which is essential for any proper system of morals.

---

[1]E.g. Geoffrey Scarre, "Interpreting the Categorical Imperative," *British Journal for the History of Philosophy*, v.6:2, pp. 223-236, and Allen Wood, "Kant on the Rationality of Morals", *Ottawa Congress on Kant in the Anglo-American and Continental Traditions*, 1976, pp.93-110. Thomas E. Hill Jr. alludes to this difficulty in "Kant's Argument for the Rationality of Moral Conduct," Pacific Philosophical Quarterly, v.66, pp. 3-23. See also Henry E. Allison, "On a Presumed Gap in the Derivation of the Categorical Imperative", *Philosophical Topics*, v.19:1, pp.2-8.

[2]While over the course of the *Groundwork*, Kant argues for several propositions about FUL, I will focus on his claim that FUL formulates the supreme principle of morality, an argument he gives in the discussion in ch. 2 (4:221-222).

In this paper, I will consider one such representative argument, that given by Wood in "Kant on the Rationality of Morals." I will begin, in Section I, by offering a "bare-bones" account of Kant. This bare-bones account will be useful to help us navigate between the differing interpretations. I will go on, in Section II, to discuss Wood's objection in great detail. His objection is quite compelling, and presents quite a challenge for those who wish to defend Kant. In Section III, I will discuss an interpretation that O'Neill offers in "Universal Laws and Ends-in-Themselves",[3] and show how her interpretation can help us make headway on this issue. While I will argue that her interpretation is subject to some difficulties, which I will discuss, her interpretation is a useful first step towards extricating Kant from these difficulties. In the last section, I will present the interpretation that I believe succeeds in rescuing Kant from Wood's objection, and yet sidesteps the difficulties that O'Neill's interpretation faced.

Section 1: Schematic overview of the argument

In this section, I will offer a schematic review of the argument that FUL formulates the supreme principle of morality, in the *Groundwork for the Metaphysics of Morals*, ch.2 (4:221-222). My intention here is to review the argument in a way that is general enough to avoid controversy. Having a schematic review of the argument will help us understand the differences between competing interpretations of Kant, as well as, much later, help us articulate how we can rescue Kant from Wood's objection.

The categorical imperative, Kant tells us, is to be contrasted with the hypothetical imperative. According to the hypothetical imperative, we are rationally required to pursue the means to our ends. While the necessity of conforming one's actions to a

[3] Onora O'Neill, "Universal Laws and Ends-In-Themselves," *Monist: An International Quarterly Journal of General Philosophical Inquiry*, v.72, pp. 341-361.

hypothetical imperative depends on one's commitment to the end so acting would obtain[4], the necessity of conforming one's actions to the categorical imperative is not in this way conditional. The categorical imperative, Kant writes, is the one imperative "which commands a certain line of conduct directly, without assuming or being conditional on any further goal to be reached by that conduct" (4:217). The categorical imperative binds us, *whatever* ends we have.

After introducing the concept of a categorical imperative, Kant proceeds to offer one formulation of it, the universal law formulation (FUL): "Act only on that maxim by which you can at the same time will that it should become a universal law" (4:222).[5] Kant thinks that the content in FUL is derived from the very concept of a categorical imperative. In his argument for FUL, he expresses this idea (that the content in FUL is derived from the very concept of a categorical imperative) by saying,

> The first part of our task is to see whether perhaps the mere concept of a categorical imperative might also give us the formula containing the only proposition that can be a categorical imperative (4:221).

And he continues,

> But if I think of a *categorical imperative*, I know right away what it contains. For since this imperative contains, besides the law, only the necessity that the maxim conform to this law, while the law, as we have seen, contains no condition limiting it, there is nothing left over to which the maxim of action should conform except the universality of a law as such; and it is only this conformity that the imperative asserts to be necessary (4:222).

He concludes, "There is therefore only one categorical imperative and it is this: 'Act only on that maxim by which you can at the same time will that it should become a universal

[4]It depends on one's commitment to the end, since the end can be given up.

[5]And his variation, the universal law of nature formulation, is "Act as though the maxim of your action were to become by your will a universal law of nature" (4:222).

law'" (4:222). Let us note that this is the main passage from which Wood draws his interpretation of Kant's argument. Therefore, rather than interpreting the passage in detail here, we will discuss this argument in detail over the course of this paper.[6]

In order to understand Kant's claim that we must act only on maxims that can become universal law (or, what is the same thing, can conform to universal law), we must first understand what it would *take* for a maxim to conform to the law, or to become a universal law. A maxim is the principle on which an agent acts. In a footnote, Kant writes,

> A *maxim* is a subjective principle of action and must be distinguished from an *objective principle*—namely, a practical law. The former contains a practical rule determined by reason in accordance with the conditions of the subject (often his ignorance or his inclinations); it is thus a principle on which the subject *acts*. A law, on the other hand, is an objective principle, valid for every rational being; and it is a principle on which he *ought to act*—that is, an imperative (4:222).

A worthy maxim, i.e. a maxim upon which it is morally permissible to act, is one in which the agent non-accidentally conforms to universal law. But unworthy maxims could be anything from "I will steal this doughnut because I am hungry" (selfish maxim) to "I will replace her speech notes to thwart her career because it gives me pleasure" (evil maxim). A maxim where one does the right action for the wrong reason ("I will rescue this baby from drowning for fame, honor and possibly a monetary reward") is also unworthy. According to what Kant says in the argument for FUL, the categorical imperative contains "the necessity that the maxim conform to the law," (i.e. the categorical imperative expresses the necessity that our maxims can be willed as universal law— since the moral law applies to all rational beings, if our maxims cannot be willed

[6]However, my focus will be on the *universal* aspect of this argument. Several other difficulties, while important, are outside the scope of this paper. See, e.g. Hill's "Analysis of Arguments," 121-122.

as universal law, then these maxims will not conform to moral law,). From this we can conclude that a maxim is morally permissible to act on if it is consistent with objective principles.[7] Our maxims, then, need to be maxims that are valid for every rational being in order to be morally permissible to act on.[8] Therefore, the categorical imperative asserts that we are to act only on maxims that can be willed to be universal law.

There are two ways that a maxim could fail to be able to be willed to be universal law. The first way has to do with the coherence of the result when the maxim is willed to be universal law: A maxim, when willed to be universal law, may result in a contradiction in conception (or a contradiction in nature, which is arguably the same). "Some actions are so constituted that we cannot even *conceive* without contradiction that their maxim be a universal law of nature, let alone that we could *will* that it *ought* to become one" (4:225). Kant thinks his examples of suicide and false-promising fail in this way. The idea, roughly, is this: Some maxims, if they were adopted as a universal law of nature, would result in a contradiction in the *system* of nature. If we cannot conceive of a world in which it is universally permissible to act according to that maxim, then it is not morally permissible for an agent to act according to that maxim.

The second way has to do with the *willing*: A maxim, when an agent attempts to will it to be universal law, may result in a contradiction in the *will* of the agent. In this kind of case, the will is committed to two opposing ends, and so when the agent wills for a maxim promoting one of these ends to be *universal*, while the will is still committed to

[7]I interpret maxims in such a way that in general we are testing whether they are universally permitted, although of course if we are never permitted not to act accordingly, then they would be universally required! See, e.g. Hill, 69.

[8]What "valid for all rational beings" consists in will vary significantly according to the interpretation of Kant. For Wood, "valid for all rational beings" will turn out to mean something like, "all rational beings have the same necessary ground to act according to the principle."

the other end, the result is a contradiction in the will. "In the case of other actions, we do not find this inner impossibility, but it is still impossible to *will* that their maxim should be raised to the universality of a law of nature, because such a will would contradict itself" (4:225). Kant thinks the maxims in his examples of failing to develop a talent and failing to help others in need fail in this way. So, perhaps the maxims in these cases *could* be adopted as a system of nature, but the agent could never *will* that they be adopted as a system of nature because it would result in a contradiction in the agent's will. If it is not possible for a rational being to will that the maxim should become a universal law (because that rational will would contradict itself due to conflicting necessary ends), then it is not morally permissible for an agent to act according to that maxim. [9]

Section 2: Wood's objection

We will soon see that Wood objects to Kant's argument that FUL is a formulation of the Categorical Imperative on the grounds that it is invalid (101). We will see that his argument is that the conclusion of this argument (that the categorical imperative asserts that we are to act only on maxims that can be willed to be universal law) is not warranted by the premises. Wood thinks that there could be some maxims that are valid for each rational being, and yet for which it is not rational for an agent to will as universal law. It is one thing, he claims, for a maxim to be valid for each rational being, and quite another for it to be universalizable (for it to be rational for an agent to will it to be universal law). For example, he argues, it is rational for an egoist to adopt the principle of self-love

[9]This discussion of the argument has been vague on purpose. Wood objects to this argument, and his objection will help us to introduce more precision into our understanding of Kant's argument.

(always act to promote your own happiness) as a maxim, but it is not rational for him to will that all agents adopt the principle of self-love as a maxim. He will conclude that the argument is fallacious.

Before we consider Wood's interpretation of this argument, we must first pause to consider some of his (and his understanding of Kant's) terminology. Reasons and grounds seem to be synonyms, for Wood. A ground is just a reason, and although Wood does not elaborate on his understanding of reasons, his argument does not appear to depend on a controversial understanding of reasons. A principle is meant to guide our action. Maxims, which are the subjective principles of the will, are the principles we *actually* act on.[10]

For Wood, objectively grounded principles are particularly crucial in Kant's argument. Objective principles, or objectively grounded principles, are the principles we *should* act on (i.e. we all have sufficient reason to act on them).[11] An objectively grounded principle is a principle which all rational beings have a sufficient ground (a sufficient reason) to act on—and it must be the *same* ground for all rational beings. This is the same, for Wood, as saying that it is universally valid.[12] It must also be a necessary ground, i.e. an *a priori* necessity. A principle on which I have one reason to act, and for

[10]For Wood, it is important that maxims articulate our grounds/reasons for action, although this may be controversial. For the purposes of this paper, however, I will grant the point.

[11]We act on an objective principle by bringing our subjective principle, i.e. our maxim, into accordance with the objective principle.

[12]But see p.104, where he alludes to an ambiguity. "Terms like 'universally valid' or 'universally rational' can be taken in either an individual sense or a collective sense. If I say that a principle is 'universally valid' or 'universally rational', this might mean that for every individual without exception, the supremely rational course of conduct for that individual consists of his following that principle. But it might also mean that any group of people in which the principle was universally followed would be a group whose behavior, considered collectively, is rational." Wood argues that this ambiguity leads to the fallacy in the argument we are considering.

which you have another ground, is not an objectively grounded principle, even if every rational being happens to have a ground to act on it.

I will reconstruct Kant's argument, as Wood sees it, here.

1. If there is a principle of morality, it is an objectively grounded principle.

   Wood writes, "What Kant wants to show about this principle of morality in his deduction of it is that it is a rational principle in the sense that every rational agent has some ground or reason for following it whatever the circumstances, and moreover a ground which is always rationally sufficient, decisive and overriding" (96).[13]

2. An objectively grounded principle is one which all rational beings have not only reason/ground to act on, but the *same* (necessary) reason/ground to act on.

   "A practical principle is objective, or a practical law, if and only if there is a ground for following it which is necessarily valid for every rational being as such" (101).

3. A maxim is *objectively* grounded (i.e. all agents have the same reason/ground to act on it; i.e. it is universally valid) if and only if it conforms to the *concept* of a principle which is objectively grounded.

   Wood writes, "But a ground of this kind cannot consist in anything except the agreement of the agent's maxim with the *concept* of a principle which is universally valid" (101).

   In other words, a maxim conforms to law if and only if that maxim is objectively grounded; a maxim conforms to universal law if and only if we all share the *same* necessary reason/ground to act on it.

4. A maxim is objectively grounded if and only if each rational being possesses a ground (a necessary ground, the *same* necessary ground) for acting on it.

[13]Here, Wood is speaking of Kant's project as a whole (including Kant's attempt to prove that there *is* a Categorical Imperative, in ch. 3), and he (Wood) is not specifically talking about the argument for FUL. Later, speaking of the argument for FUL as the first stage in a two-part project (the second stage being the attempt to prove that there is a Categorical Imperative), Wood writes, "First, Kant argues that the principle of morality is the only principle which agrees with the concept of an objective ground, so that this principle must be the foundation of every principle which is objectively grounded" (98).

This is intended as an attempt to fill out premise 3. What it takes for a maxim to conform to the concept of a principle which is objectively grounded is for each rational being to have the same necessary ground for acting on it.[14]

5. Therefore, a maxim is objectively grounded if and only if I can will it to be universally followed.

Wood thinks the move from 4 to 5 (the conclusion) is fallacious. In 4, the claim that a maxim is objectively grounded consists in claiming that I (a rational being) and *every other* rational being possess a ground for following that maxim (the ground is that the maxim conforms to law). In 5, the claim is that *I* have a ground for willing that *all* rational beings follow that maxim. Wood writes,

> The universal validity of an objective principle therefore consists solely in the fact that each rational being possesses a ground (the same ground) for following it. But it does not necessarily consist, as Kant seems to infer, in the rational desirability (for someone, presumably for anyone at all) of the state of affairs which would result if everyone did follow this principle.

And he continues,

> To put the matter more simply, Kant appears to be arguing from the premise that it is rational for *each* person to follow a certain principle, to the conclusion that it must be rational (for some person, or any person) to will that *everyone* follow this principle. But this argument is not valid. For the fact that it is rational for each person to follow a certain principle does not, by itself, permit us any conclusion at all about the rational desirability of the state of affairs which would result if everyone followed it (101).

The assertion is that there could (hypothetically, even) be maxims that all rational beings have grounds to follow, yet there are no grounds for willing these maxims to be universal

[14]Furthermore, Wood thinks that the objective ground is that all rational beings have a ground to act on the principle/maxim. It is hard to see how to understand this claim so that it is not circular or infinitely regressive. However, since Wood's objection can be made without a commitment to this claim, I have set this issue aside.

law. And the reason that there could be maxims which met one condition but not the other is that conceptually the conditions are different.[15]

Wood goes on to illustrate his objection with a counterexample. If the inference from 4 to 5 is valid, Wood thinks, then there ought to be a contradiction in claiming of *any* principle *both* that it cannot be willed universally *and* that it is an objectively grounded principle. And in his objection, Wood aims to present a case where there is not a contradiction in so claiming (and therefore the inference from 4 to 5 will be shown to be invalid, if the case is successful). The idea is as follows: if the inference from 4 to 5 is legitimated by the meanings of the concepts involved (and according to Wood, Kant thinks it is), then any principle that meets the condition that it is an objectively grounded principle *must (conceptually)* meet the condition that it can be willed universally (and vice versa). And there is a contradiction, therefore, in supposing that there could be a principle that meets one of these conditions but not the other.

Wood's counterexample is as follows: suppose there is an egoist, who claims that the principle of self-love, "Always seek your own happiness," is an objective principle. He claims that it is an objective principle on the grounds that every rational being has the same (necessary) ground to follow the principle of self-love, "Always seek your own happiness." This egoist, however, cannot will this principle to be universally followed. He thinks "if everyone acted as (in the egoist's opinion) he has an objective ground for

[15]Scarre nicely summarizes Wood's objection as follows: "…Kant may simply have confused two quite different theses about practical reasoning. The first of these theses is that rational agents will abide only by principles that are 'universalizable' in the very weak sense of being formally suitable for all rational agents; the second, that rational agents will abide only by principles that they are 'able to will' should be concurrently adopted by all other rational beings—i.e. principles that are universalizable in the stronger sense we have been considering so far" (234-5).

acting, the result would be the frustration of the egoist's own desire for happiness, along with everyone else's" (102). Given that the result would be the frustration of the egoist's own desire for happiness, the egoist agrees, in the end, that the principle of self-love, "Always seek your own happiness," cannot be willed as universal law. However, Wood claims, so far, we have been given no reason to think that the principle of self-love is not an objective principle.

The problem Wood is alluding to, in other words, is that the principle of self-love cannot be willed as a universal law, and therefore it is not the principle of morality. However, it is (by hypothesis) an objective principle. For any principle, there ought to be a contradiction, if Kant is right, in maintaining that it is an objective principle and yet is incompatible with the principle of morality. We *ought* to have a contradiction, Wood thinks, but we don't. The problem isn't just that we have a principle which is (supposedly) an objective principle, and yet one that cannot be willed as universal law. The problem, Wood thinks, is that, in this example, there is no *contradiction* in claiming that the principle of self-love is an objective principle. Wood writes,

> Kant's argument purports to show that the only possible objective ground, the only ground consistent with the *concept* of an objective ground, consists in the agreement of an agent's maxim with the principle of morality, or the possibility of willing that this maxim be universally followed. Hence if Kant's argument were valid, it would be impossible to maintain without contradiction that the principle of self-love is an objective principle, since the principle of self-love, as we have supposed, is inconsistent with the principle of morality and cannot be willed as universal law." (102-3.)

The point here is that the principle of self-love, Wood thinks, has been demonstrated to be in disagreement with the principle of morality, since the principle of self-love cannot be willed to be universally followed. Therefore, if we accept Kant's argument, it is a

principle that is seemingly not consistent with the concept of an objective ground (since it is not consistent with the principle of morality). But if that is true, then the principle of self-love cannot be an objective principle, for there would be a *contradiction* in maintaining that it is. However, Wood claims, there is no such contradiction, and so we have reason to reject the move from 4 to 5.

It is important for Wood that the egoist cannot will that this principle of self-love be followed universally. Wood's language suggests that it is the egoist's *desire* for happiness that prevents him from willing the principle of self-love to be universal law. He writes,

> The egoist concedes, then, that if everyone acted as (in the egoist's opinion) he has an objective ground for acting, the result would be the frustration of the egoist's own desire for happiness, along with everyone else's. The egoist must candidly admit, therefore, that he cannot will that everyone should follow the principle of self-love. (102)

In the remaining part of this section, I will first demonstrate why, on Kant's view, the egoist's *desires* would not be able to play this role (standing in the way of the universalization of the principle of self-love). I will then suggest how we can plausibly understand Wood's objection without this mistaken view of Kant (whether it is how Wood intended it or not).

On Kant's theory, desire-frustration cannot be the sole obstacle to a principle's being willed as universal law. We know that Kant thinks that if a maxim cannot be willed to be a universal law, the failure must be due to a contradiction (either in the will or in nature) that would result if it *were* willed universally. Kant writes,

> Some actions are so constituted that we cannot even *conceive* without contradiction that their maxim be a universal law of nature, let alone that we could *will* that it ought to become one. In the case of other actions, we do not find this inner impossibility, but it is still impossible to *will* that

> their maxim should be raised to the universality of a law of nature, because such a will would contradict itself (4:225).

But clearly there is no contradiction in willing something (universally or otherwise) that would frustrate my desires. For example, "I will that my desires be frustrated" is not contradictory in the least. Desires, for Kant, are not part of the will; therefore it is difficult to see how they could lead to a contradiction in the will. The contradiction can only result if I have adopted the object of what I desire as an *end.* This is because ends are set by the will—and therefore an end is the only kind of thing that can lead to a contradiction in the will.

However, even having adopted the object of a desire as an end is not enough to result in the contradiction in the will. It would be quite implausible of Kant to claim that all *contingent* ends must remain contradiction-free when evaluating whether a maxim could be willed universally. Contingent ends are the proper subject of *hypothetical* imperatives, not categorical imperatives. A frivolous example will help demonstrate the point. Suppose that I (as someone with a sweet-tooth) have adopted as an end that I will refrain from eating candy today, even when I desperately want to. If any end will be included in an evaluation of whether the maxim could be willed universally, given that I have adopted this end, I would be unable to will "I will eat candy when I want it" to be universally permitted, because that would result in a contradiction in my will. However, Kant surely does not want to say that it is *impermissible* for us to eat candy when we want it, just because I happen to have set as an end for myself that I will not eat candy.[16]

[16]There is still something wrong with me, however, something irrational! If I were to act on a maxim that, when willed to be universally followed, resulted in a contradiction in my contingent ends, I would be doing something irrational. Perhaps it could be said that I lack integrity. However, the accusation cannot be, if made solely on the contradiction in my contingent ends, that I violate the categorical imperative.

And, given the diversity of contingent ends among all rational beings, if all contingent ends were included in evaluations, we would have very restrictive limitations on action, and it is hard to see how in this case we could ever act without violating the categorical imperative!

Rather, for Kant, what matters for evaluating a maxim according to FUL is whether there are *rationally necessary ends* that would prevent the maxim from being willed as universal law. For example, in his discussion of the man who does not want to develop his talent, Kant writes, "*For as a rational being he necessarily wills* that all his powers should be developed, since they are after all useful to him and given to him for all sorts of possible purposes" (4:224, emphasis mine). Contingent ends that are not shared by all rational beings cannot form the basis of a categorical imperative.[17,18]

Let us now consider how we may understand Wood's counterexample so that it does not depend on desire-frustration or on a contingent end (and so that he may

[17]If all contingent ends must be considered in deciding which maxims could be willed as universal law, we would get different answers depending on who was testing the maxim. I, as someone who has willed not to eat candy, could not will that those who want candy shall eat it to be a universal law, whereas plausibly someone who has not willed to avoid candy could will that those who want candy shall eat it to be a universal law.

[18]This may seem controversial. For example, suppose I have a contingent end that not everyone could pursue—e.g. suppose my contingent end is to eat cookies (every cookie I see, even when it belongs to someone else). It seems plausible that my maxim ("I will eat every cookie I see, even when it belongs to someone else") would result in a contradiction in my will, since other people may try to take my cookies from me when they want them. In this case, my contingent end is what prevents my maxim from being willed to be universal law, and this may seem to be a counterexample to the position I have been arguing. But there is an important difference between this cookie case and the case I described earlier, where I adopt the end of refraining from eating candy, but this end does not thereby mean that eating candy cannot be willed to be universal law. It is the difference between the mere fact of a contingent end being an end preventing a maxim from being willed to be universal law (what I have been arguing against, using the candy as my example) and the contradiction in one's will when a person holds that *he* should be allowed to pursue a contingent end where no one else should be able to (the cookie case, where contingent ends *are* relevant). What's bring universalized in the cookie case is a maxim based on a contingent end, but the origin of the contradiction is not in the contingent end, but in the exception that the person is making for himself. In contrast, using the candy case, I have been arguing that the origin of the contradiction cannot be the contingent end itself.

plausibly be able to claim that the principle of self-love cannot be willed to be a universal law). By hypothesis, the principle of self-love ("Always seek your own happiness") is an objective principle. I will assume, as seems plausible, that the assumption behind the principle of self-love is that each person's own happiness is something that each rational being (by hypothesis, given that the principle of self-love is an objective principle) has the same ground for seeking.

The contradiction in willing the principle of self-love to be a universal law would be that *in doing what he has objective grounds to do* he thereby frustrates his own *ends* (he cannot achieve his ends, because in seeking them he would thereby frustrate them). By willing that *each agent* seek his own happiness, the egoist would thereby thwart his own end of seeking his own happiness. That contradiction is why the egoist cannot will the principle of self-love to be followed universally, and why it thereby cannot *be* the principle of morality. If, by trying to will what he has objective ground to will, he thereby frustrates the very thing he has objective ground to will, then it is not possible to will the principle of self-love to be universal law.[19] If Wood is correct in his interpretation of Kant, then he seems to have provided a successful counterexample to Kant's argument for FUL.[20]

[19]This example assumes that the principle of self-love is an objective principle on the basis of the objectivity of the end of happiness. However, it seems possible that the (hypothetical) objectivity of the end of happiness is better captured by some other practical principle. Perhaps there is another principle that is grounded in the end of happiness that actually *could* be willed to be universal law. Wood does not consider this possibility, but it seems relevant to the example.

[20]It is somewhat unclear whether Wood intends the violation of the categorical imperative to be a contradiction in the will of the egoist, or whether the result is a contradiction in conception/nature. The case's striking resemblance to Kant's example of the impermissibility of false promising suggests that it is a contradiction in conception/nature. However, Wood's own discussion, while not taking a clear stand, sometimes suggests that he thinks it is a contradiction in the will of the egoist. (E.g., Wood writes that the problem is the frustration of the egoists own desires.)

At this point, I think we need to step back interpretively. Wood thinks that Kant has taken advantage of an invalid inference, thereby undermining his argument for FUL. Wood argues that Kant has failed to recognize a difference between claiming that each rational being possesses a ground for following a maxim and claiming that I have a ground for willing that *all* rational beings follow that maxim. However, I believe that the source of the invalidity is in reading the argument as if it concerned our reasons or grounds for acting on maxims. As an alternative, I propose that we should view Kant's argument in FUL as solely concerned with the universal form of the categorical imperative. Kant introduces reasons and grounds when he introduces the Humanity Formulation, not when he argues for FUL. There are several benefits of this interpretation, not least of which is that it avoids reading Kant as taking advantage of an invalid inference. However, it is difficult to express this point in the language of Wood's interpretation. Therefore, in the next section I would like to delineate a different approach to understanding Kant's argument.

Section 3: O'Neill's interpretation

I want to propose in this section that if one interprets Kant in such a way that he is not concerned with the *grounds* that agents have for acting on FUL in his argument for FUL, then it is no longer natural to interpret Kant as failing to recognize the difference between a case where it is rational for *each* person to will a maxim to be universal law and a case where it is rational for each person to will that *everyone* will this maxim (and we can call these, as Scarre does, the weaker and the stronger senses, respectively, of universalizability). Wood mistakes Kant to be offering the stronger thesis in the conclusion of FUL because he takes Kant to be making claims about what agents have

grounds to do in his argument for FUL. We should focus our energies on interpreting Kant in such a way that he sidesteps this problem. Later in this section, I will consider O'Neill's interpretation, which I will argue succeeds in sidestepping the problem, even while it is subject to other difficulties.

At this point, then, I would like to step back briefly and reevaluate the assumption that Kant's argument in FUL involves any claims about what agents have *grounds* to do in his argument for FUL. Even Wood would agree that Kant does not explicitly use the word "ground" in that argument. Since Kant does not use the word "ground" (or reason, or any other obvious synonym) in the argument for FUL, the interpretive burden is (prima facie at least) on Wood. While Kant clearly takes grounds for action seriously in the argument for the humanity formulation, I will argue that in the FUL argument, for Kant, grounds are not yet relevant.

I will argue that Wood misinterprets the imperative as given in FUL when he takes it to be more content-ful than the preceding premise.[21] The conclusion of the argument to establish that FUL formulates the supreme principle of morality is not that all rational beings must have *grounds* to will that the maxim become universally followed. Kant's ultimate point, rather, is that the only kind of maxim that *could* conform to universal law (i.e. that could be necessarily valid for all rational beings) is a maxim that applied to all of us. The moral law is universal, and therefore any maxim that

[21]Remember, the preceding premise, to paraphrase Wood's interpretation, is "A maxim is objectively grounded if and only if each rational being possesses a ground (a necessary ground, the *same* necessary ground) for acting on it." The conclusion, to paraphrase Wood's interpretation, is "Therefore, a maxim is objectively grounded if and only if I can will it to be universally followed."

is not universal cannot express moral law. For Kant, we only begin to consider our grounds for acting when we have moved on to the second formulation of the categorical imperative, the humanity formulation. Our *grounds* for acting are not relevant when we are merely considering what conditions will allow us to discover which maxims have the proper form, a universal form, so that they can apply to all of us and therefore satisfy the most basic moral requirement, that moral requirements apply equally to all of us. The categorical imperative as given in FUL is meant to provide necessary (but not sufficient) criteria for the proper form of a maxim that conforms to universal law.

For now, let us consider an interpretation that will give us some insight as to how we can avoid interpreting Kant as making an invalid inference. The key to such an interpretation is to understand Kant's argument as proceeding in two stages, one given in his argument for FUL and one given in his argument for FH. The first stage is the stage where Kant discusses the necessary conditions for the proper form of maxims (he gives necessary conditions for determining which maxims are permissible, but not sufficient conditions). However, FUL is not by itself sufficient to determine moral law. As we have already seen, for Kant, rationally necessary ends are the only ends relevant for determining whether a maxim would result in a contradiction in the will. But until we know *what* ends are rationally necessary, we cannot know which maxims can be willed to be universal law. (We will learn later that the only rationally necessary end, according to Kant, is humanity.) FUL and FH are each necessary and jointly sufficient for determining which maxims are morally permissible.[22] It is in the second stage that we

[22]For an example of an opposing view, that the universal law formulation should "generate moral laws that are…independent of empirical premises such as…human ends" (p.403), see Richard Francis Galvin, "Ethical Formalism: The Contradiction in Conception Test," *History of Philosophy Quarterly*, v.8:4,

learn how these conditions can be applied, once we know what the rationally necessary end (humanity) is. But even in such an interpretation, it is important to be able to explain how these two formulations are complementary to each other.

O'Neill offers just such an interpretation of the Formulation of Universal Law (FUL) and the Humanity Formulation (FH, which she refers to as FEI, for "the Formula of the End-In-Itself"), in "Universal Laws and Ends-in-Themselves." Her interpretation emphasizes parallels between FUL and FEI/FH. FUL, on her view, is meant to help us find maxims that are compatible with the agency of all whom they would govern, and FEI emphasizes the value of the agency of all rational beings. While in the end I will argue that O'Neill's interpretation does not succeed, she makes great progress towards avoiding the objection that Wood gives, and so her interpretation will be instructive to us.[23]

Her argument can be understood as proceeding in two stages. We will first consider O'Neill's claim of equivalence between the FUL requirement that we be able to *conceive* that a maxim is universally permissible and the FEI requirement that we never treat rational beings as mere means. We will then consider her claim that the FUL requirement that we (as rational beings) be able to *will* that a maxim is universally permissible is equivalent to the FEI requirement that we treat other rational beings as ends.

October 1991. The disagreement here is over the degree to which we should expect support from FH to generate contradictions in our maxims and Galvin appears to think we do not need such support.

[23]I should note that in some of her other works, O'Neill seems less inclined to emphasize the parallels between the two formulations. See, e.g. Onora O'Neill, "Consistency in Action", *Kant's Groundwork of the Metaphysics of Morals: Critical Essays*, ed. P. Guyer. Rowman & Littlefield, Lanham, Maryland, 1998, pp.103-131. However, my aim here is to see how the lesson can be applied to help rescue Kant from Wood's objection.

O'Neill begins by considering the FUL's condition that we must be able to *conceive* that a maxim is universally permissible, if it is to conform to the categorical imperative. She writes,

> Maxims which violate strict duties are said to yield *contradictions in conception* if we try to universalize them: the very attempt to think of the maxim as universally adopted breaks down owing to some incoherence in the way the world would have to be if it was universally acted on (347).

For example, as she points out, a maxim of deceit is a maxim that would yield a contradiction in conception (or, what is arguably the same thing, a contradiction in nature). In order to deceive someone about x, that person must trust you (at least about x), yet permitting universal deceit would thereby make that sort of trust impossible (assuming that deceit would be frequent if permitted). Therefore, a maxim of deceit is the sort of thing that could not be universally permitted, for it would only prevent its own possibility if universally permitted.

O'Neill argues that violence is a similar case. She argues that it also is the sort of thing that cannot be permitted according to FUL, since any maxim that permits violence will result in a contradiction in conception/nature. She claims that violence against a victim undercuts the agency of the victim. Without agency, however, the victim is unable to will the maxims that permit the violence against them in the first place. Therefore, when we try to will maxims permitting violence as universal law, a contradiction in conception/nature results. She writes,

> However, both instrumental and brute violence undercut the agency of those whom they victimize. It is not merely that victims do not in fact will the maxims of their destroyers and coercers: they are deliberately made unable to do so, or unable to do so for some period of time. A test that demands action only on maxims that all can adopt will require that action not be based on maxims of victimizing (347).

Since a victim's agency is compromised (destroyed or damaged or restricted), the victim *cannot* will the maxim permitting the victimizing action, O'Neill claims. It is the very loss of agency that prevents the victim from willing the maxim to be universal law (agency is a precondition for exercising the will). And since the victim is unable to will the maxim, it straightforwardly follows that the maxim cannot be willed to be universal law.[24]

More generally, O'Neill believes that protecting agency is a precondition for universalizability. She writes,

> Self-sufficiency is an incoherent goal for finite rational beings; at most they can coherently aim to minimize their dependence on others. They cannot universalize maxims either of refusing to accept any help or of refusing to offer any help, since help may be needed for the survival of their agency (348).

An agent cannot universalize maxims that undermine his agency, due to resulting contradictions in his will. Pogge objects to this claim on the following grounds:

> But then, by the same token, I would have to will the use of all other (permissible) means, such as flattery or begging, if my ends should turn out to be unattainable without them. Yet this is patently false, as I can reasonably will to get along without flattery, begging, or unpaid help. The flaw in this argument is this: Even if it is necessary that agents have ends, it is not necessary that we have this or that *particular* end. Thus it is possible for us to renounce certain means, by resolving to pursue only ends attainable without them (210-11, N19).[25]

Of course, it is crucial to keep in mind that for Kant, it *is* necessary for us to have one particular end, namely, the end articulated in the humanity formulation: agency, rational nature, humanity. While any contingent end may be given up, we as agents cannot give

[24]O'Neill believes that this is a case where willing would result in a contradiction in conception, but I will argue later that this rather is a case where willing would result in a contradiction in the will.

[25]Thomas Pogge, "The Categorical Imperative", *Kant's Groundwork of the Metaphysics of Morals: Critical Essays*, ed. P. Guyer. Rowman & Littlefield, Lanham, Maryland, 1998.

up the end of protecting agency. O'Neill's point, I think, is that we may need others' help to remain agents. Pogge misses that the very survival of one's agency is not a merely contingent end that can be given up.

For O'Neill, maxims that permit compromising an agent's agency (by destroying it, damaging it, or restricting it) result in contradictions in *conception* when one attempts to will them to be universal law. Maxims that permit agents never to help other agents (maxims that permit what she calls "self-sufficiency") result in contradictions in the *will* when one attempts to will them to be universal law. O'Neill moves on to consider the parallel argument against victimization, using FEI/FH.

> In using FUL to test our maxims we check that those maxims could be acted on by all other agents; in using FEI to test our maxims we check that action on them disables no other agents from adopting them (354).

On her interpretation, treating another as a mere means (as opposed to treating him as an end) is to interfere with his agency. She writes,

> If we treat other agents as mere means, we do prevent, damage or restrict their agency…. It is not merely that we may act in ways to which they *do not* consent; we act on maxims to which they *could not* consent (353).

O'Neill emphasizes that treating another agent as a mere means thwarts that person's own *agency*. And thwarting his agency, as we saw, acts as a roadblock to willing a given maxim as a universal law. To see the parallel between FEI/FH and FUL, we only need to see that when we treat another as a mere means, and act on maxims to which he could not consent, we act on maxims that would result in a contradiction in conception/nature when willed to be universal law. She writes,

> To use another as a mere means, as Kant sees it, is to act on a maxim which the other *cannot* also adopt. This amounts to acting on a maxim that one *cannot* at the same time will as universal law (353).

And to fail to treat another as an end, on this view, commits one (rationally) to self-sufficiency—it is to attempt to act on maxims that permit failing to help other agents. Such a maxim would result in a contradiction in the will when willed to be universal law, since, for O'Neill, we cannot rationally will to be self-sufficient.

The force of Wood's counterexample derives from the alleged *compatibility* of the propositions, 1) The principle of self-love is an objective principle and 2) The principle of self-love is incompatible with the principle of morality. Kant's argument implies that the two should be contradictory, so if they are compatible then Kant's argument has failed. While we are not quite in a position to defend Kant against this counterexample yet, we have the material to see how, using O'Neill's interpretation, we can begin to develop a contradiction between 1) and 2).[26] While 1) and 2) may not seem, *prima facie* to be contradictory, we know that for Kant, there must be a *reason* why the principle of self-love is not compatible with the principle of morality. Understanding *why* the principle of self-love is not compatible with the principle of morality is what will generate a contradiction between 1) and 2). On O'Neill's interpretation, the reason that self-love is not consistent with the principle of morality must be something like, "We must never damage another's agency, and we must help the agency of others when we can." Agency is the objective end, the rationally necessary end, on her view. The principle of self-love, however, requires choosing actions that further one's own ends, not choosing actions on the grounds that they further another's ends. But if this is true, then

[26]And from the get-go, we can see that she is opposed to Wood's style of interpreting Kant: "[FUL] is often misconstrued as a claim that morally worthy maxims must be ones that we are willing, i.e. want to see universally adopted…. This may not be such deep died heteronomy as a Utilitarian pursuit of maximal satisfaction of desires; but it is heteronomy nonetheless, and Kant rejects it decisively" (346). See also her article "Consistency in Action", pp.103-131. Wood's interpretation is not necessarily heteronymous, but at the least it does make a similar mistake—that whether a maxim can be willed universally depends on what is in the *agent's* best interest.

self-love (happiness) is not an objective end, and therefore the principle of self-love is *not* an objective principle.[27] Developing a contradiction between 1) and 2) requires simply showing *why* 2) is true. Understanding why 2) is true will show us why 1) must be false.

Let us put the point in a slightly different way. We begin with the following two claims, which Wood claims are not contradictory:

1) The principle of self-love is an objective principle

2) The principle of self-love is not compatible with the principle of morality.

We can then ask the following question: Why is the principle of self-love *not compatible with* the principle of morality? Why can it not be willed to be universal law? O'Neill's answer will be that agency, *not* happiness, is the objective end, the end that is rationally necessary for all rational beings. Only maxims that are compatible with treating agency as an objective end can pass the universalizability test. Furthermore, we can plausible say that if agency is the objective end, then happiness cannot be the objective end.[28] Therefore, if agency is the objective end, then the principle of self-love is *not* an objective principle.

Let us return, with a critical eye, to O'Neill's argument that maxims that rely on victimizing are incoherent when universally permitted.[29] It is true that the actions taken

[27]I assume here that there can be only one unqualified objective end. Kant must have had something like this in mind, though, for if there are two unqualified objective ends, they potentially could conflict, which would undermine Kant's claim that we must always act to promote the objective end, and never undermine it. Therefore, I assume, if agency is an objective end, happiness thereby cannot be an objective end.

[28] Again, I am assuming here that there can be only one unqualified objective end.

[29]I do agree with O'Neill that violence cannot be universally permitted, and that Kant has good reason to maintain that violence is not permitted according to the categorical imperative. However, my reason for thinking that (for Kant) violence cannot be universally permitted is that humanity, rational nature, agency, is an objective end. Without knowing that humanity is an objective end, I do not see why it is *incoherent* to suppose that maxims permitting violence, i.e. maxims where the victimizer articulates his intention to compromise or destroy another's humanity or agency for his own gain or pleasure, are universally permitted. More on this point later.

by the victimizer, should he act on such a maxim, would undercut or destroy the agency of the victim. It is this fact, in part, that makes being victimized so frustrating and demoralizing. However, the loss of agency involved in victimization occurs only *after* the victimization has taken place. Remember, O'Neill's claim was that maxims involving victimization cannot be willed as universal law since the victim cannot *will* anything, including that the maxim be universal law (and therefore these maxims, when we try to universalize them, result in a contradiction in conception/nature). Her argument requires that the very loss of agency is what prevents the victim from willing the maxim to be universal law. Let us note that if it is *not* the loss of agency that prevents the victim from willing the maxim to be universal law, then the result would not be a contradiction in conception/nature. At most, we would have a contradiction in the will.

However, in cases of victimization, the victim's agency is undercut or destroyed only once the maxim has been acted on. There is no incoherence (even if there is implausibility) in supposing that the victim could will the maxim to be universal law *before* his agency has been undermined.[30] As a contrast case, consider Kant's claim that maxims permitting lying would result in a contradiction in conception. In that case, the very possibility of lying depends on trust; however, there could be no trust in a world where lying was permitted.[31] Kant thinks that the very idea of a world where lying is permitted is for this reason incoherent. In the case of victimization, there is no such contradiction. The fact that agency would break down *later* (after the maxim has been

[30]Consider Sophie, in "Sophie's Choice", for example, to illustrate the point that there is no incoherence in wishing that one's agency be taken away. Conceivably, she could take steps to undermine her agency in order to avoid making the choice.

[31]I do not here mean to imply that *all* lying is impermissible. But where lying is impermissible, Kant's view seems to be that it is impermissible *because* lying in such situations would not be possible if everyone did it—because we would no longer trust each other in those situations, and lying depends on such trust.

acted on) is not sufficient to make the world in which victimization is permitted incoherent *unless* agency is an objective end. (Remember, we saw earlier that *contingent* ends are not sufficient to prevent a maxim from being universalizable. Only rationally required ends, that is, objective ends, are relevant when deciding which maxims can meet the conditions set by the universal law formulation.) But if it is true that there is no incoherence in a world where the victim wills the maxim to be universal law, then the contradiction must derive from the fact that agency is a rationally required end. It is a contradiction in the *will* of the victim that stands in the way of maxims allowing victimization to be willed as universal law.

In her discussion of FEI and imperfect duties, O'Neill writes that Kant thinks that agency is so fragile that merely forbidding actions that assault agency (treat other rational beings as mere means) is not sufficient to protect it. For her interpretation, agency is the key in connecting these two formulations. She explains,

> In using FUL to test our maxims we check that those maxims could be acted on by all other agents; in using FEI to test our maxims we check that action on them disables no other agents from adopting them (354).
>
> The equivalence of treating others as ends-in-themselves and of acting on maxims that can pass the contradiction in the will test is based on the fact that both principles express that agency be secured for all (355).

As we saw earlier, however, while acting on a maxim may disable another agent from consenting to that maxim, it is quite another thing to say that the maxim in question *could not* be consented to by that very agent. All it shows is that he could not consent to it at a particular time (namely, after he had been victimized).

The problem, ultimately, is that on O'Neill's account, we only succeed in ruling out maxims which permit violence or assaults on agency on the grounds that they result

in a contradiction in the will, not that they result in a contradiction in conception/nature. If my claim that there is no incoherence (even if there is implausibility) is supposing that the victim could will the maxim to be universal law *before* his agency has been undermined is true, then we must believe that the reason a maxim allowing victimization cannot be universalized is because it results in a contradiction in the will of an agent when that agent attempts to will it to be universal law. We can *conceive* of a maxim allowing victimization (by those who are willing) being universal law—what prevents it from passing the test given by FUL is that we cannot will it without contradicting the rationally required end of protecting agency, our own and that of others.

In this section, we saw that one way around Wood's argument may be to see that the FUL and the FH formulations are importantly different—with FUL focusing on the *form* of appropriate maxims, and FH focusing on the development of an objective end. Once we interpret Kant in such a way that the grounds are not given in FUL, but come later in FH, we can see that Kant's argument will not be invalid in the way Wood supposes. O'Neill's argument was an interpretation very much in this spirit, and as such it made headway on how we should understand Kant in order to circumvent the objection. However, because of details in O'Neill's argument, I concluded that her interpretation ultimately did not succeed. In the next section, I will offer an interpretation of Kant that will build on her account while also sidestepping her difficulties.

Section 4: Interpreting Kant

We can take O'Neill's interpretation as a model for avoiding Wood's objection, but in doing so it will be important to understand Kant in such a way that we avoid the main difficulty facing her interpretation. In this section, I will outline my own

interpretation. It is my hope that this interpretation will provide us with the necessary equipment to illustrate how Kant can avoid the difficulties posed by Wood's objection. I will take a significant cue from O'Neill, while also trying to improve upon the progress she has made. I will start from the beginning, so that it is clear at each step exactly what I take Kant to be doing.

As we saw earlier, Kant gives two conditions that a maxim must meet in order for that maxim to conform to law, i.e. in order for that maxim to conform to the categorical imperative as expressed in FUL. The first condition is that we must be able to conceive of a world in which it is universally permissible to act according to that maxim.[32] "Some actions are so constituted that we cannot even *conceive* without contradiction that their maxim be a universal law of nature, let alone that we could *will* that it *ought* to become one" (4:225). Kant thinks his examples of suicide and false-promising fail in this way. The idea, roughly, is this: Some maxims, if they were adopted as a universal law of nature, would result in a contradiction in the *system* of nature. If we cannot conceive of a world in which it is universally permissible to act according to that maxim, then it is not morally permissible for an agent to act according to that maxim.[33] For Kant, the

[32]Again, I assume here that maxims are universally permitted, although they would be required if the maxim's negation was never permitted.

[33]One might wonder what exactly makes such a maxim fail—is it that *A)* it could not *be* a law of nature? Or, is it instead that *B)* we could not *will* it to be a law of nature, *because* we could not conceive of it being a law of nature? In other words, the question is about what, ultimately, the source of the failure is—is it the incoherence of the possibility that it *be* universal law, or does the failure reside in *our* inability to conceive that it could be universal law. I don't have the space here to consider this issue in great detail, but it will suffice to gesture at the possibilities. To B), the interpretation that we could not *will* it to be a law of nature without a contradiction in conception, one might object that someone could, out of ignorance, fail to realize that willing it universally would result in a contradiction of nature. Kant could reply by claiming that there are conditions of success in conceiving, or even actually willing, and that in these cases, we are not *actually* conceiving or willing universally, but we only think we are. This strategy could be more plausible, or less, depending on the details, but the important thing to note is that *even if the reason the maxim fails is because we cannot will it to be a universal law of nature*, the *reason* we cannot will it (or conceive it) to be so is still because it could not *be* a law of nature. (The only other natural response is that we *do* will the maxim to be

distinction between, on the one hand, cases concerning maxims that, willed universally, result in a contradiction in the system of nature or in conception and, on the other hand, cases concerning maxims that result in a contradiction in the will (and we will discuss the latter sort of cases shortly) lines up with the distinction between strict or narrow (perfect) duty and wide or meritorious (imperfect) duty. "We see readily that the first kind of action is opposed to strict or narrow duty, the second opposed to wide (meritorious) duty" (4:225).

The point here, contra Wood, is that *if we cannot conceive of a world where such a maxim is universally permissible, then it cannot be universally permissible in this world* (since, of course, this world is not only conceivable but actual). It would be inconceivable for this maxim to be universally permissible, by hypothesis. Again, the moral law applies universally to all rational beings. If a maxim is such that it *cannot* apply universally to all rational beings, then that maxim *cannot* conform to the law. Wood argued that Kant failed to recognize the difference between a case where it is rational for *each* person to will a maxim to be universal law and a case where it is rational for each person to will that *everyone* will this maxim. But on this interpretation, there is

---

a universal law, but that it cannot *be* a law of nature, which is just A).) Therefore, the ultimate explanation for *either* A) or B) will be that the maxim could not *be* a law of nature. The ultimate condition of success for a maxim in these sorts of cases would be that it *could* be adopted as a system of nature, and the ultimate condition of failure is that it could *not* be adopted as a system of nature. The only difference is that on one interpretation, there is the intermediate step whereby the maxim must not be able to be willed to be a law of nature. I should note that each interpretation has support in the text. For B): "Some actions are so constituted that we cannot even *conceive* without contradiction that their maxim be a universal law of nature, let alone that we could *will* that it *ought* to become one" (4:225). For A), consider Kant's discussion of why the maxims in his suicide and false promising cases each fail: e.g. (false-promising), "I then see immediately that this maxim can never qualify as a self-consistent universal law of nature, but must necessarily contradict itself" (4:223).

no stronger or weaker sense of universalizability. There is only the one sense of universalizability—that a maxim *can possibly* be universally followed. And therefore the inference (from the fact that it is rational for *each* person to will a maxim to be universal law *to* the fact that it is rational for each person to will that *everyone* will this maxim) that Wood thought he saw in Kant's argument is avoided. We will see later, in section V, below, how to address Wood's counterexample, and to develop a contradiction between the two claims that Wood argued *should* be contradictory if Kant is right, but which Wood believes are not contradictory.

The second condition provided by FUL is that we (each of us as a rational being) must be able to *will* that the maxim should become a universal law (of permission).

> In the case of other actions, we do not find this inner impossibility, but it is still impossible to *will* that their maxim should be raised to the universality of a law of nature, because such a will would contradict itself (4:225).

Kant thinks the maxims in his examples of failing to develop a talent and failing to help others in need fail in this way. So, perhaps the maxims in these cases *could* be adopted as a system of nature, but the agent could never *will* that they be adopted as a system of nature because it would result in a contradiction in the agent's will. If it is not possible for a rational being to will that the maxim should become a universal law of permission (because that rational will would contradict itself due to conflicting necessary ends) then it is not morally permissible for an agent to act according to that maxim.[34]

[34]One might wonder whether, analogously to the problem raised earlier, we can be mistaken about whether willing a maxim to be a universal law of nature would result in a contradiction in our will. That is, just as it seems possible not to know whether a maxim, if it were to become a law of nature, would result in a contradiction in the system of nature, it seems possible not to know (at least, at a given time) that willing a maxim to be a universal law of nature would result in a contradiction. While it does seem impossible to will x *and* not-x at the same time, it seem somewhat more possible to will contradictorily when the contradiction is less obvious, such as in Kant's examples. And, therefore, *some* maxims that would result in a contradiction in the will could still be willed to be universal law, if the agent was ignorant of the

The point here, again contra Wood, is that if I cannot will a maxim that would result in a contradiction of my will (rooted in a rationally necessary end) to be universally permitted, *then it cannot be valid for every rational being.* The contradiction in the will *will always be rooted in a rationally necessary end*, that is, an end that *all rational beings* necessarily share. If I cannot will the maxim to be universally followed, because of that rationally necessary end, then that maxim cannot possibly be valid for every rational being.

Until we know more, however, it is quite difficult to apply FUL. For one thing, in the absence of any knowledge about what the rationally necessary ends are, it is difficult to know which maxims a rational being would be able to will universally without contradicting his own will. In order to make progress following his discussion of FUL, therefore, Kant moves on to discuss the possibility of an objective end, that is, an end that has absolute worth. If there is a categorical imperative, Kant thinks, there must be an objective end that forms the ground of the imperative. (Relative ends, i.e. ends that are not objective but rather depend on an agent's having chosen them, are only able to support *hypothetical* imperatives.) He writes,

> For if this were not so, [if rational beings did not have absolute value as an objective end], there would be nothing at all having *absolute value* anywhere. But if all value were conditional, and thus contingent, then no supreme principle could be found for reason at all (4:229).

---

contradiction. Kant could reject this possibility, he could claim that there are conditions of success in willing, and so we are not actually willing universally when willing so results in a contradiction, or he could claim that we successfully will them to be universal law, but it results in a contradiction, and so the reason it fails the test is because it results in a contradiction in the will. However, whichever he chooses, the ultimate condition of success for a maxim of this sort is that it does not result in a contradiction in the will, and the ultimate condition of failure is that it does.

And so, if there is a categorical imperative, there must be something with absolute value, or else there could not *be* a categorical imperative, which of course is not possible given the (assumed but not yet proven) antecedent condition.

Kant postulates that there is an objective end[35], and that the objective end is *rational nature* (also sometimes referred to as humanity, or agency). Rational nature *is* an objective end, Kant thinks, but furthermore it is something that we must *regard* as an objective end. He writes, "Suppose, however, there were something *whose existence in itself* had an absolute worth, something that, as an end *in itself*, could be a ground of definite laws" (4:228). He continues, "Now, I say, a human being, and in general every rational being, *does exist* as an end in himself, *not merely as a means* to be used by this or that will as it pleases" (4:229). Kant here emphasizes that a rational being *is* an objective end, that is, each rational being is an end-in-itself.

But he continues, later, saying, "In all his actions, whether they are directed to himself or to other rational beings, a human being **must always be viewed *at the same time as an end***" (4:229, bold emphasis mine). "*Rational nature exists as an end in itself.* This is the way in which a human being necessarily conceives his own existence, and it is therefore a *subjective* principle of human actions" (4:229). In these latter two passages, Kant's emphasis is on how rational nature must be viewed *by rational beings*.[36]

[35]Although, of course, we cannot know this until we have proven that there is a categorical imperative, which must wait until Kant's chapter 3, "Final Step from a Metaphysics of Morals to a Critique of Pure Practical Reason", which is outside the scope of this paper.

[36]How we are supposed to understand Kant's argument for the claim that humanity is an end in itself is a subject of some controversy (see Hill, Analysis of Arguments, pp.123-4). On the one hand, we can understand Kant as claiming that humanity is an objective end, and is an end in itself, and the argument is an argument by elimination (that also depends on the claim that *something* has to be an objective end, if there is a Categorical Imperative). On the other hand, the conclusion is that we must *regard* humanity as an end in itself, and the argument proceeds from the way we each regard ourselves as objective ends to the fact that we must regard all others as objective ends. I think that both interpretations are correct: Kant

Kant uses his claim (which depends on there being a categorical imperative) that there is an objective end to ground another formulation of the categorical imperative, the humanity as an end formulation (FH). This imperative is, "*Act in such a way that you treat humanity, whether in your own person or in any other person, always at the same time as an end, never merely as a means*" (4:230). One fails in one's perfect duty when one treats another as a mere means, and one fails in one's imperfect duty when one fails to treat another as an end.

We now have the information we need in order to be able to understand how to apply FUL more specifically. Let us consider the first condition that FUL provides, that we must be able to *conceive* that our maxim is universally permissible, if acting on the maxim is permitted. If there is a categorical imperative, Kant thinks, then there is an objective end, and it is humanity (rational nature, agency). It is a fact then, if there is a categorical imperative, that humanity is an objective end, a fact that does not depend on our wills. We cannot conceive, Kant must think, that a maxim in which ends-in-themselves are treated as mere means could possibly be universally permitted, because we cannot conceive that a maxim in which something that has absolute worth is treated as if it does not have absolute worth (4:228) could be willed to be universal law. Therefore, any maxim that involves treating another rational being (a being that is an end-in-itself) as a mere means will not conform to the categorical imperative. It is *this* result that legitimates Kant's claim that we must never treat humanity (oneself or another) merely as a means.

---

needs both conclusions, that humanity *is* an end in itself and that we must *regard* humanity to be an end in itself. We will see why shortly.

Let us consider the second condition provided by FUL, that we must be able to will a maxim to be a universal law, if acting on the maxim is permitted. Kant thinks that we must regard humanity as an end-in-itself—we *necessarily* conceive of rational nature as an end-in-itself. Rational nature, humanity, is a *rationally necessary end* for all rational beings. Therefore, any maxim that involves failing to treat an end-in-itself as an end-in-itself will result in a contradiction in the will of the agent, and therefore will not meet the condition set by FUL. And from this we can see that any maxim that involves failing to treat another rational being (a being that is an end-in-itself) as an end-in-itself will not conform to the categorical imperative. It is *this* result that legitimates Kant's claim that we must treat humanity (oneself or another) always at the same time as an end.

Section 5: Answering Wood's objection

At this point, we are able to return to Wood's counterexample and see why it fails against Kant's argument. Remember, Wood objected that if Kant's argument for FUL was successful and valid, then there ought to be a contradiction in maintaining both of the following theses:

1. The principle of self-love ("Always seek your own happiness") is an objective principle.

2. The principle of self-love is not compatible with the principle of morality.

If 2 is true, that is, if the principle of self-love is not compatible with the principle of morality, then it must be either because willing it to be universally followed results in a contradiction in conception/nature or because willing it to be universally followed results in a contradiction in the will. In other words, if self-love is not compatible with the principle of morality, there must be a reason *why* it is not compatible with the principle of

morality—and it is this fact that we will exploit to show why 1 and 2 *do* result in a contradiction.

Suppose that in Wood's example, the reason (2) is true (that is, the reason that the principle of self-love is not compatible with the principle of morality) is because it is not possible to conceive that a maxim to seek one's own happiness is universally permitted. Well, one very convincing reason that it would not be possible to conceive such a maxim as being universally permitted is that such a maxim presupposes that *happiness* is an end-in-itself, i.e. has absolute worth.[37] Remember, we have learned that for Kant, *humanity* is the only end-in-itself. In other words, the principle of self-love requires that we take something that is *not* an end-in-itself and treat it as an end-in-itself. But it is impossible to conceive of a world where it is universally *required* to treat something as an end-in-itself when it is not.

The reason the principle of self-love cannot be conceived without contradiction to be universally followed is that we know that humanity is the only objective end, and yet the principle of self-love presupposes that happiness is an objective end. But we now are in a position to see that (1) (that the principle of self-love is an objective principle) cannot be true, since happiness *is not an objective end*.[38] So, if happiness is not an objective end, then the principle of self-love cannot be an objective principle. In other words, our

[37]A difficulty here is that this is not the reason Wood gives for supposing that the principle of self-love is not the principle of morality. However, we run into similar problems in Kant's own text. He gives one reason for supposing that false promising is not permissible in FUL, and another reason that is not equivalent for supposing that false promising is not permissible in FH. I am not sure what to say about these difficulties.

[38]Remember, the idea that the principle of self-love is an objective principle presupposes that happiness is an objective end. If happiness were *not* an objective end, then it would not be the case that every rational agent always had a necessary ground for pursuing happiness, and therefore the principle of self-love would not be an objective principle.

argument for (2) amounts to a *reductio* of (1). And therefore there *is* a contradiction in supposing both (1) and (2). So, Wood is wrong that there is no contradiction in supposing (1) and (2), at least so long as the reason (2) is true is that supposing the principle of self-love to be universally followed results in a contradiction in conception/nature.

Alternatively, suppose that the reason (2) is true is that willing the principle of self-love to be universally followed results in a contradiction in the will. Our reasoning here will be exactly parallel to the reasoning above. The contradiction in the will that appears when the egoist tries to will that the principle of self-love is universally permitted is due to the fact that the egoist (as a rational being) *necessarily* has rational nature as a necessary end, and yet the principle of self-love requires that he take *happiness* to be his necessary (ultimate) end. Taking happiness to be his necessary end would contradict his actual necessary end, which is rational nature.

But we now are in a position to see that (1) (that self-love is an objective principle) cannot be true, since happiness *is not* the egoist's (or *any* rational being's) objective end. In other words, our argument for (2) amounts to a *reductio* of (1). And therefore there *is* a contradiction in supposing both (1) and (2). So, again, Wood is wrong that there is no contradiction in supposing (1) and (2), at least so long as the reason (2) is true is that supposing the principle of self-love to be universally followed results in a contradiction in conception/nature.

Section 6: Conclusion

In this paper, I have discussed Wood's objection that Kant's argument that the Universal Law Formulation (FUL) is a formulation of the Categorical Imperative relies

on an invalid inference. We saw that at the heart of Wood's objection is the contention that Kant confuses, on the one hand, the question of what is rational for an agent to will for *himself*, and on the other hand, the question of what is rational for an agent to will for *everyone*. I have tried to show, however, that by distinguishing carefully between what Kant wants to accomplish in the universal law formulation and what he wants to accomplish in the Humanity Formulation (FH), we can interpret Kant in such a way that he avoids making the mistake Wood thinks he sees in the argument. I have argued that Kant does not expect the general point of view, as expressed in the universal law formulation, alone to pick out all and only worthy maxims. Instead, we must also rely on a universal end—humanity—and the humanity formulation expresses the role Kant thinks the universal end will play in our deliberation and maxim-selection. In this way, I have tried to defend Kant's attempt to define a central role for a general, impartial point of view in his moral theory, while also addressing the concerns of critics who think that a general, impartial point of view is not sufficient.

Mill's proof of Utilitarianism

In the previous two chapters, I have argued that Hume and Kant both take the recognition that morality involves an impartial point of view as the starting point for their investigation. It is far less clear that Mill's starting point is the same as that of Hume and Kant. However, I will argue that we can rescue the proof of utilitarianism in Chapter 4 from the traditional objections with an interpretation according to which Mill moves with us *from* a first-personal point of view *to* a general or impartial point of view. In this way, I will interpret Mill so that his approach to moral theory is more closely aligned to Hume's and Kant's approaches than we previously might have thought.

In the fourth chapter of *Utilitarianism*,[68] Mill presents an argument that he calls a "proof" of the utilitarian view. Mill's proof, broadly, amounts to a defense of two main claims. The first is that happiness is an end of human conduct, and the second is that happiness is the only end of human conduct. I interpret Mill as a teleologist[69] (for reasons that will be discussed briefly later), and accordingly, if he succeeds in demonstrating that happiness is an end in itself, and the only end in itself, he will have succeeded in demonstrating (within the teleological framework) the truth of utilitarianism. That is, if he succeeds in demonstrating that happiness is an end and the

---

[68] John Stuart Mill, *Utilitarianism*, ed. George Sher. Hackett Publishing Company, Indianapolis, Indiana, 2001.

[69] To use Wellman's term (269). The defense is not intended to be convincing to those who are not persuaded by a teleological approach to ethics. My aim in this paper is to show that, in Mill's world, given his basic moral framework, the argument is not subject to the fallacies that have been attributed to it. Carl Wellman, "A Reinterpretation of Mill's Proof," *Ethics*, v.69:4, pp.268-276.

only end in itself, then he will have succeeded in demonstrating that acts are right insofar as they promote happiness, and that acts are wrong insofar as they fail to do so.

Of the controversy that this proof has engendered, most has focused on this first claim, Mill's claim that happiness is *an* end of human conduct. As the purpose of this paper is to help defend Mill against the critics of his argument, I will focus on the same part of the proof that the critics have. While Mill goes on to argue that nothing other than happiness is an end in itself, that argument can be viewed as an argument distinct from his attempt to prove that happiness is an end in itself.

The argument that happiness is an end in itself has two main steps. Mill begins by claiming that the fact that something (in this case, happiness, but the principle is broad enough to apply to anything) is desir*ed* is evidence that it is desir*able*. Second, he moves from the fact that individual happiness is desirable (to the individual) to the fact that the aggregate happiness is desirable (to the aggregate). This argument as a whole (that happiness is *an* end of human conduct, i.e. that it is valuable and desirable), may be briefly reconstructed as follows:

P1. What is desired is desirable. (Or, a desire for X is evidence that X is desirable.)
P2. Happiness is *desired*.
P3. Happiness is *desirable* (or at least, we have evidence that happiness is desirable, from P1 and P2).
P4. Since each person's happiness is desirable, the aggregate happiness is desirable.

Critics have tended to object to P1 and to P4.[70] They claim that P1 is controversial at best, and obviously wrong at worst, and point to cases where a thing is allegedly desired

[70]See, for example: Hardy Jones, "Mill's Argument for the Principle of Utility," *Philosophy and Phenomenological Research*, v.38:3, pp.338-354. R. F. Atkinson, "J. S. Mill's 'Proof' of the Principle of Utility," *Philosophy*, v.32:121, pp.158-167. James Seth gives a good summary of classic objections, including those by Sidgwick and Dewey. For objections to P1, see pp.474-5, and for objections to P4, see

but not desirable. They claim that P4 is a *non sequitur*, and that Mill's inference relies on the fallacy of composition.

My defense of Mill will involve emphasizing the importance of the role that *points of view* play in this argument. What I will suggest is that P1 and P4 are said from very different points of view, and it is critical to understanding each premise that we understand the point of view from which it is uttered. When Mill claims that the fact that one desires something implies that thing is desirable, the implication is claimed *from the first-person point of view*. That *I* desire something is evidence *to me* that it is desirable *for me* (and the only possible such evidence). But when we move through the argument to P4, the point of view is a more general, moral point of view. The move *from* my happiness being good to me *to* the aggregate happiness being a good to the aggregate is legitimated by taking up a general, moral point of view. From the moral point of view, we draw conclusions about what is morally desirable, using what each individual desires as considerations.

With these modifications, we can understand Mill's argument as follows:

1. I desire my happiness, and I therefore view it as desirable to me, and it is a good to me.

2. Once I adopt a POV stripped of the features that are particular to me, I see that the fact that my happiness is a good to me doesn't depend on particular features either of my happiness or of me.

3. Therefore, happiness (everyone's happiness, not just mine, or yours) is a good in itself.[71]

---

pp.469-72. James Seth, "The Alleged Fallacies in Mill's *Utilitarianism*," *The Philosophical Review*, v.17:5, pp.469-488.

[71]We will see later in more detail what this means, for Mill.

4. Therefore, morality requires that we promote the general happiness (because *qua* a teleologist, Mill holds that morality requires that we promote what is an end in itself).

Mill does not explicitly refer to a general, moral, point of view, so we do not have very much guidance about what features this point of view would have. However, I believe it is sufficient to abstract away everything *specific* from our individual points of view and come to an unbiased point of view that represents what we all share, as humans. Moreover, given that this is an attempt to understand the proof, the only real requirement is to understand this "general point of view" in such a way that it succeeds in resolving the difficulties that commentators have traditionally found in the argument.

In this paper, I will begin with a brief overview of Mill's view and proof, focusing mainly on the elements that will be relevant for interpreting the proof in chapter 4. I will then turn to P1, Mill's claim that the fact that something is desired implies that it is desirable, and argue that the claim is considerably more plausible than it has been thought, once we interpret it from the first-person point of view. A discussion of P4 will follow, and again, I will argue that understanding the claim from a *moral* point of view legitimates the controversial inference (and we will pay special attention to what we will have learned about desirability in the course of our investigation into P1).

Section 1: Overview

The central claim of utilitarianism is that an action's rightness or wrongness derives from the extent to which it maximizes (or fails to maximize) happiness. Right actions are those that maximize happiness, and wrong actions are those that do not

maximize happiness.[72] This is the view that Mill undertakes to defend in the "proof" of utilitarianism. He states the view, at the outset, as follows:

> Actions are right in proportion as they tend to promote happiness; wrong as they tend to produce the reverse of happiness. By happiness is intended pleasure and the absence of pain; by unhappiness, pain and the privation of pleasure (7).

He also provides, helpfully, a gloss for what he means by happiness: it is a positive balance of pleasure over pain. Conversely, unhappiness is a positive balance of pain over pleasure.

His later argument, the proof, is solely focused on arguing that happiness is an end, the sole end, of human action. Consider, for example, what he says in his argument for the proof of utilitarianism:

> The utilitarian doctrine is that happiness is desirable, and the only thing desirable, as an end; all other things being only desirable as a means to that end (35).

The utilitarian doctrine, then, has been referred to as both of the following: one, that happiness is (and is the only thing) desirable as an end, and two, that actions are right in virtue of the extent to which they promote happiness. In order for these two claims to be equivalent, we must close the gap by understanding Mill as advancing (as Wellman calls it, 268) a teleological moral view. That is to say, we must understand Mill to be taking for granted that morality requires us to promote what is desirable and valuable, and never

[72]In chapter 5 of *Utilitarianism*, Mill seems to endorse a kind of rule-utilitarianism, and he claims that some principles of justice (those that are "expedient", or themselves maximize happiness) in some cases are more binding than just the imperative to maximize happiness. "Justice is a name for certain classes of moral rules which concern the essentials of human well-being more nearly, and are therefore of more absolute obligation, than any other rules for the guidance of life; and the notion which we have found to be of the essence of the idea of justice—that of a right residing in an individual—implies and testifies to this more binding obligation" (59). In light of this modification, we can see that in some cases, an action that would maximize happiness may contradict the more important rule of justice, and in these cases, the right action is the one that accords with the rule of justice.

to fail to do so. Or, what is the same for Mill, he must be taking for granted that morality requires us to promote the end of human action. Wellman writes, on this topic,

> Mill's position is that (I) all obligation is grounded exclusively in value. The only factor which determines the rightness or wrongness of an action is the goodness or badness which it contains or produces.

In fact, we can see Mill himself closing this gap and more or less explicitly expressing a teleological moral view in the following quotations:

> Happiness is the sole end of human action, and the promotion of it the test by which to judge of all human conduct; from whence it necessarily follows that it must be the criterion of morality, since a part is included in the whole (39).

> According to the greatest happiness principle […] the ultimate end, with reference to and for the sake of which all other things are desirable—whether we are considering our own good or that of other people—is an existence exempt as far as possible from pain, and as rich as possible in enjoyments, both in point of quantity and quality…. **This, being according to the utilitarian opinion the end of human action, is necessarily also the standard of morality, which may accordingly be defined "the rules and precepts for human conduct**" (12).

In this, Mill's framework, one has only to prove that something is desirable or valuable in order to prove that it is the standard of morality. Mill does not argue for this framework[73], and, of course, ethicists of a different persuasion would vigorously resist his assumption. For the purposes of this paper, however, we will just note this commitment of Mill's and move on. For Mill, proving the truth of utilitarianism need only involve proving that happiness is desirable, and therefore an end, the only end, of human action[74]. Once we have established that fact, for Mill, we have sufficient proof

[73]From the get-go, he claims that we only need to prove that happiness is an end in itself and the only end in itself, in order to establish the truth of utilitarianism. He appears to think it self-evident that morality is concerned with exclusively promoting ends-in-themselves.

[74]We will see later that there are different ways a thing can be desirable. Only those things that are desirable from the general point of view are ends of human action.

that utilitarianism is true. So, the first important point for us to notice going into the proof is what Mill's conclusion must be: that happiness is the only end of human action.

Before we begin our investigation of this proof, though, a little more background will be helpful. One important thing to notice going forward will be that this issue—of what *makes* an action right—is distinct, for Mill, from the issue of what *motivation* a person has to perform that action. That is, we need to distinguish Mill's argument for the truth of utilitarianism from his account of moral motivation, his account of why we act in ways we take to be moral. In the spirit of keeping these two issues distinct, Mill is at great pains to show that the reasons we have to act morally, as well as the reasons we have for *not* acting immorally, are the same for any system of morals. Utilitarianism is neither more nor less subject to any problem concerning moral motivation.

> It is a necessary part of moral philosophy to provide the answer to this question, which, though frequently assuming the shape of an objection to the utilitarian morality, as if it had some special applicability to that above others, really arises in regard to all standards. **It arises, in fact, whenever a person is called on to *adopt* a standard, or refer morality to any basis on which he has not been accustomed to rest it** (27).

He then explicates what he takes to be the sources of motivation for any moral system. He divides these sources into two types: internal and external sanctions.

> The principle of utility either has, or there is no reason why it might not have, all the sanctions which belong to any other system of morals. Those sanctions are either external or internal (28).

External sanctions are, as might be inferred from the name, sanctions that are imposed from outside a person in an attempt to keep his actions in line with morality. These can include any anticipated physical or moral punishment from god or man (e.g. hell, jail, fines, loss of friendship or respect, etc).

The internal sanction, in contrast, is a "subjective feeling in our own minds", i.e., our conscience. Mill describes the feeling as follows.

> The internal sanction of duty, whatever our standard of duty may be, is one and the same—a feeling in our own mind; a pain, more or less intense, attendant on violation of duty, which in properly cultivated moral natures rises, in the more serious cases, into shrinking from it as an impossibility. **This feeling, when disinterested and correcting itself with the pure idea of duty, and not with some particular form of it, or with any of the merely accessory circumstances, is the essence of conscience…** (28-29).

It is our conscience, Mill thinks, that moves us to act morally, and also that which moves us to refrain from acting immorally. Insofar as this is true, it does not matter to which system of morals our conscience is attached. Our conscience will push us to do what we regard as *right* and as our *duty*. If we regard utilitarianism as true, then our conscience will push us to maximize happiness, and restrain us from acting in such a way that would not maximize happiness.

> The ultimate sanction, therefore, of all morality (external motives apart) being a subjective feeling in our own minds, I see nothing embarrassing to those whose standard is utility in the question, what is the sanction of that particular standard? We may answer, the same as of all other moral standards—the conscientious feelings of mankind. **Undoubtedly this sanction has no binding efficacy on those who do not possess the feelings it appeals to; but neither will these persons be more obedient to any other moral principle than to the utilitarian one** (29).

We can see, then, that Mill regards this issue (what causes us to act morally) to be entirely separate from the issue of what the correct moral system is. Utilitarianism may be true, another moral theory may be true—either way, it is our conscience that pushes us to do what we take to be right.[75] So, the second important point for us to notice going

[75]One might concede this much to Mill, and still think that it is the obligation of particular moral systems to account for, not our *motivation* to be moral, but for our *reason* to be moral. There are, after all, plenty of people that abandon Hume's thesis that something can only be a reason if it is motivating. Mill seems to think that reason and motivation are more or less the same, however. In light of the fact that (as I will later

forward is that Mill's proof, while presented in the idiom of "desire", does not concern the question whether and why we have moral motivation. It is not meant to address the question: why be moral? Rather, that idiom is meant to signal only Mill's *teleological* framework: again, that happiness is the sole *end* of human action. Mill's proof is meant to answer questions about what makes right acts right, and not to answer questions about what reason or motivation *we* have to perform right acts.

Let us turn now to consider an overview of Mill's proof of utilitarianism. As we have just seen, this proof is not intended to convince us each individually that we have *reason* to promote the general happiness. Mill has been quite careful to distinguish what he takes to be the *criteria* from what he takes to be the *reason to obey* morality. His proof should be regarded as a philosophical investigation, in which Mill intends to discover what the moral principle is, the principle according to which our actions are properly judged right or wrong (the "*summum bonum*," as he says). He repeatedly emphasizes that he intends to discover the "criteria of morality," and this should be contrasted with any reasons we have for acting morally. Consider the following passage.

> They say it is exacting too much to require that people shall always act from the inducement of promoting the general interest of society. **But this is to mistake the very meaning of a standard of morals and confound the rule of action with the motive of it.** It is the business of ethics to tell us what are our duties, or by what test we may know them; but no system of ethics requires that the sole motive of all we do shall be a feeling of duty (18).

In the proof, what Mill takes himself to need to provide is twofold. First, he must demonstrate that happiness is an end of human conduct. Second, he must demonstrate

---

argue) it is more charitable to Mill to understand his "proof" as if it is *not* intended to provide us with reasons, I will set this issue aside.

that happiness is the only end of human conduct.[76] Given that, as I am interpreting him, he is a teleologist, if he succeeds in demonstrating that happiness is an end in itself, and the only end in itself, he will have succeeded in demonstrating the truth of utilitarianism—that acts are right insofar as they maximize happiness, and that acts are wrong insofar as they fail to do so.

Most of the controversy surrounding Mill's proof has focused on the first claim, the part of the proof in which Mill argues that happiness is *an* end of human conduct. While Mill goes on to argue that nothing *other* than happiness is an end in itself, the later argument can be viewed as distinct from his attempt to prove that happiness is an end in itself. As the purpose of this paper is to help defend Mill against the critics of his argument, we will focus on the same part of the proof that they have: Mill's argument that happiness is an end in itself. As we saw in the introduction, this argument has two main steps. First, he claims that the fact that something (happiness) is desired is evidence that it is desirable. Second, he moves from the claim that individual happiness is desirable to the individual to the claim that the aggregate happiness is desirable to the aggregate. Again, as we saw in the introduction, this argument (that happiness is *an* end of human conduct, i.e. that it is valuable and desirable), may be briefly reconstructed as follows:

P1. What is desired is desirable. (Or, a desire for X is evidence that X is desirable.)
P2. Happiness is *desired*.
P3. Happiness is *desirable* (or at least, we have evidence that happiness is desirable, from P1 and P2).
P4. Since each person's happiness is desirable, the aggregate happiness is desirable.

[76]By "end", he means "end in itself", thereby excluding intermediate ends, those which are ends merely for the sake of some further end.

Let us now consider the controversial premises (P1 and P4) in turn.

Section 2: P1

We can begin with P1, in which Mill claims that the fact that something is desired is evidence that it is desirable. Mill writes,

> The only proof capable of being given that an object is visible is that people actually see it. The only proof that a sound is audible is that people hear it; and so of the other sources of our experience. In like manner, I apprehend, the sole evidence it is possible to produce that anything is desirable is that people actually desire it (35).

Prima facie, this passage contains an elementary flaw of reasoning. The analogy between visibility and audibility, on the one hand, and desirability seems flawed. Whereas "visible" and "audible" just mean "able to be seen" and "able to be heard," respectively, desirable does not mean "*able* to be desired." It means, rather, something like "*worthy* of desire." So, even if it is true that the only proof of an object's visibility is that it is seen, and that the only proof of an object's audibility is that it is heard, it does not seem to follow that the only proof for an object's desirability is therefore that it is desired. That something is desired would be a proof that it is capable of being desired, but that is not the sense of 'desirable' that Mill needs. He needs a proof that happiness is *worthy* of desire, and the analogy with vision and hearing cannot, *prime facie*, earn him that. Furthermore, apart from the failure of this analogy, it just does not seem true that a thing's being *desired* implies that thing's *desirability*. Again, *prima facie*, we seem to be able to come up with many examples where an undesirable object (an object not *worthy* of being desired) is *desired* nonetheless.

Interpreters of Mill rightly set to work by trying to find a reading of the text that can resolve these *prima facie* difficulties. While some of them do an admirable job of

partially resolving the difficulties, I don't believe any of the interpretations to date have succeeded entirely. The interpretation I will propose later in this paper is an attempt to build on what I take to be the most successful of these (the "sufficient condition" interpretation), and amend it so that the remaining difficulties *are* resolved. I will show both that there is good sense to be made of Mill's analogy between seeing, hearing, and desiring, *and* that these purported examples of items that are unworthy of being desire, but are desired nonetheless can be shown to miss their mark once we understand Mill correctly. To begin, we must consider the previous attempts at understanding Mill charitably.

Some interpreters of Mill defend this premise, P1, by claiming that Mill's view is a type of psychological hedonism. On this sort of interpretation, Mill is understood as arguing that happiness (or pleasure) is, *in principle*, the only kind of thing that we are able to desire. The proof, then, can be reconstructed as follows:

> 1. If something is desirable, then we must be able to desire it.
> 2. We are only able to desire pleasure/happiness.
> 3. *Something* is desirable.[77]
> Conclusion: Therefore, pleasure/happiness is desirable.

James Seth, for example, is a proponent of this interpretation of Mill. Consider this passage from "The Alleged Fallacies in Mill's 'Utilitarianism'":

> And we must admit that the truth of the doctrine of psychological Hedonism carries with it the negation of any non-hedonistic theory of the Good, or the desirable in the sense of what ought to be desired. While we cannot say that what we are able to desire is, as such, what we ought to desire, we must admit that what we ought to desire is what we are able to

[77]As Hardy Jones points out, this premise may not be explicitly stated: This "interpretation begins with the weak relation between the desired and the desirable. If nothing else is desired, there is no evidence that anything else is desirable. In order validly to infer that happiness is desirable, another premise must be added: (3) Something is desirable" (344).

> desire**. It follows that, if pleasure is the only thing that we can desire, what we ought to desire cannot be anything other than pleasure** (476).

This interpretation seems to resolve the objections by sidestepping them. The analogy between visibility and audibility, on the one hand, and desirability, on the other hand, no longer does any inferential work. If that is true, then objections that attempt to undermine the analogy by showing that the visible/seen relationship is very different than the desirable/desired relationship are no threat to the argument. They may be appropriate criticisms of his presentation, but Mill's proof remains unscathed. Another virtue of this interpretation is that it seems to avoid objections that depend on counterexamples of things that are desired yet not desirable. If Mill is a psychological hedonist, it is senseless to say that there are cases where something is desired that is not desirable, since we are only able to desire happiness. In these ways, this interpretation circumvents the objections.

I believe, however, there are good textual reasons to reject the psychological hedonism interpretation as an interpretation of Mill's proof. Hardy Jones, in "Mill's Argument for the Principle of Utility," rightly points out that psychological hedonism requires an unnatural reading of the text, since on this interpretation we know from the outset that happiness is the only thing we can possibly desire. Mill goes on to argue for several pages *after* the passage in question that nothing *but* happiness is desirable as an end. He takes great pains to prove that virtue and money, for example, are not desirable as an end, but only desirable as a means to an end.[78] Yet, his doing so seems

[78]Of course, he doesn't stop there. His point ends up being that these things (money, virtue) are *part* of happiness, and as such they are desirable as an end. But until we know that they are part of happiness, we cannot regard them as desirable as an end.

inexplicable, if his earlier claim indeed depends on happiness/pleasure truly being the only thing that we are even *capable* of desiring as an end.

> [Mill] believes himself to have shown that happiness is a good, by the discussion of paragraph three, before he even begins to consider the view that humans are capable of desiring only happiness. He regards this claim as the basis for the conclusion that happiness is the only (intrinsically) desirable end. But the argument provided by the PH interpretation requires it even for the minimal conclusion that happiness is desirable. The account imposes an unnatural, circuitous reading of the text (345).

Furthermore, Jones points out, Mill's claim that the sole evidence that x is desirable is that it is desired, on this interpretation, plays *no* role in the argument. Remember, Mill claims that "the sole evidence it is possible to produce that anything is desirable is that people do actually desire it." But on this interpretation, our discovery that happiness is desirable is not deduced from the fact that we desire it. Jones writes,

> What Mill states as a premise, or at least a basis, for his argument forms no part of the support for its conclusion. The essential point is not that happiness is desired, but that it is the only thing capable of being desired…. Either the interpretation is mistaken, or Mill has incoherently represented his own argument (345).

Jones has provided two good objections on textual grounds. For our purposes, they are sufficient grounds to set the psychological hedonism interpretation of the proof aside.[79]

Let us investigate another interpretive strategy. On this sort of strategy, we must understand Mill as claiming that desire is *evidence* for *desirability*. The view that Mill is positing an evidentiary relationship between desire and desirability seems very plausible out of the gate, since it fits right in with Mill's language ("evidence," "proof," etc…). Supposing this is right, it is important to see that there are still *two* possible ways of

[79]Here I do not mean to take a stand on whether or not Mill is a psychological hedonist. His later argument, that happiness is the *only* thing desirable in itself, would certainly seem to point in that direction. My point here is rather that the argument we are currently considering, and P1 specifically, does not depend on psychological hedonism.

understanding Mill's claim to be a claim about *evidence*: desire can be either sufficient evidence of desirability or insufficient evidence of it.[80] If one interprets Mill according to the "sufficient evidence" (SC) view, then one understands him as claiming that the fact that something is desired is sufficient evidence that it is desirable (worthy of desire). If one interprets him, on the other hand, according to the "inconclusive evidence" (IE) view, then one understands him to be claiming that the fact that something is desired is evidence, albeit inconclusive (i.e. defeasible) evidence that it is desirable.

Clearly, the inconclusive evidence view is a weaker version of this sort of interpretation than the sufficient condition view. The main reason to accept it is that one is inclined toward an interpretation in which Mill is claiming an evidentiary relationship between desire and desirability, and yet one believes the *sufficient* condition claim of the SC view is too strong.[81] There are, however, independent textual reasons for resisting the IE interpretation. Certainly, as Jones points out, Mill writes as if he has established more than just *inconclusively* that happiness/pleasure is desirable. Consider the following quotations:

> "…we have not only all the proof which the case admits of, **but all which it is possible to require,** that happiness is a good" (35).
>
> "If the opinion which I have now stated is psychologically true—if human nature is so constituted as to desire nothing which is not either a part of happiness or a means of happiness—we can have no other proof, **and we can require no other,** that these are the only things desirable" (39).

[80]Here I continue to follow Jones' taxonomy of interpretations.

[81]I believe the main reason for defending an "inconclusive evidence" interpretation is because one finds the alleged counterexamples (cases where a thing is purported to be desired but yet is not desirable) to be successful in demonstrating that desire is not sufficient evidence of a thing's desirability—and yet that Mill must be putting forward a view according to which desire is evidence of a thing's desirability, in which case it is clear that desire is *insufficient* evidence of desirability.

His claims that we can require no further proof would appear to be false if he had proven only inconclusively that happiness is desirable. Since I believe the doubts that would cause us to prefer IE to SC can be addressed, let us turn to the SC interpretation.

According to the SC interpretation, when Mill says "the sole evidence it is possible to produce that anything is desirable is that people do actually desire it," he means that not only is it the sole evidence, but also that it is sufficient evidence.[82] Desiring x is sufficient evidence that x is desirable. Of course, as the argument proceeds, we would discover that we desire happiness, and therefore we would have sufficient evidence that happiness is desirable. As we saw earlier, Mill writes as if he believes he has proven that happiness is desirable, a fact which supports the sufficient condition interpretation.

The main virtue of the sufficient condition interpretation, namely that it seems to be the most natural reading of the passage in question, also turns out to be, at least on the face of it, a vice. That it is the most natural reading of the text seems to subject it, more than other interpretations, to the objections we were considering. In fact, the objections take the SC interpretation as their main target. The analogy between visibility/audibility and desirability seems to provide much of the force behind the claim that desires are sufficient evidence for desirability (in the way that hearing or seeing something is sufficient evidence that it is audible or visible). And, of course, if we were persuaded

[82]Wellman offers a version of the SC interpretation according to which desire is sufficient for desirability in the case of *ends* but not means. I don't find this strategy to be particularly promising (it is an odd distinction to hold for this principle, and doesn't find much support in the text). Furthermore, while this may rescue Mill from difficulties arising from P1, the resulting interpretation is subject to Jones' objection to P4 (which we will discuss later). Jones' objection is that Mill does not demonstrate that aggregate happiness is the same as the end prescribed by *utilitarianism*, and on Wellman's view, this will be true.

that we can desire things that are not desirable, this interpretation does not appear to have assuaged our concerns.

Raphael expresses the worries just discussed (and against which he defends Mill) as follows:

> The alleged fallacy is as follows. Mill talks as if the word "desirable" had the same relation to the word "desired" as the word "visible" has to the word "seen." Now "visible" means "*able* to be seen;" "desirable" means "*worthy* of being desired." If a thing is seen, it follows that it can be seen. If a thing is desired, it follows analogously that it can be desired, but not that it ought, or is worthy, to be desired (347).

Jones expresses the last point, that we can desire things that are not desirable, by arguing that the claim that desire for x is sufficient for demonstrating x's desirability "appears to be obviously false," saying, "Mill must have recognized that a thing's being desired is not sufficient for it to be desirable" (341-2).

This objection appears to rely largely on the contention that there are things that we desire that are yet undesirable. Therefore, we will do Mill the favor of considering the plausibility of these counterexamples. Note that if these counterexamples do not succeed, (that is, if there is nothing of which it is true to say that it is desired and yet not desirable), then Mill's principle (that desire is sufficient evidence for desirability) will have been vindicated. If there is nothing that is desired yet is undesirable, then Mill's claim that desiring x is sufficient evidence for x's desirability will be true.[83]

We can begin with the easiest to dismiss. Jones argues that Mill's own writings themselves can function as a counterexample, saying that Mill's own ethical work "may be understood as aimed at getting people to stop desiring undesirable things and to start desiring desirable ones" (342). Jones' point is that Mill would not bother trying to

[83]Of course, textually, we will still have to explain the analogy between desirability and visibility.

convince us of the truth of utilitarianism, if he were not trying to convince us that some of what we desire is undesirable. Mill's own writings, the reasoning goes, belie the truth of the claim that desiring implies desirability. If we only desired desirable things, we would not need moralists to try to point us in the proper direction. We do desire undesirable things, however, and therefore desiring cannot imply desirability.

Notice the different meanings of "desirable" and the attendant different standards according to which something is worthy of being desired. One such standard is a *moral* one, and while "desirable" *can* in some contexts mean "morally desirable", it certainly does not *have* to mean that. For example, we can say that sunny weather is desirable, but this clearly has nothing to do with morality. Someone who agrees with it does not agree on moral grounds, and someone who disagrees equally does not disagree on moral grounds. Or, we can say that a house is located in a desirable location, and this does not mean that it is in a *morally* desirable location. Examples abound—many things from different facets of life are desirable without thereby being morally desirable.

In fact, at this stage of the argument, it is hard to see how to interpret Mill charitably while also interpreting him as claiming that a thing's being desired implies that it thereby morally desirable. For one thing, this would seem to be an awfully controversial premise for Mill to employ *with no defense at all*. This way of reading 'desirability' would have it that Mill's *very first premise* is that a thing's being desired *implies* that that thing is a moral end. It is *more* controversial than the *prima facie* reading of the claim that desire implies desirability (and therefore doesn't seem very charitable!). Furthermore, once Mill has claimed that desire implies moral desirability and that happiness is desired, he would have cut off any hope of arriving at utilitarianism. For

once we discover that I desire my happiness, and therefore that my happiness is *morally* desirable,[84] we seem to give up any hope of arriving at the moral desirability of the aggregate happiness. That is, on this reading, Mill's very first premise would already *contradict* his conclusion!

If we do not interpret Mill as meaning *morally* desirable, in this premise, then the sense of desirable used in Jones' objection (morally desirable) is a different sense than what Mill has in mind in his "proof." That is, it is plausible to say that he intends "desirable" in a broader sense that is not limited to *morally* desirable. Jones is right that Mill's own writings are an attempt to convince us that certain things are *morally* desirable, even when we don't desire them. But if Mill doesn't mean *morally* desirable when he claims that desire implies desirability, then the fact that we desire things that aren't *morally* desirable is not a successful counterexample.

There are other potential counterexamples, however. Even if Mill does not mean that desiring implies *moral* desirability, one might still object that *other* standards of desirability make this claim of Mill's equally implausible. That is, if Mill's claim is not that a thing's being desired implies that it is morally desirable, it must still be that a thing's being desired implies that it is *in some sense* desirable, and whatever sense we appeal to will be equally subject to counterexample. Surely, one might object, on any *plausible* reading of 'desirable', there are things that we desire that are not desirable. For example, we are familiar with cases in which people desire to undermine or destroy important goods, goods that we recognize as desirable. Consider the following case. A

[84]Instead, I will later argue that we should interpret Mill as claiming that each person's happiness is desirable *to him*, but not that each person's happiness is *morally* desirable. We do not get that happiness is *morally* desirable until later, when in P4 we adopt the moral point of view. This "morally desirable" argument will come in section 3, on P4.

developer desires to raze an important national forest, a forest which contains many endangered species, in order to build condominiums and turn a huge profit. In this case, razing the forest is clearly undesirable, because razing the forest compromises things that are desirable. The developer desires to raze the forest, but we are not able to infer that razing the forest is desirable.[85]

Examples such as this do seem abundant and compelling. To see how Mill can deal with them nonetheless, it will be helpful to step back and take a bird's eye view of his proof. Mill begins by saying that our senses and our internal consciousness judge the first premises of knowledge, and he goes on to imply that it is the same for the first premises of our conduct (i.e. for moral principles). He writes,

> To be incapable of proof by reasoning is common to all first principles, to the first principles of our knowledge, as well as to those of our conduct. But the former, being matters of fact, may be the subject of a direct appeal to the faculties which judge of fact—namely, our senses and our internal consciousness. Can an appeal be made to the same faculties on questions of practical ends? Or by what other faculty is cognizance taken of them? (35).

In this passage, Mill claims that we learn of first principles of knowledge through our senses and our internal consciousness, and he implies that we learn of first principles of conduct (moral principles) in the same way—through our senses and our internal consciousness. This passage is a clue, I think, that we must interpret his premise (that being desired is evidence of desirability) from a *first-person* point-of-view. That is, the kind of appeal that Mill makes concerning the first principle of knowledge is an appeal to each of us to consult *our own* senses and internal consciousness for evidence of the truth

[85]For those that prefer to side with the developer, the case can easily be reversed. The environmentalists desire to preserve the national forest and the endangered species that live in it, yet preserving the forest is not desirable. So, either way, there are things that are desired and yet not desirable.

of such principles. We each must *experience* the truth of these principles for ourselves in order to have their truth "proven" to us. What is important here is not the epistemology of first principles of knowledge that Mill endorses, however, but rather the *mode of appeal* that he issues. Senses and internal consciousness are experienced *first-personally*. To "prove" a first principle requires not an impersonal, third-person "proof", but rather demonstration *to a person* that the first principle is correct. It requires a person to *come to see* that it is correct. If I am right that this is a clue, then the first step in understanding Mill's claim about the kind of evidence that we can have for a thing's desirability will be to reformulate that claim as making an analogous appeal to the *experience* of something *as* desirable. So, we should render that claim as follows: the sole (and sufficient) evidence it is possible to provide *me* that something is desirable[86] is that *I* desire it.

With this in mind, let us return to the case of the developer. In this case, the wedge between what is desired and what is desirable is driven by a disagreement between two *separate* parties. The developer has the desire to raze the national forest, and the environmentalist finds that result undesirable. One party has the desire, and the other party claims the desire is not for something that is desirable. With this in mind, we can see that the developer case is not a successful counterexample. The force of this counterexample comes from *us* seeing that the *developer* desires something that is not desirable. This is not the proper form, however. In order to respect the newly uncovered form of *Mill's* standard of evidence for such things, we need a case in which the same person both desires x *and* finds x undesirable. We must pick a different case.

[86]We will see later that "desirable" will need to be qualified.

What we will see, in picking a case that involves the first-person point of view, is that once I desire something, I have no reason to think that it is not desirable (to or for me). This claim (that once I desire something, I have no reason to think that it is not desirable, to or for me) is what supports the premise that desire is sufficient evidence for desirability.[87] I will argue that supposed counterexamples to this view rely on a confusion between finding something desirable *all-things-considered* and finding it desirable *per se*. But so long as we do not equivocate, we will see that desiring something *necessitates* finding it desirable.[88]

Consider, then, the case of a reluctant smoker: "I, a reluctant smoker, can desire a cigarette, or desire to smoke, *even as* I find smoking and cigarettes undesirable." This is a case of a person who desires something *and yet* finds it undesirable, and so seems to be of the proper form. Both the desire and the undesirability of the cigarettes are within the

[87]For Mill so interpreted, there is no external standard of desirability. In other words, one might worry that we have only established a connection between my desiring something and my finding it desirable, but not the needed connection between desiring something and its *being* desirable. However, we saw at the outset that Mill intends to rely on parallels between first principles of knowledge and first principles of conduct. We can only rely on our senses and our internal consciousness to learn what is desirable. Therefore, there is no further proof (for Mill) that something is desirable than that I desire it and find it desirable.

[88]Mill does not give an analysis of "desire" in the proof, and it would be tricky to do so on his behalf. However, we may not want to count "mere urges" as desires, since they seem relevantly different. For example, if someone cuts me off in traffic, I may have the urge to crash into his car; however, I don't necessarily find crashing into his car desirable. However, we don't usually think of this as a case of "desire." I don't necessarily *want* to crash into his car, and it feels very different to me than other more standard desires. Therefore, I think we have reason to interpret Mill's use of "desire" in such a way that it does not include these kinds of "mere urges." C.f. Scanlon, paraphrasing and endorsing Warren Quinn's point (38): "…although we may sometimes have such urges, the idea of such a purely functional state fails to capture something essential in the most common cases of desire: desiring something involves having a tendency to see something good or desirable about it" (Thanks to Douglas MacLean for pointing this out.) T.M. Scanlon, *What we Owe to Each Other*, Belknap Press of Harvard University Press, Cambridge, Massachussets, 1998.

smoker's own purview. It also seems like a plausible enough case.[89] Is it a counterexample to the sufficient-condition interpretation of Mill's proof?

I will argue that it is not, but to see why, we must delve more deeply into this case. We must investigate *why* the reluctant smoker desires smoking and yet finds it undesirable. What makes him think that smoking is undesirable? Why is he a *reluctant* smoker? Some common reasons for this kind of conflict would be because the reluctant smoker desires cigarettes, on the one hand, and health, smelling nice and/or saving money on the other. In fact, it seems quite plausible that if he *doesn't* desire health, smelling nice, or saving money, then he will not agree that such things are desirable. And if he does not find them desirable, then he will not, in turn, agree that smoking is *not* desirable. The only way to make sense of him finding smoking undesirable, that is, is to imagine him as desiring *other*, incompatible things, and therefore finding these other, incompatible things themselves to be desirable.

This lesson is instructive. We must be careful to distinguish a thing's being desirable *all things considered* from a thing's being desirable *per se.* The claim we have been puzzling over in Mill is that desire implies desirability, and it is important not to equivocate. If what we have just said is right, then what these kinds of counterexamples show is simply that to desire something *per se* does not imply that that thing is desirable *all things considered.* Surely, it is uncharitable to assume that Mill intended to deny this! That is, Mill's claim cannot plausibly have been that desire for something *per se* implies that thing's desirability *all things considered.* If the example is one where something is desired *per se,* then we need to evaluate whether it is desirable *per se,* and not all-things-

[89]And we can think of plenty of similar cases! I, the reluctant dieter, desire chocolate cake even as I find it undesirable (for health reasons). Many compelling cases, familiar experiences, fit this form.

considered. Only if the example is one where something is desired *all-things-considered* would it be fair to evaluate whether it is *desirable* all-things-considered.

In the case of the smoker, the complaint can be fleshed out as follows: The smoker desires smoking *per se*, but does not find it desirable *all-things-considered*. We have seen that using this fact to leverage the case as a counterexample requires a kind of equivocation. What we need to find out, to test the true threat of this case for Mill's proof, is whether the smoker finds smoking desirable *per se*, or, whether he desires smoking *all-things-considered*. Here, our choice comes down to the following two interpretations of the desire/desirability connection.

1. Desiring x *per se* implies that x is desirable *per se.*
2. Desiring x *all-things-considered* implies that x is desirable *all things considered.*

As it turns out, Mill does go on later to prove that nothing *else* is desirable in itself, which *will* establish that happiness is desirable all-things-considered as well as *per se*. Because of this, I believe we should understand Mill as defending the first claim—that desiring x *per se* implies that it is desirable *per se*.

I have not said anything yet about what it is *to be desirable*. This was on purpose. I don't believe that we *need* to say anything about what it is to be desirable, on this view. All that matters, for Mill's view, is that one cannot desire something and yet find it undesirable (*per se*). This is not a logical claim, or meant to be true in virtue of the definition. Rather, I believe that for Mill, this is a psychological fact about us as humans.

If I desire something, *per se*, and from the first-person point of view, I will also see it as desirable, *per se* and from the first-person point of view[90,91].

Admittedly, the claim that something is desirable is often taken to mean that it is desirable all things considered; furthermore it is also often taken to imply that such a thing is desirable from an impartial point of view. But *Mill's* claim need not concern either of these senses of 'desirable'. In fact, the explanation of the smoker case that we have just considered provides a paradigm example of the kind of "proof" that Mill thinks we must give of a thing's "desirability" and also of what it means for something to *be* desirable. It is a case of a person experiencing psychic conflict among his desires. He desires to smoke a cigarette, but he also desires to rid himself of an unhealthy habit. The desire to rid himself of an unhealthy habit is what makes smoking seem undesirable to him—undesirable, once *all* of his desires are taken into consideration. But, taken alone, insofar as he desires to smoke a cigarette, he also finds it desirable[92].

Earlier, we worried that if a thing's being desired implies that it is in some sense (even if not *morally*) desirable, the principle would remain subject to counterexample. We worried that on any *plausible* reading of 'desirable', there are things that we desire that are not desirable. But what we have just seen is that Mill has closed the gap between

[90]We can say that, for Mill, a thing that is desirable is a good. He uses that inference later in the proof, in P4. This doesn't seem to shed any real light on the matter at hand, but we should keep it in mind.

[91]One might wonder how to explain seeing something as desirable yet lacking the corresponding desire. For example, my husband finds it desirable to follow baseball, and he can get me to see the desirability of it even though I never come to share his desire. In a nutshell, Mill can appeal here to the difference between experiencing a desire and empathizing with someone else's experience of a desire. So what I come to understand is how it is desirable *for my husband, from his point of view*. I do this by empathetically feeling his desire, but I do not desire it and it is therefore not desirable for me.

[92]We may ask further metaphysical questions, such as, is the thing desirable *because* it is desired? Is the desire just a sign that it is desirable? Mill does not need to be able to answer these questions. So long as it is true that desiring implies desirability (keeping in mind the appropriate conditions), it does not matter for the view *why* desiring implies desirability, or in virtue of what the implication is true.

desiring and finding something desirable. Insofar as we accept Mill's first-person standard of evidence, the principle seems to be able to fend off any potential counterexample. Furthermore, with this standard of evidence in hand, we can see how the analogy between desirability and visibility is a helpful analogy. Desiring x *is* the only evidence we can have that x is desirable (*for me*), in the same way that seeing x is the only evidence we can have that x is visible (*to me*).

We now have the material to dismiss Jones' second objection. He points out that there is conflict and confusion among human desires, and he claims that if SC were correct, there would thus be conflict and confusion among what is desirable. He writes,

> …there is good reason to believe that Mill would have realized that either x or y could be undesirable though desired by A and B. Many individuals' personal ends are not consonant with the ultimate end of morality, the most important test of all human conduct (and thus of all human ends). If they lack such consonance, surely they are not desirable (342).

But we have seen that Mill's claim needs to be evaluated from a first-person perspective. Conflict and confusion *between* people will not suffice to object to this premise. Conflict and confusion between people would only suffice to object to the premise if we were evaluating this from a point of view other than first-personally. For Mill, all that is required to convince *me* that something is desirable is to get me to desire it. Insofar as I desire it, I also find it desirable (for me)[93].

Furthermore, conflict and confusion among desirable ends for a given individual is only problematic if we mean desirable *all-things-considered*. However, we have

[93]Here, we should not think of this as an evaluative process. It is true that prudential evaluations, for example, involve weighing pros and cons and deciding what (all things, or at least most things, considered) would be best for one. But here, I mean to refer to a much more basic phenomenon. Desiring something just *is* finding it desirable (for me). And in order to find it desirable, I must come to desire it first. The experience of desiring something is the same as the experience of finding something desirable, in the same way that the experience of seeing something is the same as the experience of finding it visible.

already seen good reason for interpreting Mill as meaning desirable *per se*. If this is the case, however, then we can see that there is no threat in taking two incompatible things each as being desirable *per se.* The reluctant smoker is a good example of this! He finds many incompatible things each desirable *per se.* So long as we don't claim that they are all desirable all-things-considered, we will not be saying anything controversial, at least in Mill's mind.

In this section, I have argued that Mill's claim (that desiring something is evidence for its desirability) is more plausible than it has been given credit for. The problem has been that many readers interpreted "desirable" as being a much more loaded term than it needs to be for Mill's purposes. For each person, it is plausible to say (at least, Mill thought it plausible to say) that once he desires something, he finds that thing to be desirable (*per se*). And given the lengths to which Mill goes to compare our experience of something as desired/desirable with our experience of something as seen/visible or heard/audible, it makes sense to understand his claim as being limited to the first person experience. I have attempted to argue that counterexamples cannot apply to this principle when properly understood, since it is plausible to believe, so long as we don't equivocate between things that are desired *per se* and things that are desired all things considered, that desiring something involves seeing it as desirable.

Section 3: P4

Let us now proceed to discuss the second controversial move in Mill's proof, P4: since each person's happiness is desirable, the aggregate happiness is desirable. At this stage in the argument, we will assume that the preceding section was convincing and that Mill has established that each person's happiness is desirable to that person (desirable *per*

se). The claim that we will be discussing in the second half of this paper, then, is found in the following passage:

> This [each person desires his own happiness], however, being a fact, we have not only all the proof which the case admits of, but all which it is possible to require, that happiness is a good: that each person's happiness is a good to that person, and the general happiness, therefore, a good to the aggregate of all persons (35-36).

The standard objection here is that Mill's inference relies on the fallacy of composition. A's happiness may be a good to A, and B's happiness may be a good to B, the objection goes, but we can't thereby conclude that {A's happiness + B's happiness} is a good to {A + B}. Raphael, for example, summarizes the objection as follows (before defending Mill against it):[94]

> Here it is commonly said, Mill commits the fallacy of composition or division. In his conclusion, he takes collectively the objects of all persons' desires, with the phrase "the general happiness," but takes the persons themselves distributively, interpreting the word "aggregate" to mean each of all the persons. From the fact that A's happiness is desirable for A, and B's happiness desirable for me, it does not follow that A's-happiness-plus-B's-happiness is desirable for A and desirable for B. Mill believes he has proved this conclusion (349).[95]

The objection is meant to draw our attention to two facts: First, the aggregate of all persons may not have a good, being that it is not, itself, a person. And second, while Mill may have succeeded in demonstrating that *my* happiness is a good to me, he did not succeed in demonstrating that just any happiness (and therefore the general happiness) is a good to *me*. To put this point another way, Mill's conclusion allegedly relies either on one or both of the following claims:

---

[94]And Seth, in "The Alleged Fallacies in Mill's "Utilitarianism," provides an overview of several of the main examples of this objection, pp.469-472.

[95] D. Daiches Raphael, "Fallacies in and about Mill's *Utilitarianism*," *Philosophy*, v.30:115, pp.344-357.

> C1. The aggregated good is good for the aggregate (implying that the aggregate *has* a good).
>
> C2. The happiness of multiple people can be aggregated into a general happiness and both the individual and the aggregate happiness are equally good for each person.

These claims are implausible and the objection, in drawing our attention to these two features of the argument, is meant to convince us to resist Mill's conclusion.

Mill does say that happiness is a good to the aggregate of all persons; let us note, however, that what he must mean is that happiness is a *good* [96] (and not just a good-for-you or a good-for-me). Mill doesn't *need* to establish the claim that the aggregate of all persons actually has a good, much less that the aggregate has a good *and* that it is the happiness of the aggregate of all persons. Utilitarianism does not actually depend on either of these claims. Mill's claim, rather, is that happiness, no matter *whose happiness*, is a good. It is a good, not just for you, or me, or whoever—it is good in itself. With this, we can set aside any concern that P4 implies that the aggregate has a good (C1).

Assuaging concern about C2 will be more involved. Remember, in P4, from the fact that my happiness is a good to *me*, we infer that the general happiness is a good in itself (this is our spin on Mill's "good for the aggregate"). For Mill, this is a significant transition: it is a move from the *first-person* point of view to a more general point of view, a moral point of view. In P1, we evaluate whether our own happiness is a good *to us,* and in doing so we necessarily evaluate it from the first-person point of view. To say that something is a good to me involves evaluating it from my own point of view and determining that it is desirable, from my point of view.

[96] We will see later that this claim is equivalent to happiness being desirable from the moral point of view.

In contrast, evaluating whether something is good in itself, for Mill, involves evaluating it from a different point of view, a point of view that is unbiased and objective.[97] This point of view could be called a general or a moral point of view. In P1, I realize that my happiness is a good to me, but when I adopt a general, moral point of view, I come to see that there is nothing special about *me* that makes *my* happiness special, more important or more desirable than anyone else's happiness. I realize that my happiness is no more special or desirable than another's happiness; I realize that *each person's* happiness is equally desirable and important. For Mill, this transition is what legitimates the move from an individual's happiness being good for him to the aggregate happiness being a good in itself.[98] It would not be true if said from the first-person point of view. From the first-person point of view, the fact that it is *my* happiness is exactly what makes my happiness more special and more important, to me. As we saw in the previous section, in fact, it is *only* that something is desired *by me* that can ever count as evidence that it is desirable at all.

The fact that Mill has already addressed what reason/motivation we have to be moral can provide us with our first clue about how to understand Mill's claim in P4. Specifically, it cannot be about why *I* should be moral, but rather about what morality requires. For P4, what is required is not a *first-person* evaluation of what is desirable, but rather an evaluation from a distinctly *moral* point of view. From the moral point of view, I must step outside of my own perspective, into a more general point of view, as Hume

[97]It is still a point of view, nonetheless.

[98]My argument for this reading of P4 is mainly a philosophical, rather than exegetical one. The interpretation of P1 with which we have been working, which does find ample textual as well as philosophical support, suggests that "points of view" play a significant role in Mill's thought and the inference in P4, so interpreted, is more reasonable than other competing interpretations of the inference.

would say. From this more general, moral, point of view, we see that I desire happiness, and you do, and we all do. But given that we are all equal and that we all have equal moral standing, from the moral point of view, our happiness is equally desirable[99].

Let's consider an analogy. Take a city planner. In his personal life, he has personal desires and personal ends. It is in his best interest for his neighborhood to remain in the residential zone, because that will keep the value of his house at its current healthy level. It will also keep the neighborhood relatively quiet, which will keep his life more peaceful and relaxing. He has small children, and he would prefer for the neighborhood to avoid the extra traffic that would result from increased business in the area. From his own personal point of view, therefore, it is desirable for the neighborhood to remain zoned residential.

Suppose, however, that he is on a committee to evaluate re-zoning his neighborhood from residential to mixed residential/business. There are clear overriding benefits *to the city* for the neighborhood to be re-zoned as a mixed zone, which will help draw revenue to the city and provide many services to its residents that will make the city a better city and a better place to live. As a city planner, he must evaluate the zoning proposal form a different point of view, a more general point of view that considers the effect on all of the city's residents, and not just on him. From this more general point of view (the "city planner" point of view), he recognizes the desirability (to the residents of the city as a whole) of the adoption of the re-zoning proposal.

[99]Mill doesn't give us much guidance here about what would prompt us to take up this moral point of view. However, it may just be that we take it up because we are reading ch. 4 of *Utilitarianism*, and so we are engaged in an investigation about ethics. If we are engaged in an ethical investigation, then clearly (we might suppose he would think) we need to take up the moral point of view. Another reason might be that we want to engage our conscience in the proper way. Our conscience motivates us to do what it thinks is right; perhaps it also motivates us to verify that what it thinks is right actually *is* right.

Similarly, from my own, personal point-of-view, I desire my *own* happiness. However, from the moral point of view (from my point of view *qua* member of the moral community), I recognize the desirability of aggregate happiness (happiness, no matter whose happiness it is). Once I am in the moral point of view, I can recognize that while my happiness is desirable, it is not so because of anything special about me. And seeing that, I also see that your happiness is desirable, and so is everyone else's. Once we recognize that we have moved from the individual, first-person point of view, into the moral point of view, we no longer need to worry about the difference between *my* happiness and *your* happiness.

One might worry that this implies that we come to desire the aggregate happiness, and that this is implausible, but here we can see a key difference between the moral point of view for Hume and for Mill. On Hume's view, we adopt the general point of view, and proceed to feel sentiments from that point of view. For Mill, in contrast, the tight connection between desire and desirability is true from the first-person point of view, but not necessarily from the moral point of view. For Mill, we don't need to feel desires while in the moral point of view. Rather, when we adopt the moral point of view, we draw conclusions that are based on everyone's desires.

We saw, briefly, earlier that the fact that Mill had already addressed the issue of moral motivation in the third chapter is a clue that in *this* argument, the conclusion will not be about what *I* have reason to do, but rather, what morality requires. So while a trouble-maker may try to resist by saying, "I don't desire the general happiness, and therefore Mill has given me no reason to maximize it," this would be to mistake the

standard of conduct with the *motive* to act in accordance with it.[100] We have already seen that Mill's proof is not intended to provide us with a reason to do anything, including maximize happiness. Mill does not intend his proof to move any of us to action, except insofar as we are already motivated to be moral. Raphael makes this point by saying,

> The common interpretation takes Mill to be trying to prove, in one sentence, that each man ought to desire the happiness of all. Mill would hardly think this could be dealt with in one sentence. He has in fact dealt with it at some length in Chapter III, where he has given a genetic explanation of how men do or may come to adopt the universalistic principle. Sympathy, fostered by social life and education, can lead us to share the desires of others, to desire (and so to call desirable) what others desire for themselves, i.e. their happiness. (350).

We are moved to do what is right because of our conscientious feelings (and external sanctions like punishment), according to Mill. On this interpretation, then, the proof is intended to convince us that our conscience ought to adopt the standard given by utilitarianism, but not to provide us with a further reason to promote the general happiness.

We now can see why C2, the other part of the objection that accuses Mill of succumbing to the fallacy of composition, is misguided. The objection interprets Mill's inference as moving from *my* happiness being a good to me to the *aggregate* happiness being a good to me (because it is a good to the aggregate). We have seen, however, that this latter move is incorrect. Mill does not need to prove that the aggregate happiness is a good to me—only that it is a good in itself. Once it is established that it is a good in itself, my conscience will take over and give me a reason to promote it: my conscience, if operating correctly, will make me desire its promotion.

[100]To paraphrase Mill.

Let us see how this interpretation, by understanding the conclusion of the argument being from the general point of view, can handle another objection. Hardy Jones objects to Mill's proof on the grounds that the sum of each person's happiness is very different than the maximum aggregate happiness. He writes, "The conclusion of the argument is that a certain collection of individual goods is desirable. But *that* collection is not the end specified by the utilitarian principle as the ultimate end." Mill, he claims, needs to hold that the general happiness is the sum of all individual goods, and he objects, saying, "the crucial premise is false. Mill's argument does not show that the maximum happiness, the "second best", is a good" (348). In other words, in P1, Mill claims that *my* happiness is a good, and *your* happiness is a good, and *his* happiness is a good. And maybe Mill has demonstrated that happiness, collectively, is a good. However, aggregated happiness is a very different thing than collective happiness. Since the thesis of utilitarianism is that an act is right insofar as it promotes aggregated happiness, and not a collection of the happiness of individuals, there is a gap between the conclusion of Mill's proof and the thesis of utilitarianism.

However, according to my interpretation, Jones misunderstands the "crucial" premise. Mill's point is not that we sum up the value of many individual goods. Rather, the point is that once we have adopted the moral point of view, we see, within Mill's teleological framework, that there is no reason to keep these goods separate. If pleasure is good and pain is bad, and it doesn't matter (morally) *whose* pleasure and pain it is, then utilitarianism follows. We may begin by realizing that my own happiness, as an individual, is a good. However, Mill does not thereby believe that the ultimate (if unattainable) good is my happiness + your happiness + his happiness (and he doesn't

merely settle for the aggregate happiness). Rather, as we have seen, from the general point of view, it is irrelevant *whose* happiness it is. Jones mistakenly believes that the utilitarian *settles* for second best, because he doesn't realize that Mill's conclusion is said from the moral point of view.[101]

Section 4: Conclusion

In this paper, I have attempted to defend Mill by emphasizing the importance of the role that different *points of view* play in the proof. I have argued that it is essential to Mill's proof that P1 is said from the first-person point of view, whereas P4 is said from the moral point of view. When Mill claims that the fact that I desire x implies that x is desirable, the implication is claimed *from the first-person point of view*. That *I* desire something is evidence *to me* that it is desirable *for me*. But when we move through the argument to P4, we move to a more general, *moral* point of view. The move *from* my happiness being good to me *to* the aggregate happiness being a good to the aggregate is legitimated by taking up a general, moral point of view.

It is interesting, then, that the "standard" interpretation seems to have understood the proof in precisely the reverse way it is intended. Most proofs interpret P1 as being said from a *general* point of view, and therefore interpret the standard for desirability as being that desire implies that something is desirable more generally. Similarly, they interpret P4 as being said from a *first-personal* point of view, and therefore complain that Mill has not succeeded in demonstrating that *I* have reason to promote the aggregate

[101]In an earlier footnote, I alluded to this objection and claimed that Wellman is not able to use this strategy to escape objection. On Wellman's interpretation, he is not able to abstract from the first-person point of view to the moral point of view in the same way, because the fact that *my happiness* is an end proves that *my happiness* is desirable as an end. Each individual's happiness has intrinsic value—and since these values can conflict, as they do in real life, Jones is right that on this view there are two different goods at issue, one is the good in the "proof", and one is the good prescribed by the utilitarian view.

happiness. However, paying appropriate attention to key passages, we saw that P1 is more plausibly interpreted as offering a first-personal standard of evidence for what is desirable *for me*. We also saw that since Mill addressed moral motivation in chapter 3, this proof is more plausibly interpreted as addressing the *standard* of ethics, and therefore P4 should be interpreted as a claim about what is desirable *from the moral point of view*.

Let us revisit Mill's proof, taking stock of the key lessons we learned along the way. First, we learned that P1 should be understood as a first-personal standard of evidence for what is desirable. Insofar as we accept Mill's first-person standard of evidence, anything that I desire will be seen to be desirable (*per se*, at least). We know that I desire my happiness, and therefore I view it as desirable to me, as a good to me. We then saw that, for Mill, in order to move *from* what is desirable to me *to* what is morally desirable, we must correspondingly move *from* a first-person point of view *to* a moral point of view. From the moral point of view, I see that the fact that my happiness is a good to me doesn't depend on particular features either of my happiness or of me. Therefore, from the moral point of view, happiness (everyone's happiness, not just mine, or yours) is a good in itself. Within a teleological framework, this means that morality requires that we promote the general happiness.

I am not a utilitarian, and I do not take myself to have established the truth of utilitarianism. But within a teleological framework, granting Mill certain plausible premises, I have argued that Mill's proof is considerably more plausible than it has been given credit for being.

Bibliography

1. Allison, Henry E., "On a Presumed Gap in the Derivation of the Categorical Imperative," *Philosophical Topics,* v.19:1, pp.2-8.

2. Atkinson, R. F., "J. S. Mill's 'Proof' of the Principle of Utility," *Philosophy,* v.32:121, pp.158-167.

3. Aune, Bruce, *Kant's Theory of Morals*. Princeton University Press, Princeton, New Jersey, 1979.

4. Cohon, Rachel, "The Common Point of View in Hume's Ethics," *Philosophy and Phenomenological Research,* v.57:4, pp.827-850.

5. Galvin, Richard Francis, "Ethical Formalism: The Contradiction in Conception Test," *History of Philosophy Quarterly*, v.8:4, October 1991.

6. Hill, Thomas E. Jr., "Kant's Argument for the Rationality of Moral Conduct," *Pacific Philosophical Quarterly*, v.66, pp. 3-23.

7. Hume, David, *An Enquiry Concerning the Principles of Morals*, ed. J. B. Schneewind. Hackett Publishing Company, Indianapolis, Indiana, 1983.

8. ---*A Treatise of Human Nature*, ed. David Fate Norton and Mary J. Norton. Oxford University Press, 2000.

9. Jensen, Henning, "Hume on Moral Agreement", *Mind*, New Series, v.86:344, pp. 497-513.

10. Jones, Hardy, "Mill's Argument for the Principle of Utility," *Philosophy and Phenomenological Research*, v.38:3, pp.338-354.

11. Kant, Immanuel, *Groundwork for the Metaphysics of Morals*, trans. Arnulf Zweig, ed. Thomas E. Hill Jr., Arnulf Zweig. Oxford University Press, 2002.

12. Korsgaard, Christine M., "Kant's Formula of Universal Law," *Pacific Philosophical Quarterly*, v.66, pp.24-47.

13. Mackie, J. L., *Hume's Moral Theory*. Routledge & Kegan Paul, London, 1980.

14. Mawson, Tim, "Mill's Proof," *Philosophy*, v.7:301, pp. 375-405.

15. Millgram, Elijah, "Mill's Proof of the Principle of Utility," *Ethics*, v.110:2, pp.282-310.

16. Mill, John Stuart, *Utilitarianism*, ed. George Sher. Hackett Publishing Company, Indianapolis, Indiana, 2001.

17. O'Neill, Onora, "Consistency in Action," *Kant's Groundwork of the Metaphysics of Morals: Critical Essays*, ed. P. Guyer. Rowman & Littlefield, Lanham, Maryland, 1998, pp.103-131.

18. --- "Universal Laws and Ends-In-Themselves," *Monist: An International Quarterly Journal of General Philosophical Inquiry*, v.72, pp. 341-361.

19. Paton, H. J., *The Categorical Imperative: A Study in Kant's Moral Philosophy*. University of Pennsylvania Press, Philadelphia, Pennsylvania, 1947.

20. Pogge, Thomas, "The Categorical Imperative," *Kant's Groundwork of the Metaphysics of Morals: Critical Essays*, ed. P. Guyer. Rowman & Littlefield, Lanham, Maryland, 1998.

21. Raphael, D. Daiches, "Fallacies in and about Mill's Utilitarianism," *Philosophy*, v.30:115, pp.344-357.

22. Scanlon, T.M., *What we Owe to Each Other*, Belknap Press of Harvard University Press, Cambridge, Massachussets, 1998.

23. Scarre, Geoffrey, "Interpreting the Categorical Imperative," *British Journal for the History of Philosophy*, v.6:2, pp. 223-236.

24. Seth, James, "The Alleged Fallacies in Mill's Utilitarianism," *The Philosophical Review*, v.17:5, pp.469-488.

25. Spence, G. W., "The Psychology Behind J. S. Mill's 'Proof'," *Philosophy*, v.43:163, pp.18-28.

26. Wellman, Carl, "A Reinterpretation of Mill's Proof," *Ethics*, v.69:4, pp.268-276.

27. Wood, Allen, "Kant on the Rationality of Morals", *Ottawa Congress on Kant in the Anglo-American and Continental Traditions*, 1976, pp.93-110.