Reading Kant's *Groundwork*[1]

J. David Velleman
New York University

The Good Will

The overall strategy of Kant's moral theory is to derive the content of moral obligations from the very concept of an obligation. Kant thinks that we can figure out what morality requires by analyzing the very idea of being morally required to do something. Where I am using the word 'obligation' or 'requirement', Kant uses the German word *Pflicht*, which is usually translated into English as "duty" — an unfortunately antiquated term for what he has in mind. Sticking to Kant's terminology, however, we can say that the strategy is to figure out *what duty requires* by analyzing *what duty is*.

Kant says, "[W]e shall set before ourselves the concept of **duty**, which contains that of a good will though under certain subjective limitations and hindrances" (397, 10).[2] What he means is that the concept of duty is the concept of a requirement to do something whether or not we want to, indeed even if we want not to. When Kant speaks of "a good will ... under certain subjective limitations and hindrances", he means a will that does something just because of being required to and despite wanting not to. Kant reasons that if we couldn't act that way, then we couldn't be required to act that way, and so we wouldn't have any duties at all. He concludes that if we have duties, then we must be capable of

---

[1] To appear in *Moral Philosophy,* ed. George Sher (Oxford: Routledge, projected 2012). This essay supersedes my "Brief Introduction to Kantian Ethics", which appeared in my book *Self to Self; Selected Essays* (New York: Cambridge University Press, 2006), pp. 16–44. I hope that it is more faithful to both the letter and the spirit of the text — which is not to say that it is especially faithful to either one. My understanding of the *Groundwork* has profited from discussion with many graduate students who have attended my undergraduate Ethics course as teaching assistants or auditors, including: Jonny Cottrell, Mihailis Diamantis, Grace Helton, Shieva Kleinschmidt, Colin Marshall, Nick Riggle, Ang Tong; and especially Nandi Theunissen, who provided detailed written comments on an earlier draft, and Melis Erdur, who provided a key element of the interpretation (see note 14, below). Finally, I'm indebted to Kyla Ebels Duggan for comments on the penultimate draft.

[2] References are, first, to the standard Academy Edition of the Groudwork and then to the Cambridge Texts edition, edited and translated by Mary Gregor (Cambridge: Cambridge University Press, 1977). Central Kantian terms are set in bold at their first appearance.

acting on them in opposition to our desires. (Where we speak of desires, Kant speaks of **inclinations**.)

Note that this conception of a good will is not what we ordinarily have in mind when using the term. In our minds, the term 'good will' connotes benevolence, as in the charitable organization of that name. For Kant, however, a good will is not a will that does good;[3] it's a will that does *right*. A will that does right by overcoming its inclinations is what Kant means by "a good will under subjective limitations and hindrances".

One of the most puzzling claims in Kant's *Groundwork* is that there is a special value in doing one's duty when one doesn't want to, a value that doesn't attach to doing one's duty when it's something one wants to do anyway. He draws a contrast between two shopkeepers, one of whom gives correct change because he wants to attract more customers, and the other of whom would prefer to shortchange his customers but doesn't solely because it's forbidden. The latter shopkeeper acts solely **from duty**, whereas the former acts **in accordance with duty** but *from* inclination. For this reason, according to Kant, the act of the latter shopkeeper has "moral worth", whereas the action of the former does not (397, 11).

Kant says that the moral worth of actions in which duty overcomes inclination is a value "higher" than that of actions in which duty and inclination coincide. Many readers take this statement to mean that acts of moral worth are better than or preferable to acts that satisfy both duty and inclination — as if it would be better if we were averse to doing our duty. This interpretation cannot be right. In what sense, then, can the value of moral worth be "higher" than other values?

The answer begins with the fact that although Kant denies moral worth to acts that merely accord with duty, he does credit such acts with other modes of value. Speaking of a case in which someone satisfies his inclinations in doing his duty, Kant says that "an action of this kind . . . deserves praise and encouragement but

[3] "A good will is not good because of what it effects or accomplishes, because of its fitness to attain some proposed end … " (394, 8).

not esteem" (398, 11).[4] In other words, the difference between acts done in accordance with duty but from inclination, on the one hand, and acts done from duty alone, on the other, is that they merit different kinds of appreciation — praise and encouragement in the former case, esteem in the latter. A clue to how these modes of appreciation differ can be found in Kant's statement that moral worth is not just the highest value but "incomparably the highest". Thus, esteem must regard its object as higher in value without regarding it as better, since 'better' is a term of evaluative comparison and so cannot apply to what is incomparable. The question is how a value can be higher than other values without being better.

Think of the value that siblings have in the eyes of their parents. The parents do not cherish any one of their children more than the others, but not because they compare the children to one another and find them equally valuable; rather, they value each child as special and hence as not to be rated or ranked against the others. Indeed, to rank their children in value would already be to devalue them, by disregarding their "specialness", which bars such comparisons. In other words, the parents regard each child as incomparably valuable. When Kant says that moral worth is incomparably higher than other values, he is saying that it is special, precisely in the sense that it must not be ranked against other values, not even as better. Kant uses the term 'esteem' for a mode of appreciation that doesn't rank its objects but regards them as special; by contrast, praise and encouragement are comparative modes of appreciation.

The Structure of the Will

Kant says next that an action performed solely from duty "has its moral worth *not in the purpose* to be attained by it but in the maxim in accordance with which it is decided upon"; he also says that its worth "can lie nowhere else *than in the principle of the will*" (400, 130). As if to explain these statements, he continues:

[4] The notion of esteem also appears here: "We have, then, to explicate the concept of a will that is to be esteemed in itself and that is good apart from any further purpose …" (397, 10). The notion that we have other attitudes of approval toward inclinations appears here: "I cannot have respect for an inclination as such, whether it is mine or that of another; I can at most in the first case approve it and in the second even love it … " (400, 13).

"For, the will stands between its *a priori* principle, which is formal, and its *a posteriori* incentive, which is material, as at a crossroads; and since it must still be determined by something, it must be determined by the formal principle of volition as such when an action is done from duty, where every material principle has been withdrawn from it" (*ibid.*). Unfortunately, this explanation seems to cast only shadow on the subject, not light.

Let's start with the concepts of an **incentive** and a **principle of volition**. An incentive is just something that we want, and our wanting it is what Kant calls an inclination. Kant describes incentives as *a posteriori* because they arise in experience. We learn by experience that something is attainable, and then we experience an inclination to attain it.

Kant says that the will stands between its *a posteriori* incentive and an *a priori* principle of volition. What is that? Kant never really explains, but here is a plausible hypothesis. When we find ourselves with an inclination, we consider whether it gives us a reason for acting. Suppose we know that in order to attain the object of our inclination — our incentive — we would have to take certain steps. We may then consider whether our inclination gives us sufficient reason to take those steps. What we are considering is whether there is a valid principle endorsing the inclination as a good enough reason for the action.[5]

Recall Kant's example. A shopkeeper wants to increase his profits without working any harder. The only way to do so is to start shortchanging his customers. He asks himself whether a desire for easy gain is a good enough reason for shortchanging his customers. In asking this question, he is formulating a principle of reasoning, like this: "A desire for easy gain is a good enough reason to shortchange customers." Having formulated the principle, he is considering whether to base his decision on it — whether, that is, to make it the principle of his volition.

Kant says that a principle of volition is *a priori*. He means that it is not derived from experience. In that respect, it is like all other principles of reasoning. The

[5] Kant himself does not use the concept of a reason for acting. Nevertheless, many contemporary interpreters regard the concept as essential to an understanding of Kant's view.

principle of non-contradiction, for example, is not gleaned from experience. No observation or experiment tells us that the conjunction of a statement and its negation (*p and not-p*) must always be false: we know it without having to investigate. Similarly, it is *a priori* that a desire for an end is a reason for taking the means required for attaining it.[6] If the only means to easy gain is to shortchange one's customers, then it is *a priori* that a desire for the one is a reason for the other.

But is it also *a priori* that the desire is a *good enough* reason? That's the crucial question, to which we as yet don't know the answer. We'll come back to it shortly.

The passages quoted above contain one more concept that calls for explanation — the concept of a **maxim**. A maxim is the content of a possible or actual decision. The shopkeeper in our example is wondering whether to shortchange his customers in the interest of easy gain. "To shortchange my customers for easy gain" is the content of a decision that he is wondering whether to make, and the content of a possible decision is what Kant calls a maxim. If the shopkeeper decides to shortchange his customers for easy gain, then he will have "adopted" that maxim. Adopting a maxim is how the will decides to act.

Here, then, is a diagram of the "crossroads" at which the will of the shopkeeper stands:

A priori	**THE WILL**	*A posteriori*
Principle: "A desire for easy gain is good enough reason to shortchange customers."	Maxim: "to shortchange my customers in the interest of easy gain."	Inclination: desire for easy gain (an incentive).

[6] Some philosophers argue that desiring an end is not a reason for adopting the means to it. (See, for example, John Broome, "Wide or Narrow Scope?", *Mind* 116 [2007]: 359–370.) Rather, they argue, there is a rational requirement *either* to adopt the means *or* to give up the end. I disagree, although I must of course allow for cases in which giving up the end rationally dominates adopting the means. In my view, a desire for the end is indeed *a* reason for adopting the means, but it is not necessarily a *good enough* reason. There may be better reasons against adopting the means, and reasons against adopting the means for an end are also reasons for giving up the end. The difference between these views is that, in mine, there is a presumption in favor of adopting the means to a desired end, although that presumption can be overridden.

The inclination on the right-hand side is presented to the will in experience. The will considers a maxim citing that inclination as a reason ("in the interest of easy gain") for taking an action ("to shortchange my customers"). Adopting the maxim would be rational only if the inclination was in fact a good enough reason for taking the action. And if it was, the agent would have an *a priori principle* to that effect, as shown on the left-hand side. In considering whether to adopt the maxim, then, the will is considering whether there is such a principle on the basis of which to adopt the maxim, as the will is inclined to do.

Assume for the sake of simplicity that a desire for easy gain is the shopkeeper's one and only inclination: it's all he wants in life, at least for the moment. Giving correct change to his customers will therefore entail acting contrary to every inclination he has. If he does give correct change, in that case, he will be performing an act of moral worth. (We are assuming that giving correct change is the right thing to do. It *is* the right thing, of course, but we don't yet know why it is, and so we can only assume so.)[7] But what reason can the shopkeeper have for giving correct change? The only inclination he has is a reason for giving *incorrect* change. He has no inclination that can be cited as a reason for doing the opposite. His reason, and the principle endorsing it, must therefore come from somewhere else. But from where?

The Formula of Universal Law

In reply to this question, Kant says that in an act of moral worth, "there is left for the will nothing that could determine it except objectively the *law* and subjectively *pure respect* for this practical law" (400–401, 13–14). The sudden appearance of **law** at this point is puzzling, but in order to solve the puzzle, we must press ahead to an even greater puzzle. Here it is:[8]

[7] In making this assumption, we are following Kant's own method of proceeding from ordinary moral consciousness to moral philosophy. (See the title of Part I of the *Groundwork.*)

[8] 420–421, 31: "When I think of a *hypothetical* imperative in general I do not know beforehand what it will contain; I do not know this until I am given the condition. But when I think of a *categorical* imperative I know at once what it contains. For, since the imperative contains, beyond the law, only the necessity that the maxim be in conformity with this law, while the law contains no condition to which it would be limited,

> But what kind of law can that be, the representation of which must determine the will, even without regard for the effect expected from it, in order for the will to be called good absolutely and without limitation? Since I have deprived the will of every impulse that could arise for it from obeying some law, nothing is left but the conformity of actions as such with universal law, which alone is to serve the will as its principle, that is *I ought never to act except in such a way that I could also will that my maxim should become a universal law.* [402, 14–15]

The only way to make sense of this passage is to think through the prior puzzle that it is meant to solve.

Let's revisit the shopkeeper as he considers shortchanging his customers. He considers this option by considering whether to adopt the maxim "to shortchange my customers in the interest of easy gain". Now, since we are assuming that it would be wrong to shortchange his customers, we must also assume that his inclination is not good enough reason for doing so. We must therefore assume that the agent is somehow blocked from finding an *a priori* principle endorsing his inclination as a sufficient reason. Why he is blocked remains to be explained. For the moment, however, the question is what principle of volition he can have instead. The principle we're looking for will be the one that determines him to perform an act of moral worth by giving correct change, out of duty and contrary to his one and only inclination. What can that principle be?

The answer is hiding in plain sight. If the agent is to perform an act of moral worth, then his action will have to be determined by something other than inclination: it will have to be determined by a principle. In most cases, an agent's principle tells him that his inclination is a good enough reason for acting. But the agent in the present case has no such reason, since his one and only inclination, we are assuming, isn't good enough. All he has, then, is the *idea* of a good enough reason, with nothing to satisfy it, a concept with no instances. If his action is to be determined by a principle, as it must if he is to act contrary to inclination, then his principle will have to be fashioned out of the materials at hand, hence out of the mere idea of a good enough reason. What principle can be

---

nothing is left with which the maxim of action is to conform but the universality of a law as such; and this conformity alone is what the imperative properly represents as necessary."

fashioned out of *that?* It will have to be the principle of *not acting without a good enough reason.*

Thus, when the shopkeeper lacks a principle endorsing his inclination as a good enough reason to give short change, he must gain access to a principle telling him *not* to act on that inclination — not to give short change — precisely because his inclination is not a good enough reason to do so. And how can he *not* give short change, except by giving *correct* change? His act of moral worth must therefore be based on the principle of not acting without a good enough reason.

But what constitutes a good enough reason? Whatever constitutes a reason must be covered by a principle of reasoning, a principle that sanctions the transition from that reason to its conclusion — or, in this context, to the action for which it is a reason. In this context, then, a good reason for acting is an inclination for which there is a principle validating it as such. Not to act without a reason therefore amounts to not acting without an inclination for which there is a validating principle.

This principle bears a resemblance to Kant's statement that "*I ought never to act except in such a way that I could also will that my maxim should become a universal law"*. The shopkeeper is considering a maxim that proposes an action (giving short change) and a reason for that action (a desire for easy gain). He looks for a principle to validate the reason proposed in his maxim. At this point, he might be described as trying to turn the maxim into a law — for example, by turning the maxim "to shortchange my customers for easy gain" into the principle "A desire for easy gain is a good enough reason to shortchange customers." We don't yet see why he might fail in his attempt to turn his maxim into a law, but we have seen that *if* he fails, he will be left with the bare concept of a good enough reason, plus a principle directing him not to adopt a maxim that doesn't propose one, or in other words, a maxim whose proposed reason is not validated by a principle of reasoning. Thus, he ought not to act on the maxim if he cannot turn it into a

law, exactly as Kant says.[9]

Kant thinks of himself at this point as having established a string of conditional conclusions:

i. If we have any duties, then we must be able to do something that goes against our inclinations — to do it just because it's required.

ii. If we act in opposition to our inclinations, then our action must be determined instead by a principle.

iii. If our principle doesn't endorse any inclinations as good enough reasons, then it must contain the mere, un-instantiated idea of a good enough reason.

iv. If a principle contains no more than the un-instantiated idea of a good enough reason, then it must be the principle of *not* acting without a good enough reason — or in other words, without a reason endorsed as such by a principle of reasoning.

Stringing these conditionals together, we get the conclusion that *if* we have any duties, then in some cases we must act on a last-ditch principle of volition, which tells us not to act without a principle validating the reason proposed in our maxim. This alst-ditch principle of volition must be the one on which we proceed when we go against all of our inclinations, as we do when our action has moral worth.

You may not have realized it, but we have now derived the content of our duty from the very concept of a duty. For we have discovered the one and only principle on which we can act against our inclinations, as duty sometimes requires us and must then enable us to do. When we act against our inclinations, what requires and enables us to do so is the principle of not acting without a reason endorsed by a principle of reasoning. So we have one and only one duty — that is, if we have any duties at all. Our duty is to act on a maxim only if we

[9] The formula "Act only on that maxim which you can at the same time will to be universal law" is called the **Categorical Imperative**. Kant and his interpreters spend much time explaining why it is an imperative and in what sense it is categorical. My interpretation skips over those issues, thus departing from both the text and its standard interpretations.

also have a principle endorsing the reason specified in that maxim as sufficient for the specified action.

Looking back on the argument thus far, we can see why Kant was so interested in acts of moral worth, in which duty must overcome all inclination. These acts are the ones in which duty must be the sole determinant of our will, with inclination playing no part. But if inclination plays no part in determining our will, then all there is to determine it is a principle of volition, which must therefore be the embodiment of our duty. And there turns out to be only one principle of volition that can determine our will to oppose all of our inclinations — namely, the principle of not acting on inclinations without a principle of volition endorsing them as good enough reason to act. That principle must therefore embody our one and only duty. I will therefore call it the *principle of duty* (though that's not Kant's term).

Many questions remain. Why does Kant say that we must be able to *will* that our maxim become universal law? And what does he mean in calling the presumptive law *universal?* Most importantly, why can't the shopkeeper will his easy-money maxim to become a universal law? Let's start with this last question.

Contradictions in Conception

Earlier we considered the law of non-contradiction as a principle of reasoning. We said that it is *a priori* in the sense that it isn't gleaned from observation or experimentation or any other kind of experience. Its validity is obvious to us simply upon reflection.

What's more, the validity of this principle is obvious to any creature[10] that is capable of reasoning. A creature cannot reason unless it regards the law of non-contradiction as valid. A creature that had to learn the validity of this law from experience wouldn't be able to learn it at all, because learning it from experience would require reasoning that already treated the law as valid.

[10] I find the word 'creature' less awkward than the word 'being'. Strictly speaking, however, a creature is a *created* being, and Kant's theory applies also to beings that aren't created — for example, to God.

The fact that the validity of this principle is obvious to any reasoning creature is also obvious to any such creature. In short, everyone knows that the principle is valid, everyone knows that everyone knows it, and so on. The principle's validity is, as we say, *common knowledge* among creatures capable of reasoning. That's the sense in which the principle is universal.

Being universal in this sense is essential to the authority of principles and essential to their being *a priori.* Imagine that when you reflected on the law of non-contradiction, you saw it as valid for your reasoning but you weren't sure whether others would see it as valid for theirs. You would have to wonder whether you shouldn't be looking at the matter from their point-of-view instead of your own. For all you knew, your point-of-view on the validity of this principle might be like a literal, physical point-of-view, from which some things are visible that aren't visible from other points-of-view, and vice versa. You would have to think: Maybe other people can see a problem with the principle that I can't see. The principle would lack authority over your reasoning, since you could always imagine getting around it by resorting to a different point-of-view.

In reality, the principle of non-contradiction has authority in your eyes because you can see that there is no getting around it — no vantage point from which it doesn't hold, or from which their appears to be a vantage point from which it doesn't hold, and so on. It has authority, in short, because it is common knowledge among reasoners, yourself included.

Note that the universality of this principle is not represented in the principle's content. The principle of non-contradiction doesn't say that contradictions are false *for everyone*, or that *no one* should accept a contradiction, or that *anyone* who considers a contradiction should reject it. The principle is universal because everyone finds it valid for his own reasoning, and knows that everyone likewise finds it valid. It's universal, in other words, because its validity is common knowledge.

Principles of practical reasoning can be universal in the same sense. One example is the principle of instrumental reasoning, which says that having an

end is a reason for adopting means to its attainment.[11] Everyone knows, and everyone knows that everyone knows, that the shopkeeper's end of easy gain is a reason for him to adopt some sufficient means to his end.

But the instrumental principle says only that having an end is *a* reason for adopting the means. The question remains, in any particular case, whether having the end is a *good enough* reason — good enough to act on, that is. What the shopkeeper needs is a principle to the effect that his end provides, not just a reason for giving short change, the means that he has proposed, but a reason that's good enough.

Suppose that the latter were a universal principle whose validity was common knowledge — evident to all, as was evident to all, and so on. In that case, it would be obvious to any profit-seeking shopkeeper that his desire for easy gain was a good enough reason for shortchanging his customers; and its being obvious to him would be obvious to all of his customers. But then all his customers would find it obvious that their own financial interests were good enough reason to count their change carefully — in which case, shortchanging them would be impossible. So if the end of easy gain were a good enough reason for the shopkeeper to shortchange his customers, then he would not be able to shortchange them, after all. Shortchanging customers and having good enough reason for doing so are incompatible.

We have now discovered why the easy-money maxim cannot become a principle of reasoning, or in Kant's terms, a universal law. If there were such a principle, then the maxim would be a proposal to do something that was obviously impossible, and so the principle would be self-defeating. The shop-keeper who considers the easy-money maxim therefore finds himself without a viable principle, and he must fall back on the principle of duty, which tells him not to act on his inclination, because he has no principle validating it as a good

[11] This principle roughly corresponds to Kant's **Hypothetical Imperative**. But it is not in imperatival form, and it uses the concept of a reason for acting — two respects in which it departs from the text and from standard interpretations. It is also controversial among contemporary theorists of practical reasoning. See note 6, below.

enough reason for acting.

Suppose I want you to think that I am the author of the leading textbook on Kantian ethics. Since there is no evidence of my having written any such book, I will just have to tell you that I have, in the hope that you will believe me. My maxim will then be as follows: "to tell people that I authored a book in order to get them to believe it". If my maxim became a universal law, it would read as follows: "A desire to get people to believe something is good enough reason for telling it to them".

If there were such a law, then everyone would know, and would know that everyone knew, that wanting to be credited with authorship of a book was good enough reason for claiming it. Yet if I claim something, and you understand what I say, then you will already know that I want you to believe it. And if you also knew that this end, by itself, was good enough reason for making the claim, and that I knew it too, then you would know that I would still make the claim even if I didn't think it was true. In that case, you wouldn't believe me, and so making the claim wouldn't be a means of convincing you, after all. Thus, if the reason stated in my maxim were endorsed by a universal law, then it wouldn't *be* a reason for the proposed action. Wanting people to believe something would not really be a reason for claiming it if it were endorsed as a good enough reason by an *a priori* principle of reasoning.[12] Such a principle would therefore be self-defeating, and so there cannot be one. Lacking a viable principle on which to tell you this lie, I must fall back on the principle of duty, which forbids me to tell it.

Next suppose that a friend asks me to keep his valuables safe while he goes off to climb Mount Everest.[13] And suppose that he is killed in an avalanche,

[12] Having discovered that my maxim cannot be fashioned into a universal principle, I might try to revise it. I might say, "Oh, I don't care whether you *believe* what I'm saying; I just like the sound of it, and you happen to be nearby." I now have a maxim that can be universalized. But I have taken a clearly illegitimate step. I have revised my maxim in order to get around the fact that it can't be universalized; and I have done so by trying to believe something that I know to be untrue. This step has a maxim of its own, which cannot be universalized.

[13] This example is adapted from one that appears, not in the *Groundwork,* but in Kant's *Critique of Practical Reason* (27, 27). See also Kant's essay "On the Proverb: That May be True in Theory, But Is of No Practical Use", in *Perpetual Peace and Other Essays,* trans. Ted Humphrey (Indianapolis: Hackett

leaving no record of what he had deposited with me for safekeeping. Suppose, finally, that I would like to keep his valuables for myself, despite his having heirs to whom I could return them. My maxim would go like this: "to conceal the fact of a deposit in the interest of keeping it for myself". A universal law fashioned from this maxim would go like this: "Wanting to keep a deposit is good enough reason for concealing it."

If there were such a law, then everyone would know, and would know that everyone one knew, that wanting to keep a deposit was good enough reason for concealing it. But then prospective depositors would know that their prospective trustees would have good enough reason for concealing their deposits, and so they would either make no deposits or leave a record of them with reliable proxies. If wanting to keep a deposit were good enough reason for concealing it, then concealing a deposit would be impossible, since no deposit would be made without being recorded. Again, a principle endorsing the inclination as a sufficient reason would be self-defeating; hence there can be no such principle; and so I am left with the principle of duty, which tells me not to act without a principle.

When I find that I cannot fashion a maxim into a principle of reasoning, I am thrown back upon a principle of last resort, which tells me not to act without a principle.[14] As we have seen, this last-ditch principle embodies the one and only duty I have, if I have any duties at all. Oddly enough, my duty turns out to be a principle on which I act as a last resort, when I can find no other principle to act on.

Publishing, 1983, 61–92, pp. 69–70 . It is significant in that the example cannot be interpreted as relying on the consequences of a universal practice, since the practice in this case would be one of concealment that would never be discovered.

[14] The main idea of this section is due to Melis Erdur. It constitutes a significant departure from standard interpretations of Kant, according to which the Categorical Imperative is always present, at least implicitly, in a moral agent's practical reasoning. According to Erdur's interpretation, the Imperative comes into play only when the agent's inclinations must be resisted for the sake of duty — hence only when it provides the basis for an action of moral worth. In the present interpretation, the Imperative then comes into play as a principle of volition, here described as the "principle of last resort". What is always present in an agent's practical reasoning, rather than the Categorical Imperative, is the agent's respect for his own autonomy, which moves him to act on principles of reasoning rather than merely on inclinations. (See the sections titled "Autonomy" and "An End in Itself", below.)

That's why Kant draws such a sharp distinction between acting from duty alone and acting in accordance with duty but out of inclination. In the latter sort of action, my duty is out of sight, even out of mind. I have a principle endorsing my inclinations as good enough reasons, and I act on that principle, by acting for those reasons, without a second thought. Only when I cannot fashion a principle endorsing my inclinations do I fall back on duty as my principle of volition. Only then does my action have **moral content**, as Kant puts it (398, 11), for only then is my action informed by duty as its principle.[15] Otherwise, my action is not *about* morality at all.

You might think that even when I act on a principle endorsing my inclinations as providing sufficient reason, I am also acting partly from duty, which enjoins me never to act without such a principle. Not so. I am indeed acting in accordance with duty, because I have a principle that endorses my inclinations, and such a principle is just what duty requires me to have. But I do not act on the basis of duty's requirement until I find myself without a principle to endorse my inclinations, at which point the command of duty becomes my principle, on the basis of which I resist my inclinations. Provided that I have a principle endorsing those inclinations, however, I can follow them on the basis of that principle, without a thought for my duty.[16]

Of course, there must be something that induces me to make duty my principle when all else fails, but it cannot be duty itself. I don't have a duty to make duty my principle when all else fails, since I couldn't act from such a duty unless I had already made duty my principle. Why, then, do I make duty my principle when all else fails? We are not yet in a position to answer this question.

---

[15] Note that the German phrase *Moralischer Wert* can be translated "moral significance": actions performed on the principle of duty have moral significance, precisely because they have moral content.

[16] This element of the interpretation begins to answer an objection that some commentators have raised to Kant's theory. Their objection is that the theory requires an agent always to have at least one eye on his duty, thereby entertaining "one thought too many". (The phrase comes from Bernard Williams, "Persons, Character and Morality", in *Moral Luck: Philosophical Papers 1973–1980* [Cambridge: Cambridge University Press 1981], p. 119.) Under the present interpretation, duty does not come into view until it is needed to restrain the agent from following his inclinations. Of course, it remains to be explained why the principle of duty, or any ordinary principle, enters into the agent's thinking, to begin with. That explanation will be provided in the sections titled "Autonomy" and "An End in Itself", below.

Note, in any case, that acts with moral content are always acts of omission, acts of not doing something because I cannot frame a principle endorsing my inclinations as providing good enough reason to do it. If a shopkeeper reluctantly gives correct change out of duty, he is, strictly speaking, not giving short change, as he is inclined to do; if I reluctantly inform my friend's heirs of his deposit, I am, strictly speaking, not following my inclination to conceal it. Acting from duty is always a matter of not acting from inclination, on the grounds that my inclinations do not give me good enough reason to act.

Willing the Law

In each of these cases, the defeated reason really is a reason for acting. Wanting easy gain really is a reason for shortchanging customers; wanting people to believe something really is a reason for telling it to them; wanting to keep a deposit really is a reason for concealing it. Liars and crooks are not mistaken to treat these ends as reasons for lying and stealing. Their mistake is in treating these reasons as good enough.[17] Wanting people to believe something isn't a good enough reason for asserting it unless one believes it to be true. Wanting money or valuables isn't a good enough reason for holding on to them unless they don't belong to anyone else.

These extra conditions — that valuables don't belong to others, or that an assertion is believe to be true — are needed for the sufficiency of reasons based on the inclinations of these agents. Thus, an agent's deliberations are not complete if he has merely surveyed his inclinations and balanced up the reasons they provide; he must also consider whether there are any additional conditions required for the sufficiency of those reasons.

How, then, can an agent tell when the reasons he has found are sufficient under the circumstances? Reasons for acting carry no seal of sufficiency

[17] In the following sections, Kant's notion that agents "will the law" is interpreted in terms of what might be called the "sufficiency clause" in principles of volition. According to this interpretation, agents do not decide what counts as a reason, but they do decide whether their proposed reasons are "good enough" for their proposed action: they decide that their reasons are good enough by terminating deliberation, which is something that they must simply do. This interpretation of "willing the law" is non-standard, and there is no direct evidence for it in the text.

guaranteeing that they're good enough. Nor do they carry on their face any indication of which further conditions, if any, would be needed to make them sufficient. It seems as if the agent must simply call a halt to his deliberations at some point and declare "Good enough!"

The necessity of calling a halt to deliberation also arises in the context of consequentialist theories, such as Utilitarianism, where the agent could in principle go on forever imagining alternative actions and their possible consequences. Continuing this process or stopping it are themselves alternative actions whose consequences the agent could go on imagining forever. And the process of deliberating whether to stop deliberating about whether to stop deliberating — that process could go on forever, too. At some point, the agent must simply stop deliberating and act.

The problem is that whether to terminate deliberation appears to be arbitrary, because it cannot be based on deliberation. Consequentialists have no satisfactory solution to the problem. Kant thinks that he has one.

We have already seen a part of Kant's solution. When one declares that reasons are good enough, one purports to state a principle of reasoning, which must be common knowledge among all reasoners. And some purported principles could not be common knowledge, since their being so would undermine them, by making the specified action impossible or canceling the specified reasons. If an agent proposes to perform that action for those reasons, he will encounter an obstacle to declaring them good enough, namely, that his declaration cannot embody a principle of reasoning, because such a principle would be self-defeating.

What about the remaining reasons for acting — the ones for which a declaration of their sufficiency *could* embody a principle of reasoning? It's not enough that there could be such a principle. The mere possibility of an *a priori* principle declaring these reasons to be good enough cannot make them so; there must actually be such a principle that is common knowledge among reasoners.

As we have seen, however, reasons do not bear any obvious mark of sufficiency. How can there be common knowledge about something that isn't obvious?

Intractable as this problem may seem, it belongs to a class of problems that can in fact be solved.[18] Here is an analogous case. Suppose that a demonstration has been called for noon, but the organizers neglected to say where it will take place. Everyone wants to gather with the others in one place, but there has been no public announcement about where to gather. Each person in town will go wherever he thinks the others will go, but he knows that each of the others will be guided likewise, by where he thinks that others will go. How can anyone figure out where to go? What's needed is common knowledge of a gathering place, where everyone can go with confidence that everyone will go there.

Now suppose that there is a square in the middle of town. Obviously, that's where everyone will gather, but not because it has been publicly designated as the gathering place. Everyone will gather in the square simply because there is common knowledge that everyone wants to gather in one place and the square is the most salient place for a gathering. It doesn't have a "Gather Here" sign that everyone can see, but everyone can see that it sticks out, in the eyes of those wanting to gather, as if suggesting itself as site for their demonstration. Thus, common knowledge of a universal desire to converge, plus a uniquely salient point of convergence, can produce common knowledge that convergence will occur at that point.

In this example there is an ultimate, physical convergence, as demonstrators flock to the town square. But there is also a prior, intellectual convergence consisting in common knowledge about where the physical convergence will occur. Demonstrators converge physically because they know that the square is where everyone will converge, that everyone knows it, and so on. So before they

[18] In what follows, "willing the law" is interpreted as a matter of terminating deliberation by declaring one's reasons to be good enough, where that declaration qualifies as a principle of reasoning — hence a law — because it is common knowledge among reasoners. How such a declaration can be common knowledge is then explained in terms of a co-ordination problem among reasoners who seek common knowledge about when reasons are good enough. And the *willing* involved in willing the law is taken to consist in the *willingness* of reasoners to converge on some coordination points rather than others. This interpretation has virtually no support in the text, but it does manage to reconstruct many features of Kant's view — for example, his notion of a Kingdom of Ends.

meet in the square, they have already reached a meeting of the minds about where they will meet.

Practical reasoners are in a somewhat similar situation — except that the ultimate convergence they seek is just a meeting of minds. Each reasoner aims to stop deliberating at a point where his reasons are endorsed as sufficient by a principle that is common knowledge. Like the townspeople who need a gathering point that's common knowledge, practical reasoners need principles of reasoning that are common knowledge; and like the townspeople, they need to arrive at common knowledge without any obvious signpost or signal.

Practical reasoners aren't hoping to converge on a single principle; rather, they are hoping that, with respect to any proposed principle, they will converge on either accepting or rejecting it. They must therefore hope that for any proposed principle, either acceptance or rejection will be the uniquely salient point of convergence. If either acceptance or rejection is the salient point of convergence for a given principle, then their converging on that point will already be common knowledge, given common knowledge of their need to converge, just as the salience of the town square already produces common knowledge that the square is where people will gather. And if it is common knowledge that reasoners will converge on accepting a principle, then the principle itself will be common knowledge, and so it will qualify as a genuine principle of reasoning; whereas if their converging on its rejection is common knowledge, then it won't qualify as a principle of reasoning.

How could acceptance or rejection of a principle achieve salience as a point of convergence among reasoners? Well, rejection of a principle can be salient if convergence on acceptance would obviously be undesirable.[19]

Here is an analogy. Suppose that our imaginary town has both a courthouse square and a market square, equally salient as places for people to gather. And imagine that the market square is obviously too small for a demonstration: convergence on that square would cause a riot and a stampede. The market's

[19] If the principle would be self-defeating, then all of this reasoning about convergence is unecessary, since there cannot be a self-defeating principle.

obvious undesirability as a place for demonstrators to converge produces common knowledge that it won't be the site of the demonstration. In the same way, there can be principles of reasoning whose acceptance would be obviously undesirable as point for practical reasoners to converge.

An example will help.

Contradictions in the Will

Suppose that you and I find ourselves in circumstances where each would lose something by cooperating with the other, no matter what the other does, but would lose even more from the other's failure to cooperate. The cooperation at issue might be helping to harvest one another's fields or (to invoke the relevant cliché) merely scratching one another's backs. In these circumstances, neither of us has anything to gain from helping the other, whether or not the other helps us, and both of us therefore face the prospect of the other's refusing to help. We might wish that we could escape the dilemma through an exchange of mutually dependent offers of the form "I will cooperate if you will." Unfortunately, the resulting agreement would generate a second-order dilemma, since each of us would lose by following through on the agreement, though he would lose even more from the other's refusal to follow through.

It is common knowledge between us that self-interest gives both of us reasons against cooperating. But are reasons of self-interest sufficient? Might self-interest be trumped by other considerations? We aren't sure, and each is fairly sure that the other isn't sure, either, since self-interest doesn't bear a mark of sufficiency on its face. Hence neither of us is sure whether his reasons of self-interest are sufficient, not only because of his own doubts but also because of the doubts that he suspects are entertained by the other, which prevent the necessary principle from becoming common knowledge. Yet in order to find good enough reason for choosing one way or the other, we must have common knowledge as to what those reasons are; otherwise, nothing will count as good enough reason either way, given that what counts as good enough reason must be common knowledge.

Now, it is common knowledge between us that, in order to have sufficient reasons for choosing one way or the other, we need to arrive at common knowledge about what such reasons might be in circumstances like ours. And where there is common knowledge of a need to coordinate, coordination can occur, provided that there is a salient point of convergence whose salience as such is common knowledge. What is the salient point of convergence as to the reasons whether to cooperate in this case?

Well, self-interest gives each of us reason to prefer that, if we are to converge somewhere, then we converge on cooperating, since each will gain more from the other's cooperation than he will lose by his own. Refusing to cooperate is like the market square that would be an undesirable meeting place on which to converge, and whose undesirability as such is common knowledge. So cooperation is like the courthouse square — the most salient point of convergence as to what there is good enough reason to do in such a dilemma. Starting from a lack of common knowledge as to whether reasons of self-interest are trumped by other considerations, we arrive at common knowledge that they are, because it is common knowledge between us that we need to arrive at common knowledge on the question, and answering in the affirmative is the salient point of convergence.

Here is another example. As you walk down the street, you pass a beggar who asks for a dollar. You believe that he is genuinely in need, but you are saving up for a vacation and decide not to help him. Do you have a principle on which to make this decision?

Your maxim in this case is "to withhold help from a needy person in the interest of saving money". The corresponding principle would be "Wanting to save money is good enough reason for withholding help from someone in need." Is that principle common knowledge?

It certainly could be common knowledge.[20] Refusing to help someone in need would still be possible if there were common knowledge that the desire to save was a good enough reason for such a refusal; and the desire would still be a reason for refusing even if its being a good enough reason were common knowledge. So the proposed principle wouldn't be self-defeating.

But whether it *is* a principle of reasoning isn't obvious. The desire to save money is obviously *a* reason for refusing to help someone in need, but it may or may not be good enough: there is nothing on the face of the desire to produce common knowledge that it is or is not a sufficient reason. You and your fellow practical reasoners are therefore on your own, as it were, when it comes to arriving at common knowledge on the subject. But like the townspeople who want to gather for a demonstration, you can reach a meeting of the minds on your own.

You and other reasoners need to have common knowledge as to whether a desire to save is good enough reason for refusing to help others in need. You therefore need to converge either on accepting or on rejecting a principle to that effect. So you must consider whether the salient point of convergence is to accept the principle or to reject it. But everyone would find it undesirable for there to be convergence on the principle that a desire to save was a good enough reason for refusing to help the needy. For everyone would see that such convergence would entail that he himself would suffer in case of need.[21] No one would want everyone to converge on refusing to help him because of a desire to save.

The universal undesirability of converging on acceptance of this principle is obvious. Rejecting the principle is therefore salient as the point of convergence for agents who seek a meeting of the minds, just as avoiding the market square is salient for the townspeople who seek to gather for a demonstration. There is

---

[20] 424, 33: "Some actions are so constituted that their maxim cannot even by *thought* without contradiction as a universal law of nature, far less could one *will* that it *should* become such. In the case of others that inner impossibility is indeed not to be found, but it is still impossible to *will* that their maxim be raised to the universality of a law of nature because such a will would contradict itself."

[21] See also 415, 26: "There is . . . *one* end that can be presupposed as actual in the case of all rational beings . . . and therefore one purpose that they not merely *could* have but that we can safely presuppose that they all actually *do have* by a natural necessity, and that purpose is *happiness*."

consequently common knowledge that the principle of stinginess fails to qualify as a principle of reasoning, and hence that the reason proposed in your maxim of stinginess is not a good enough reason, after all.[22]

There is one important difference between these Kantian examples and the example of converging on the town square. In that case, knowledge of the salient point of convergence depended on empirical information. In order to know that the town had a square, and that it was the obvious place to gather, you would need to have seen the town, or read about it, or learned about it in some other way. So what was common knowledge among the townspeople couldn't be common knowledge among all reasoners wherever they might live.

But a principle of reasoning that specifies what counts as a sufficient reason for acting must be common knowledge among all creatures capable of acting for reasons — that is, among all agents — just as the principle of logic must be common knowledge among all thinkers. That's why the Kantian cases rely only on concepts such as 'self-interest' and 'need', which any rational agent must have.

---

[22] You might have a somewhat weaker maxim of stinginess, for which the corresponding principle would not be undesirable as a point of convergence. It's the maxim "in the interest of saving money, to do only my fair share of charity". If the corresponding principle were common knowledge, then everyone would judge himself to have sufficient reason for refusing to give, but only after he had given his fair share. And if everyone thought he had reason to stop giving only at that point, then no one would be left in need. (Leave aside for present purposes what counts as a "fair share".)

So is it common knowledge that a desire to save money is a good enough reason to do no more than one's fair share of charity? That principle could be common knowledge without undermining itself; and its being common knowledge would not be undesirable. Hence rejecting the principle isn't salient as a point of convergence. But what about accepting it? Surely, there are cases in which a universal desire to converge is frustrated by the lack of a salient point. (What if the town had no central square?)

Practical reasoning is not such a case. Principles of reasoning are proposed for the sake of endorsing inclinations as providing good enough reason to act. Everyone wants to follow his inclinations, conditionally on their providing sufficient reason for doing so. Accepting principles that endorse inclinations is therefore the salient point of convergence, by default, barring conditions that shift salience to rejecting them. When practical reasoners seek common knowledge about the sufficiency of reasons, accepting them as sufficient is the salient point on which to converge.

Thus, although you may be asking yourself whether to scratch the back of someone who has just scratched yours, or whether to put money into a beggar's cup, you needn't know anything about backs or cups in order to find the answer; indeed, if you do know anything about them, you mustn't rely on it. You must think about your situation in terms that any rational agent could understand. In the first case, the relevant description of your situation is simply that each of two people could gain something by refusing to act but would lose even more if the other refuses; in the second case, it's that an agent has needs that others could alleviate. The specific, real-world details are irrelevant.

The last few sections have offered an interpretation of Kant's notion that we must be able to will the principle of reasoning that corresponds to our maxim — or, as he puts it, that we must be able to will our maxim as a universal law. When Kant expresses this notion, he seems to be suggesting that an agent can will a principle of reasoning into existence — that an agent can conjure up such a principle by an act of will. This suggestion would be absurd. A principle of reasoning must be common knowledge among all reasoners, and no individual agent can create common knowledge just by willing it to exist.

But Kant doesn't mean to suggest otherwise. When he says that we must be able to will our maxim as a universal law, he means that practical reasoners must be willing and able to converge on the relevant principle, given the aim of common knowledge about the sufficiency of reasons. If convergence on the principle is neither impossible nor obviously undesirable, then the principle is salient as a point of convergence, agents therefore converge on it, and so it really is common knowledge. And we must act on principles that are common knowledge and that consequently qualify as genuine principles of reasoning.

Thus, we don't create common knowledge of a principle by an act of will; rather, a principle *is* common knowledge because we and other agents would be *willing* to converge on it in our search for common knowledge about reasons for acting. The will enters into creating principles of practical reasoning indirectly,

by making them salient as points of spontaneous convergence among practical reasoners.

Freedom

Let us now return to a question that we have three times postponed, namely, why we act on principles of reasoning in the first place, including, as a last resort, the principle of duty. In order to answer this question, we have to delve into Kant's theory of freedom.

We are concrete beings whose behavior is governed by laws of physics, chemistry, biology, and psychology. Yet if we were no more than such beings, we wouldn't be agents, or so Kant believes. We would still move around, but the wind and waves move around, too. What distinguishes agents from the wind and waves is that agent choose how to act, and Kant thinks that choosing how to act requires **freedom** from the laws governing concrete cause-and-effect.

As before, let us consider this possibility in the case of theoretical reasoning before turning to the practical case. Why, when I am asked the sum of 2 and 2, do I say "Four"? If you consider me from the outside, as a human organism, you will say that I have a brain, hooked to ears on one end and a mouth on the other, like a flesh-and-blood computer hooked to input and output devices. The question goes in my ears, the nerve impulses go round and round in my brain, and the answer comes out of my mouth. The question produces my answer by a mechanism of cause-and-effect.

From my internal perspective, however, the process looks quite different. I don't submit your question to some inner mechanism and wait for the answer to emerge. If I did, I couldn't be sure the answer was right until I had checked the mechanism for loose connections, whereas I know the sum of 2 and 2 *a priori.* From my perspective, I add 2 and 2 by thinking *in the abstract* about the number 2, the function of addition, and the number 4. And the answer 4 isn't issued to me by any mechanism; I *volunteer* the answer, so to speak, and I volunteer the answer "Four" because it's the right one.

This 'because' is not the 'because' of causation. The 'because' of causation does appear in your explanation of my answer. From your perspective, I answer "Four" because of the process by which nerve impulses are channeled through my brain — "because" of this process in the sense of being caused by it. But when I say that I answer "Four" because it's the right answer, I mean that I volunteer that answer because I grasp the relations among abstract objects — the numbers 2 and 4, and the function of addition — which don't participate in cause-and-effect. My 'because' indicates what justifies my answer, not what causes it.

Turning now to the case of practical reasoning, we can see that Kant's view characterizes me as free in the same sense, at least when I act on a principle. When I consider a proposed reason for acting, I must ask myself whether it is validated by a principle of reasoning. If I answer Yes, it's because reasoners would converge on that answer, and here "because" expresses justification, not causation. There isn't an actual convergence of reasoners answering Yes in unison and causing me to give that answer, too; there is just my own abstract reasoning that justifies my volunteering that answer. Acting on a principle thus enables me to win free of cause-and-effect — from my own, internal perspective, at least, if not from that of an external observer.

Kant says that freedom is autonomy.[23] What he means, according to my interpretation, is that we are free insofar as we act on justifications, which consist in abstract principles, which Kant calls laws. The process of framing principles on which to act is what Kant calls *giving ourselves a law.* And giving a law to oneself is autonomy in the etymologically correct sense of the term, since *autos* is Greek for "self" and *nomos* is Greek for "law". That's why Kant says that freedom is autonomy.

We are now in a position to answer a question that we earlier postponed, about my motive for acting on principles. Why do I look for a principle to validate the

[23] "What, then, can freedom of the will be other than autonomy, that is, the will's property of being a law to itself?" (51, 447).

reason proposed in my maxim, and why, failing such a principle, do I fall back on the last-ditch principle forbidding me to act? The answer is that I aspire to be a free agent, which I can become by acting on justifications, embodied in self-given principles, which in Kant's terminology are self-given laws. Thus, I aspire to freedom in the form of autonomy. I aspire to freedom, in the form of autonomy, out of respect for the agent I would be if I were autonomous and therefore free, and for the agent I already am in having that aspiration. I act on principle, then, out of self-respect.

In sum. The principle of duty tells me not to act on inclinations without a principle endorsing them as sufficient reasons for acting. That principle becomes the principle on which I act (or, as the case may be, refrain from acting) when I can find no principle endorsing my inclinations as reasons. I act on a principle when I can frame one, and when I can't, on the principle of duty, out of respect for the free and autonomous agent I could potentially be and the actual agent who has that potential.

An End in Itself

Respect for myself cannot stop at my own freedom. I can perhaps desire freedom just for myself; I can like my own potential for freedom. But respect is an attitude toward a person as embodying some ideal; it is therefore an attitude ultimately toward the ideal, which other persons can embody as well. I cannot truly respect myself as potentially free unless I respect that potential wherever it appears, in others as well as myself; otherwise, I am merely desiring or liking, not respecting.

Thus, my motive for acting on principles, including the principle of duty, is an attitude toward freedom as an ideal, which can be embodied by anyone. Acting without respect for freedom wherever the potential for it appears is therefore incompatible with the motive out of which I act on principles, including, when all else fails, the principle of duty.

On the basis of such reasoning, or something like it, Kant concludes that I have a duty to respect freedom, or the potential for freedom, wherever it appears.[24] (Kant almost always speaks of respect for autonomy.) Kant even claims that the duty to respect autonomy is just another form of the duty not to follow inclinations without a principle for doing so. Here I think that Kant cannot be right about the implications of his own theory. For if my reconstruction of the theory is correct, then respecting autonomy is not really a duty; it is rather the motive from which I act on principles, including the principle of duty.[25] Not to respect autonomy wherever it appears is thus to lack the only motive for doing our duty — a lack that constitutes a vice rather than a violation. According to the present interpretation, then, Kant's theory specifies both a rule, of not acting without a principle, and a virtue, of respecting autonomy, which requires not acting without a principle.

There are various cases in which the requirements of morality are better explained in terms of the virtue than in terms of the rule. These are cases in which there is no obvious point of convergence on the relevant principle of volition.

Consider paternalism, in which one agent is inclined to pre-empt another's choices for the latter's good. One agent may be inclined, for example, to deprive another of tempting but potentially harmful opportunities, or to withhold information about them. ("Don't tell her that he called: he's no good for her." "Don't offer to sell it to him: he can't afford it.") Is there an obvious point at which agents would converge on adopting or rejecting the principle of

[24] 428–429, 37–38. See also 431, 39: "[T]he ground of all practical lawgiving lies (in accordance with the first principle) *objectively in the rule* and the form of universality which makes it fit to be a law . . . ; *subjectively,* however, it lies in the *end;* but the subject of all ends is every rational being as an end in itself (in accordance with the second principle) . . .. And 437–438, 45: "The principle, so act with reference to every rational being (yourself and others) that in your maxim it holds at the same time as an end in itself, is thus at bottom the same as the basic principle, act on a maxim that at the same time contains in itself its own universal validity for every rational being."

[25] 436, 43–44: "All maxims have, namely, (1) a *form,* which consists in universality; and in this respect the formula of the moral imperative is expressed thus: that maxims must be chosen as if they were to hold as universal laws of nature; (2) a *matter,* namely an end, and in this respect the formula says that a rational being, as an end by its nature and hence as an end in itself, must in every maxim serve as the limiting condition of all merely relative and arbitrary ends."

paternalism? Maybe not: some people would prefer to be deprived of dangerous opportunities, others would prefer to decide for themselves. So the principle of paternalism appears to be one for which there would be no convergence on the question of its validity.

Here the virtue of respecting autonomy can fill the gap. Respecting the person's autonomy entails allowing him to choose for herself, without paternalistic interference. Since creatures who act on principles do so out of respect for autonomy, there is after all a point of convergence among them on the principle of paternalism; for there is common knowledge that creatures who act on principles do so out of respect for autonomy, and will therefore converge on rejecting this one.

Conclusion

One last step and we'll be done.

You may recall that in deriving the content of our duties from the very concept of duty, we discovered what our duties would be *if we had any*. If there is anything to which the concept of duty applies, then we know what it must say. But is there such a thing?

For Kant, this question take a particular form, drawn from the contrast drawn earlier between causation and justification. As we saw, my giving 4 as the sum of 2 and 2 can be viewed, from an external perspective, as issuing from a causal process or, from my internal perspective, as based on a justification involving abstract relations among numbers and functions. Kant labels these perspectives as **phenomenal** and **noumenal**, respectively. He says, with good reason, that we cannot understand how these perspectives can be reconciled, and yet cannot give up either one. For Kant, then, the question whether we are free or governed by cause-and-effect is an insoluble mystery. And if we were governed by cause-and-effect, we couldn't really act on justifications, we couldn't act on principles, and so we could have no duty to do so, much less a reason for respecting ourselves as capable of doing so. Another way of putting this point is that if we were governed by cause-and-effect, we wouldn't be capable of performing acts of

moral worth — the capability that Kant adduced in Part I as the prerequisite of our having duties.

Kant now faces a deep and troubling problem. If we aren't free, then we have no duties, and we can never know whether we are free, because we can never dispense with the phenomenal perspective, which is incompatible with the perspective of freedom. Hence *we can never know whether we have duties.*

Not to worry. Kant thinks that it doesn't matter whether we really have duties. In Kant's view, our every act is a bid for freedom, an attempt to be free — perhaps a futile attempt, but an attempt that we cannot help making. So we cannot help acting *as if* we could be free, hence as if we have duties. And the necessity of acting as if we have duties is just as good as having them.

This completes my reconstruction of Kant's moral theory as it is laid out in the *Groundwork.* In summary, it goes like this.

If I have any duties, they require me to do things even I want to avoid them — indeed, even if avoiding them is all I desire. So duties would forbid me from doing what it is my every desire to do. But I can't be forbidden from doing something if I can't help doing it. If I have any duties, then, I must be able to refrain from doing what I want to do — indeed, to refrain from it precisely because it's forbidden. Question: How could I refrain from an action when I have no desire to refrain, because my every desire is to indulge? Answer: I could refrain if I had a principle to determine my will contrary to my desires. Question: What principle could I have for acting contrary to my desires? Answer: The principle would have to be that I mustn't act without a principle endorsing my desires as good enough reasons for acting. Question: How could wanting to do something — indeed, wanting to do it more than anything — fail to be a good enough reason for doing it? Answer: Something's being a good enough reason for an action must be common knowledge among reasoners who consider whether it is. In some cases, something's being a good enough reason for an action would undermine that very reason, or would make the action impossible. (Call these cases contradictions in conception.) In other cases, common

knowledge as to whether something is a good enough reason for an action would have to result from spontaneous convergence among reasoners, and its being a good enough reason is a salient point of avoidance, not convergence. (Call these cases contradictions in the will.) Those two kinds of cases are the ones in which wanting to do something isn't good enough reason to do it; and so those must be the cases in which I have duties, if I have any. Question: But do I have any? Answer: You can't avoid assuming that you do, because you respect yourself as possibly free of your inclinations. And by the way: consistency requires you to respect others, too. 1

It is an absolutely audacious theory, aiming as it does to derive the content of our duties from the mere concept of a duty. In my view, the theory accounts for many of our deepest moral convictions. Whether it is true, or perhaps even the whole truth about morality, is a controversial question.