CHAPTER EIGHTEEN

MOTIVES MATTER

ANOTHER WAY TOWARDS MORALITY

Fulfilling our duty

On the one hand, as we've seen, one can form an objective universal standard for ethical action based on *consequences*. On the other, we'll see this chapter, one can form such a standard based on *duty* or proper motivation.

And there are always *other* hands…

The worry that gives rise to ***deontological*** (duty-based) ethics is simply this: What if consequences don't go our way? What if we do what seems to us the absolutely correct thing, and all still goes to heck? If consequences turn out bad, then the action turns out to be bad, too. But what if we did all in our power to do the right thing? Are we still acting immorally, even so?

Another worry that gives rise to deontology we've already seen. There are some things that we intuitively believe are wrong no matter what. We don't think these things can ever be justified in terms of how they affect everyone involved. We think, rather, that they're wrong no matter the consequences for anyone else (or even ourselves, for that matter).

LAWS AND PRINCIPLES ARE NOT FOR THE TIMES WHEN THERE IS NO TEMPTATION: THEY ARE FOR SUCH MOMENTS AS THIS, WHEN BODY AND SOUL RISE IN MUTINY AGAINST THEIR RIGOUR ... IF AT MY CONVENIENCE I MIGHT BREAK THEM, WHAT WOULD BE THEIR WORTH?

(CHARLOTTE BRONTË)

READING QUESTIONS

As you study this chapter, use these questions for critical thinking and analysis.

- How does Kant argue that logic cannot be in any way dependent upon our experiences?
- How does Kant argue that the only thing that is good without limitation is the good will? What other things do we often think of as good, and how do they ultimately fail to measure up?
- How does Kant argue that the purpose of human rationality isn't happiness?
- What are the four different kinds of motives from which one can decide to act? Which one(s) can be properly called morally praiseworthy? Why can't the others be so called? What are they missing?
- How does Kant get to his definition of duty?
- What is a maxim? What would it mean to suppose that maxim were willed to be a universal law?

continued…

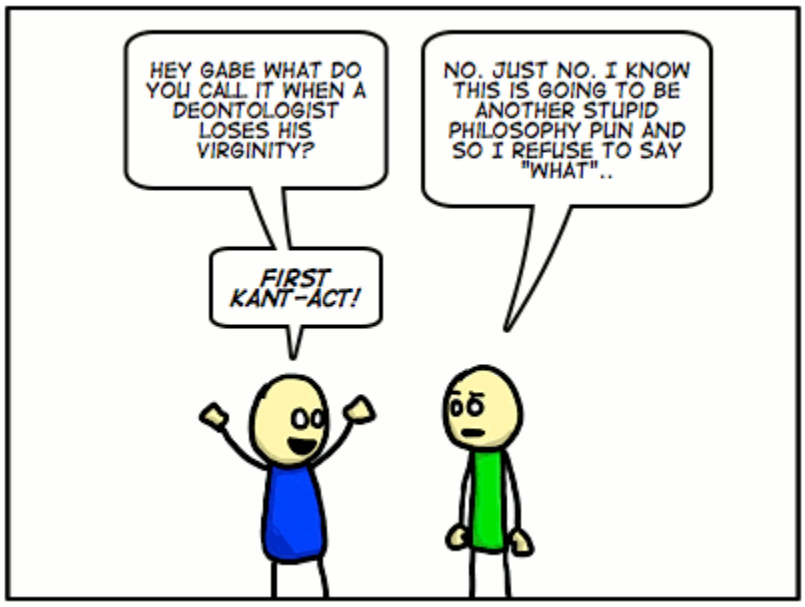
FOUNDATIONS

The following are some key ideas and concepts we'll deal with in this chapter:

- A possible world is one that contains no logical contradictions, or incoherencies. The first formulation of the Categorical Imperative—called the *Formula of Universal Law* (or FUL)—works with this concept. It tells us to create a conjunction* between our maxim of action and a description of a world where it were a universal law that everyone always acted from that maxim. If the two create a contradiction (there is no possible world where they can both be the case), then the action is immoral.
- A hypothetical imperative is dependent upon desired ends, thus always conditional. A categorical imperative is never conditional, thus is universal.
- If a moral theory is to be *non-consequentialist*, then it cannot ever take consequences into account as a part of the moral consideration.
- A *false negative* is a case where something is forbidden when (it seems) it should be allowed. A *false positive* is a case where something is allowed when (it seems) it should be forbidden. These are often found when a theory conflicts with common intuitions or considered moral judgments.

TASKS AND CRITICAL QUESTIONS

This chapter contains ***one task*** and ***at least three critical questions***.† There is ***one team project***.



READING QUESTIONS, *continued*.

- Kant argues that since we cannot be certain of the consequences of an action we must act on logic. How does his application of the Formula of Universal Law (FUL) work? Explain this in a carefully written paragraph, as if writing to a friend who has not taken this class.
- What is an imperative? What is the difference between *hypothetical* and *categorical* imperatives? Give examples of each.
- How is the Categorical Imperative like the Greatest Happiness Principle? How is it different?*
- What one thing does Kant argue is of absolute value? How does this affect the Categorical Imperative?
- How do the Formulation of Universal Law (FUL) and the Formulation of Humanity (FH) relate to each other? How do they together justify the Formulation of Autonomy (FA)?

continued…

* If you struggled with this question, start thinking meta-ethically! Whereas the Categorical Imperative (and all its formulations) focus entirely on having the proper motive, the GHP focuses entirely on having the right consequences. But both of them serve as the supreme principle of morality.

* Remember the truth conditions of a conjunction from chapter 6.

† Depending on how your reading assignments are broken up, it is possible for there to be more.

THE CONTENT AND PURPOSE OF CHAPTER EIGHTEEN

Absolute Standards

Deontology gives us absolutes. It says that certain things are always wrong no matter how they might affect society; certain things are right regardless the consequences. But by what standard can we determine such absolutes? Immanuel Kant gives us the ***Categorical Imperative*** (what I'll call the CI), a universal standard that has nothing to do with consequences. He explains the CI by presenting it in a variety of *formulations*, each of which amounts to the same thing, logically. We will focus on *three* formulations of the CI, which I'll only present here in Kant's words (we'll get them more carefully after your reading).

> ***The Formula of Universal Law*** (FUL): ''Act only in accordance with that maxim through which you can at the same time will that it become a universal law'' (4:421);
>
> ***The Formula of Humanity*** *as End in Itself* (FH): ''*Act so that you use humanity, as much in your own person as in the person of every other, always at the same time as end and never merely as means*'' (4:429; and
>
> ***Formula of Autonomy*** (FA): ''the idea of the will of every rational being as a will giving universal law'' (4:431)

These three formulations are best understood as three legs of the same stool—three aspects of the same universal principle. All of them show us the Categorical Imperative, though they each tease out a specific perspective of the CI.

In the unpacking of the CI, Kant will first introduce us to the notion of a ***good will***, from which all proper ethical behavior arises (and we'll explore why other motivations don't quite make the ethical grade), and then make a distinction between ***hypothetical imperatives*** and the *categorical imperative*.

Hypothetical imperatives are innumerable—they are only each an imperative for somebody on the hypothesis that this somebody wants some certain thing. If you want x, then you must do y. In contrast, the CI is singular—it is an imperative not based on what one wants, but based on the fact that you belong to the category *human being*. Thus, it is a universal imperative—command—law. The CI is a universal law based on pure reason.

We'll unpack Kant's deontology and make certain we understand how it works, then challenge it with a problem that the philosopher Christine Korsgaard attempts to resolve by adding even more nuance to the theory, in much the same way we saw Hare and Singer giving Utilitarianism more nuance in order to respond to worries that arose there.

You will need to read the following very carefully. Kant is precise—so much so that sometimes he can be confusing to our less-than-exact way of thinking. Read slowly and carefully, and prepare a Critical Question

READING QUESTIONS, *continued*.

- What does Kant mean by a 'realm of ends'?
- What is Schiller's Objection? How does it work, and does it correctly understand Kant's theory? Explain why or why not.
- Compare and contrast the FUL with the Golden Rule or your mom's tendency to say "if everyone jumped off a bridge, would you need to do it, too?"
- What is the problem of false negatives? How can Kant respond?
- What is the problem of false positives? How can Kant respond?
- How does Kant argue that there must be something objectively valuable? How does this argument support the Formula of Humanity (FH)?
- What are five attractions of deontology?
- Explain the rigorism problem. How can Kant respond to it? How can't he respond?
- How does Korsgaard modify Kant's deontology to answer the rigorism objection?
- Explain Korsgaard's double-level theory. What are the two levels? How does this maintain using the Categorical Imperative as the supreme principle of morality?
- Can we *ever* lie to somebody who knows we're lying to them, according to Korsgaard?

over *each assigned reading*. You'll find that later CQs can be informed by earlier reading assignments, and this certainly makes Kant more approachable.

GROUNDWORK FOR THE METAPHYSICS OF MORALS

Immanuel Kant[*]

Preface

Ancient Greek philosophy was divided into three sciences: **physics**, **ethics**, and **logic**. This division is perfectly suitable to the nature of the thing and one cannot improve upon it, except only by adding its principle, in order in this way partly to secure its completeness and partly to be able to determine correctly the necessary subdivisions.

All rational cognition is either *material*, and considers some object, or *formal*, and concerns itself merely with the form of the understanding and of reason itself and the universal rules of thinking in general, without distinction among objects. Formal philosophy is called **logic**, but material philosophy, which has to do with determinate objects and the laws to which they are subjected, is once again twofold. For these laws are either laws of **nature** or of **freedom**. The science of the first is called **physics**, and that of the other is **ethics**; the former is also named 'doctrine of nature', the latter 'doctrine of morals'.

NOTES

Logic can have no empirical part, i.e., a part such that the universal and necessary laws of thinking rest on grounds that are taken from experience; for otherwise it would not be logic, i.e., a canon for the understanding or reason which is valid for all thinking and must be demonstrated. By contrast, natural and moral philosophy can each have their empirical part, because the former must determine its laws of nature as an object of experience, the latter must determine the laws for the will of the human being insofar as he is affected by nature—the first as laws in accordance with which everything happens, the second

[*] Published 1766. Footnotes are Kant's, unless otherwise specified. Translation by

as those in accordance with which everything ought to happen, but also reckoning with the conditions under which it often does not happen.

One can call all philosophy, insofar as it is based on grounds of experience, *empirical*, but that which puts forth its doctrines solely from principles *a priori*, *pure* philosophy. The latter, when it is merely formal, is called *logic;* but if it is limited to determinate objects of the understanding, then it is called *metaphysics*.

In such a wise there arises the idea of a twofold metaphysics, the idea of a *metaphysics of nature* and of a *metaphysics of morals*. Physics will thus have its empirical but also a rational part; and ethics likewise; although here the empirical part in particular could be called *practical anthropology*, but the rational part could properly be called *morals.* [...]

Since my aim here is properly directed to moral philosophy, I limit the proposed question only to this: whether one is not of the opinion that it is of the utmost necessity to work out once a pure moral philosophy which is fully cleansed of everything that might be in any way empirical and belong to anthropology; for that there must be such is self-evident from the common idea of duty and of moral laws. Everyone must admit that a law, if it is to be valid morally, i.e., as the ground of an obligation, has to carry absolute necessity with it; that the command 'You ought not to lie' is valid not merely for human beings, as though other rational beings did not have to heed it; and likewise all the other genuinely moral laws; hence that the ground of obligation here is to be sought not in the nature of the human being or the circumstances of the world in which he is placed, but *a priori* solely in concepts of pure reason,* and that every other precept grounded on principles of mere experience, and even a precept that is universal in a certain aspect, insofar as it is supported in the smallest part on empirical grounds, perhaps only as to its motive, can be called a practical rule, but never a moral law.

Thus not only are moral laws together with their principles essentially distinguished among all practical cognition from everything else in which there is anything empirical, but all moral philosophy rests entirely on its pure part, and when applied to the human being it borrows not the least bit from knowledge about him (anthropology), but it gives him as a rational being laws *a priori*, which to be sure require a power of judgment sharpened through experience, partly to distinguish in which cases they have their application, and partly to obtain access for them to the will of the human being and emphasis for their fulfillment, since he, as affected with so many inclinations, is susceptible to the idea of a pure practical reason, but is not so easily capable of making it effective *in concreto* in his course of life.

NOTES

* The term *a priori* literally means "from before," and refers to anything that comes prior to or without any reference to experience. It refers to things whose truth value is only dependent on logic. [Kurle note]

NOTES

Thus a metaphysics of morals is indispensably necessary not merely from a motive of speculation, in order to investigate the source of the practical principles lying *a priori* in our reason, but also because morals themselves remain subject to all sorts of corruption as long as that guiding thread and supreme norm of their correct judgment is lacking. For as to what is to be morally good, it is not enough that it *conform* to the moral law, but it must also happen *for the sake of this law*; otherwise, that conformity is only contingent and precarious, because the unmoral ground will now and then produce lawful actions, but more often actions contrary to the law. But now the moral law in its purity and genuineness (which is precisely what most matters in the practical) is to be sought nowhere else than in a pure philosophy; hence this (metaphysics) must go first, and without it there can be no moral philosophy at all; that which mixes those pure principles among empirical ones does not even deserve the name of a 'philosophy' (for this distinguishes itself from common rational cognition precisely by the fact that what the latter conceives only as mixed in, it expounds in a separate science), still less of a 'moral philosophy', because precisely through this mixture it violates the purity of morals and proceeds contrary to its own end. […]

The Good Will

There is nothing it is possible to think of anywhere in the world, or indeed anything at all outside it, that can be held to be good without limitation, excepting only a **good will**. Understanding, wit, the power of judgment, and like *talents* of the mind, whatever they might be called, or courage, resoluteness, persistence in an intention, as qualities of *temperament*, are without doubt in some respects good and to be wished for; but they can also become extremely evil and harmful, if the will that is to make use of these gifts of nature, and whose peculiar constitution is therefore called *character*, is not good. It is the same with *gifts of fortune*. Power, wealth, honor, even health and that entire well-being and contentment with one's condition, under the name of *happiness*, make for courage and thereby often also for arrogance, where there is not a good will to correct their influence on the mind, and thereby on the entire principle of action, and make them universally purposive; not to mention that a rational impartial spectator can never take satisfaction even in the sight of the uninterrupted welfare of a being, if it is adorned with no trait of a pure and good will; and so the good will appears to constitute the indispensable condition even of the worthiness to be happy.

Some qualities are even conducive to this good will itself and can make its work much easier, but still have despite this no inner unconditioned worth, yet always presuppose a good will, which limits the esteem that one otherwise rightly has for them, and does not permit them to be held absolutely good. Moderation in affects and passions, self-control,

and sober reflection not only are good for many aims, but seem even to constitute a part of the *inner* worth of a person; yet they lack much in order to be declared good without limitation (however unconditionally they were praised by the ancients). For without the principles of a good will they can become extremely evil, and the cold-bloodedness of a villain makes him not only far more dangerous but also immediately more abominable in our eyes than he would have been held without it.

The good will is good not through what it effects or accomplishes, not through its efficacy for attaining any intended end, but only through its willing, i.e., good in itself, and considered for itself, without comparison, it is to be estimated far higher than anything that could be brought about by it in favor of any inclination, or indeed, if you prefer, of the sum of all inclinations. Even if through the peculiar disfavor of fate, or through the meager endowment of a stepmotherly nature, this will were entirely lacking in the resources to carry out its aim, if with its greatest effort nothing of it were accomplished, and only the good will were left over (to be sure, not a mere wish, but as the summoning up of all the means insofar as they are in our control): then it would shine like a jewel for itself, as something that has its full worth in itself. Utility or fruitlessness can neither add to nor subtract anything from this worth. It would be only the setting, as it were, to make it easier to handle in common traffic, or to draw the attention of those who are still not sufficiently connoisseurs, but not to recommend it to connoisseurs and determine its worth.

There is, however, something so strange in this idea of the absolute worth of the mere will, without making any allowance for utility in its estimation, that despite all the agreement with it even of common reason, there must nevertheless arise a suspicion that perhaps it is covertly grounded merely on a high-flown fantasy, and that nature might have been falsely understood in the aim it had in assigning reason to govern our will. Hence we will put this idea to the test from this point of view.

In the natural predispositions of an organized being, i.e., a being arranged purposively for life, we assume as a principle that no instrument is to be encountered in it for any end except that which is the most suitable to and appropriate for it. Now if, in a being that has reason and a will, its *preservation*, its *welfare*—in a word, its *happiness*—were the real end of nature, then nature would have hit on a very bad arrangement in appointing reason in this

NOTES

NOTES

creature to accomplish the aim. For all the actions it has to execute toward this aim, and the entire rule of its conduct, would be prescribed to it much more precisely through instinct, and that end could be obtained far more safely through it than could ever happen through reason; and if, over and above this, reason were imparted to the favored creature, it would have served it only to make it consider the happy predisposition of its nature, to admire it, to rejoice in it, and to make it grateful to the beneficent cause of it, but not to subject its faculty of desire to that weak and deceptive guidance, and meddle in the aim of nature; in a word, nature would have prevented reason from breaking out into *practical use* and from having the presumption, with its weak insight, to think out for itself the project of happiness and the means of attaining it; nature would have taken over the choice not only of the ends but also of the means, and with wise provision would have entrusted both solely to instinct.

In fact we also find that the more a cultivated reason gives itself over to the aim of enjoying life and happiness, the further the human being falls short of true contentment; from this arises in many, and indeed in those most practiced in the cultivated use of reason, if only they are sincere enough to admit it, a certain degree of *misology*, i.e., hatred of reason; for after reckoning all the advantages they draw, I do not say from the invention of all the arts of common luxury, but even from the sciences (which also seem to them in the end to be a luxury of the understanding), they nevertheless find that they have in fact only brought more hardship down on their shoulders than they have gained in happiness, and on this account in the end they sooner envy than despise human beings of the more common stamp, who are closer to the guidance of mere natural instinct and do not permit their reason much influence over their deeds and omissions. And we must admit this much, that the judgment of those who very much moderate the boastful high praise of the advantages that reason is supposed to supply us in regard to happiness and contentment with life, or who even reduce it below zero, is by no means morose or ungrateful toward the kindness of the world's government; but rather these judgments are covertly grounded on the idea of another aim for their existence, possessing much greater dignity, for which, and not for their happiness, reason has been given its wholly authentic vocation, and to which, therefore, as a supreme condition, the private aims of the human being must for the most part defer.

For since reason is not sufficiently effective in guiding the will safely in regard to its objects and the satisfaction of all our needs (which it in part itself multiplies), and an implanted natural instinct would have guided us much more certainly to this end, yet since reason nevertheless has been imparted to us as a practical faculty, i.e., as one that ought to have influence on the *will*, its true vocation must therefore

be not to produce volition *as a means* to some other aim, but rather to produce a *will good in itself*, for which reason was absolutely necessary, since everywhere else nature goes to work purposively in distributing its predispositions. This will may therefore not be the single and entire good, but it must be the highest good, and the condition for all the rest, even for every demand for happiness, in which case it can be united with the wisdom of nature, when one perceives that the culture of reason, which is required for the former, limits in many ways the attainment of the second aim, which is always conditioned, namely of happiness, at least in this life, and can even diminish it to less than nothing without nature's proceeding unpurposively in this; for reason, which recognizes its highest practical vocation in the grounding of a good will, is capable in attaining this aim only of a contentment after its own kind, namely from the fulfillment of an end that again only reason determines, even if this should also be bound up with some infringement of the ends of inclination.

But now in order to develop the concept of a good will, to be esteemed in itself and without any further aim, just as it dwells already in the naturally healthy understanding, which does not need to be taught but rather only to be enlightened, this concept always standing over the estimation of the entire worth of our actions and constituting the condition for everything else: we will put before ourselves the concept of **duty**, which contains that of a good will, though under certain subjective limitations and hindrances, which, however, far from concealing it and making it unrecognizable, rather elevate it by contrast and let it shine forth all the more brightly.

Acting from the Motive of Duty

I pass over all actions that are already recognized as contrary to duty, even though they might be useful for this or that aim; for with them the question cannot arise at all whether they might be done *from duty*, since they even conflict with it. I also set aside the actions which are actually in conformity with duty, for which, however, human beings have immediately *no inclination*, but nevertheless perform them because they are driven to it through another inclination. For there it is easy to distinguish whether the action in conformity with duty is done *from duty* or from a self-seeking aim. It is much harder to notice this difference where the action is in conformity with duty and the subject yet has besides this an *immediate* inclination to it. E.g., it is indeed in conformity with duty that the merchant should not overcharge his inexperienced customers, and where there is much commercial traffic, the prudent merchant also does not do this, but rather holds a firm general price for everyone, so that a child buys just as cheaply from him as anyone else. Thus one is *honestly* served; yet that is by no means sufficient for us to believe that the merchant has proceeded thus from duty and from principles of honesty; his

NOTES

Three different motives for helping your neighbor fix her flat

NOTES

advantage required it; but here it is not to be assumed that besides this, he was also supposed to have an immediate inclination toward the customers, so that out of love, as it were, he gave no one an advantage over another in his prices. Thus the action was done neither from duty nor from immediate inclination, but merely from a self-serving aim.

By contrast, to preserve one's life is a duty, and besides this everyone has an immediate inclination to it. But the often anxious care that the greatest part of humankind takes for its sake still has no inner worth, and its maxim has no moral content. They protect their life, to be sure, *in conformity with duty*, but not *from duty*. If, by contrast, adversities and hopeless grief have entirely taken away the taste for life, if the unhappy one, strong of soul, more indignant than pusillanimous or dejected over his fate, wishes for death and yet preserves his life without loving it, not from inclination or fear, but from duty: then his maxim has a moral content.

To be beneficent where one can is a duty, and besides this there are some souls so sympathetically attuned that, even without any other motive of vanity or utility to self, take an inner gratification in spreading joy around them, and can take delight in the contentment of others insofar as it is their own work. But I assert that in such a case the action, however it may conform to duty and however amiable it is, nevertheless has no true moral worth, but is on the same footing as other inclinations, e.g., the inclination to honor, which, when it fortunately encounters something that in fact serves the common good and is in conformity with duty, and is thus worthy of honor, deserves praise and encouragement, but not esteem; for the maxim lacks moral content, namely of doing such actions not from inclination but *from duty*. Thus suppose the mind of that same friend of humanity were clouded over with his own grief, extinguishing all his sympathetic participation in the fate of others; he still has the resources to be beneficent to those suffering distress, but the distress of others does not touch him because he is sufficiently busy with his own; and now, where no inclination any longer stimulates him to it, he tears himself out of this deadly insensibility and does the action without any inclination, solely from duty; only then does it for the first time have its authentic moral worth. Even more: if nature had put little sympathy at all in the heart of this or that person, if he (an honest man, to be sure) were by temperament cold and indifferent toward the sufferings of others, perhaps because he himself is provided with particular gifts of patience and strength to endure his own, and also presupposes or even demands the same of others; if nature has not really formed such a man into a friend of humanity (although he would not in truth be its worst product), nevertheless would he not find a source within himself to give himself a far higher worth than that which a good-natured temperament might have? By all means! Just here begins the worth of

character, which is moral and the highest without any comparison, namely that he is beneficent not from inclination but from duty.

To secure one's own happiness is a duty (at least indirectly), for the lack of contentment with one's condition, in a crowd of many sorrows and amid unsatisfied needs, can easily become a great *temptation to the violation of duties*. But even without looking at duty, all human beings always have of themselves the most powerful and inward inclination to happiness, because precisely in this idea all inclinations are united in a sum. Yet the precept of happiness is for the most part so constituted that it greatly infringes on some inclinations and yet the human being cannot make any determinate and secure concept of the sum of satisfaction of them all, under the name of 'happiness'; hence it is not to be wondered at that a single inclination, which is determinate in regard to what it promises and the time in which its satisfaction can be obtained, can outweigh a wavering idea; and the human being, e.g., a person with gout, could choose to enjoy what tastes good and to suffer what he must, because in accordance with his reckoning, here at least he has not sacrificed the enjoyment of the present moment through expectations, perhaps groundless, of a happiness that is supposed to lie in health. But also in this case, if the general inclination to happiness does not determine his will, if for him, at least, health does not count as so necessary in his reckoning, then here, as in all other cases, there still remains a law, namely to promote his happiness not from inclination but from duty, and then his conduct has for the first time its authentic moral worth.

It is in this way, without doubt, that those passages in scripture are to be understood in which it is commanded to love our neighbor and even our enemy. For love as inclination cannot be commanded; but beneficence solely from duty, even when no inclination at all drives us to it, or even when natural and invincible disinclination resists, is *practical* and not *pathological* love, which lies in the will and not in the propensity of feeling, in the principles of action and not in melting sympathy; but the former alone can be commanded.

The second proposition is: an action from duty has its moral worth *not in the aim* that is supposed to be attained by it, but rather in the maxim in accordance with which it is resolved upon; thus that worth depends not on the actuality of the object of the action, but merely on the *principle of the volition*, in accordance with which the action is done, without regard to any object of the faculty of desire. It is clear from the preceding that the aims we may have in actions, and their effects, as ends and incentives of the will, can impart to the actions no unconditioned and moral worth. In what, then, can this worth lie, if it is not supposed to exist in the will, in the relation of the actions to the effect hoped for? It can lie nowhere else *than in the principle of the will*, without regard to the ends that can be effected through such action; for

NOTES

NOTES

the will is at a crossroads, as it were, between its principle *a priori*, which is formal, and its incentive *a posteriori*,* which is material, and since it must somehow be determined by something, it must be determined through the formal principle in general of the volition if it does an action from duty, since every material principle has been withdrawn from it.

The third proposition, as a consequence of the first two, I would express thus: *Duty is the necessity of an action from respect for the law.* For the object, as an effect of my proposed action, I can of course have an *inclination*, but *never respect*, just because it is merely an effect and not the activity of a will. Just as little can I have respect for inclination in general, whether my own or another's; I can at most approve it in the first case, in the second I can sometimes even love it, i.e., regard it as favorable to my own advantage. Only that which is connected with my will merely as a ground, never as an effect, only what does not serve my inclination but outweighs it, or at least wholly excludes it from the reckoning in a choice, hence only the mere law for itself, can be an object of respect and hence a command. Now an action from duty is supposed entirely to abstract from the influence of inclination, and with it every object of the will, so nothing is left over for the will that can determine it except the *law* as what is objective and subjectively *pure respect* for this practical law, hence the maxim† of complying with such a law, even when it infringes all my inclinations.

The moral worth of the action thus lies not in the effect to be expected from it; thus also not in any principle of action which needs to get its motive from this expected effect. For all these effects (agreeableness of one's condition, indeed even the furthering of the happiness of others) could be brought about through other causes, and for them the will of a rational being is therefore not needed; but in it alone the highest and unconditioned good can nevertheless be encountered. Nothing other than the *representation of the law* in itself, *which obviously occurs only in the rational being* insofar as it, and not the hoped-for effect, is the determining ground of the will, therefore constitutes that so pre-eminent good which we call 'moral', which is already present in the person himself who acts in accordance with it, but must not first of all be expected from the effect.‡

* The term *a posteriori*, like the term *a priori* refers to knowledge or truth claims as they relate to human experience. But whereas *a priori* (literally, 'from before') things have truth or meaning without reference to experience, *a posteriori* (literally 'from after') things have truth or meaning only by means of experience or by reference to empirical testing. [Kurle note]

† A *maxim* is the subjective principle of the volition; the objective principle (i.e., that which would serve all rational beings also subjectively as a practical principle if reason had full control over the faculty of desire) is the practical *law*.

‡ One could accuse me of merely taking refuge behind the word *respect* in an obscure feeling instead of giving a distinct reply to the question through a concept of reason. Yet even if respect is a feeling, it is not one *received* through influence but a feeling *self-effected* through a concept of reason and hence specifically distinguished from all feelings of the first kind, which may be reduced to inclination or fear. What

The Moral Law

But what kind of law can it be, whose representation, without even taking account of the effect expected from it, must determine the will, so that it can be called good absolutely and without limitation? Since I have robbed the will of every impulse that could have arisen from the obedience to any law, there is nothing left over except the universal lawfulness of the action in general which alone is to serve the will as its principle, i.e., I ought never to conduct myself except so *that I could also will that my maxim become a universal law*. Here it is mere lawfulness in general (without grounding it on any law determining certain actions) that serves the will as its principle, and also must so serve it, if duty is not to be everywhere an empty delusion and a chimerical concept; common human reason, indeed, agrees perfectly with this in its practical judgment, and has the principle just cited always before its eyes.

Let the question be, e.g.: When I am in a tight spot, may I not make a promise with the intention of not keeping it? Here I easily make a distinction in the signification the question can have, whether it is prudent, or whether it is in conformity with duty, to make a false promise. The first can without doubt often occur. I do see very well that it is not sufficient to get myself out of a present embarrassment by means of this subterfuge, but rather it must be reflected upon whether from this lie there could later arise much greater inconvenience than that from which I am now freeing myself, and, since the consequences of my supposed *cunning* are not so easy to foresee, and a trust once lost to me might become much more disadvantageous than any ill I think I am avoiding, whether it might not be more *prudent* to conduct myself in accordance with a universal maxim and make it into a habit not to promise anything except with the intention of keeping it. Yet it soon occurs to me here that such a maxim has as its ground only the worrisome consequences. Now to be truthful from duty is something entirely different from being truthful out of worry over disadvantageous consequences; in the first case, the concept of the action in itself already contains a law for me, whereas in the second I must look around elsewhere to see which effects might be bound up

NOTES

I immediately recognize as a law for me, I recognize with respect, which signifies merely the consciousness of the *subjection* of my will to a law without any mediation of other influences on my sense. The immediate determination of the will through the law and the consciousness of it is called *respect*, so that the latter is to be regarded as the *effect* of the law on the subject and not as its *cause*. Authentically, respect is the representation of a worth that infringes on my self-love. Thus it is something that is considered as an object neither of inclination nor of fear, even though it has something analogical to both at the same time. The *object* of respect is thus solely the law, and specifically that law that we *lay upon ourselves* and yet also as in itself necessary. As a law we are subject to it without asking permission of self-love; as laid upon us by ourselves, it is a consequence of our will, and has from the first point of view an analogy with fear, and from the second with inclination. All respect for a person is properly only respect for the law (of uprightness, etc.) of which the person gives us the example. Because we regard the expansion of our talents also as a duty, we represent to ourselves a person with talents also as an *example of a law*, as it were (to become similar to the person in this) and that constitutes our respect. All so-called moral *interest* consists solely in *respect* for the law.

NOTES

with it for me. For if I deviate from the principle of duty, then this is quite certainly evil; but if I desert my maxim of prudence, then that can sometimes be very advantageous to me, even though it is safer to remain with it. Meanwhile, to inform myself in the shortest and least deceptive way in regard to my answer to this problem, whether a lying promise is in conformity with duty, I ask myself: Would I be content with it if my maxim (of getting myself out of embarrassment through an untruthful promise) should be valid as a universal law (for myself as well as for others), and would I be able to say to myself that anyone may make an untruthful promise when he finds himself in embarrassment which he cannot get out of in any other way? Then I soon become aware that I can will the lie but not at all a universal law to lie; for in accordance with such a law there would properly be no promises, because it would be pointless to avow my will in regard to my future actions to those who would not believe this avowal, or, if they rashly did so, who would pay me back in the same coin; hence my maxim, as soon as it were made into a universal law, would destroy itself.

"Now, what should our policy on honesty be?"

Common Morality and Philosophy

Thus I need no well-informed shrewdness to know what I have to do in order to make my volition morally good. Inexperienced in regard to the course of the world, incapable of being prepared for all the occurrences that might eventuate in it, I ask myself only: Can you will also that your maxim should become a universal law? If not, then it is reprehensible, and this not for the sake of any disadvantage impending for you or someone else, but because it cannot fit as a principle into a possible universal legislation; but for this legislation reason extorts immediate respect from me, from which, to be sure, I still do not have

insight into that on which it is grounded (which the philosopher may investigate), but I at least understand this much, that it is an estimation of a worth which far outweighs everything whose worth is commended by inclination, and that the necessity of my actions from *pure* respect for the practical law is what constitutes duty, before which every other motive must give way because it is the condition of a will that is good *in itself*, whose worth surpasses everything.

Thus in the moral cognition of common human reason we have attained to its principle, which it obviously does not think abstractly in such a universal form, but actually has always before its eyes and uses as its standard of judgment. It would be easy here to show how, with this compass in its hand, it knows its way around very well in all the cases that come before it, how to distinguish what is good, what is evil, what conforms to duty or is contrary to duty, if, without teaching it the least new thing, one only makes it aware of its own principle, as Socrates did; and thus that it needs no science and philosophy to know what one has to do in order to be honest and good, or indeed, even wise and virtuous. It might even have been conjectured in advance that the acquaintance with what every human being is obliged to do, hence to know, would also be the affair of everyone, even of the most common human being.

Here one cannot regard without admiration the way the practical faculty of judgment is so far ahead of the theoretical in the common human understanding. In the latter, if common reason ventures to depart from the laws of experience and perceptions of sense, then it falls into sheer inconceivabilities and self-contradictions, or at least into a chaos of uncertainty, obscurity, and inconstancy. But in the practical, the power of judgment first begins to show itself to advantage when the common understanding excludes from practical laws all sensuous incentives. It then even becomes subtle, caviling with its conscience, or with other claims in reference to what is to be called right, or even in wanting sincerely to determine the worth of actions for its own instruction, and, what is most striking, it can in the latter case do so with just as good a hope of getting things right as any philosopher might promise to do; indeed, it is almost more secure in this even than the latter, because the philosopher has no other principle than the common understanding, but the philosopher's judgment is easily confused by a multiplicity of considerations that are alien and do not belong to the matter and can make it deviate from the straight direction. Would it not accordingly be more advisable in moral things to stay with the judgment of common reason, and bring in philosophy at most only in order to exhibit the system of morals all the more completely and comprehensibly, and its rules in a way that is more convenient for their use (still more for disputation), but not in order to remove the common human understanding in a practical

NOTES

NOTES

respect out of its happy simplicity, and through philosophy to set it on a new route of investigation and instruction?

There is something splendid about innocence, but it is in turn very bad that it cannot be protected very well and is easily seduced. On this account even wisdom—which consists more in deeds and omissions than in knowledge—also needs science, not in order to learn from it but in order to provide entry and durability for its precepts. The human being feels in himself a powerful counterweight against all commands of duty, which reason represents to him as so worthy of esteem, in his needs and inclinations, whose satisfaction he summarizes under the name of 'happiness'. Now reason commands its precepts unremittingly, without promising anything to inclinations, thus snubbing and disrespecting, as it were, those impetuous claims, which at the same time seem so reasonable (and will not be done away with by any command). From this, however, arises a *natural dialectic*, that is, a propensity to ratiocinate against those strict laws of duty and to bring into doubt their validity, or at least their purity and strictness, and, where possible, to make them better suited to our wishes and inclinations, i.e., at ground to corrupt them and deprive them of their entire dignity, which not even common practical reason can in the end call good. Thus *common human reason* is impelled, not through any need of speculation (which never assaults it as long as it is satisfied with being mere healthy reason), but rather from practical grounds themselves, to go outside its sphere and to take a step into the field of *practical philosophy*, in order to receive information and distinct directions about the source of its principle and its correct determination in opposition to the maxims based on need and inclination, so that it may escape from its embarrassment concerning the claims of both sides and not run the risk of being deprived, through the ambiguity into which it easily falls, of all genuine ethical principles. Thus even in common practical reason, when it is cultivated, there ensues unnoticed a *dialectic*, which necessitates it to seek help in philosophy, just as befalls it in its theoretical use; and therefore the first will find no more tranquillity than the other anywhere except in a complete critique of our reason. […]

Acting according to the Concept of Law

But now in order to progress by natural steps in this work not merely from the common moral judgment (which is here worthy of great respect) to the philosophical, as has already been done, but also from a popular philosophy, which goes no further than it can get through groping by means of examples, up to metaphysics (which is not any longer held back by anything empirical and, since it must cover the entire sum total of rational cognition of this kind, goes as far as ideas, where even examples desert us), we must follow and distinctly exhibit

the practical faculty of reason from its universal rules of determination up to where the concept of duty arises from it.

Everything in nature works in accordance with laws. Only a rational being has the faculty to act *in accordance with the representation* of laws, i.e., in accordance with principles, or a *will*. Since for the derivation of actions from laws *reason* is required, the will is nothing other than practical reason. If reason determines the will without exception, then the actions of such a being, which are recognized as objectively necessary, are also subjectively necessary, i.e., the will is a faculty of choosing *only that* which reason, independently of inclination, recognizes as practically necessary, i.e., as good. But if reason for itself alone does not sufficiently determine the will, if the will is still subject to subjective conditions (to certain incentives) which do not always agree with the objective conditions, in a word, if the will is not *in itself* fully in accord with reason (as it actually is with human beings), then the actions which are objectively recognized as necessary are subjectively contingent, and the determination of such a will, in accord with objective laws, is *necessitation*, i.e., the relation of objective laws to a will which is not thoroughly good is represented as the determination of the will of a rational being through grounds of reason to which, however, this will in accordance with its nature is not necessarily obedient.

The representation of an objective principle, insofar as it is necessitating for a will, is called a 'command' (of reason), and the formula of the command is called an **imperative**.

All imperatives are expressed through an *ought* and thereby indicate the relation of an objective law of reason to a will which in its subjective constitution is not necessarily determined by that law (a necessitation). They say that it would be good to do or refrain from something, but they say it to a will that does not always do something just because it is represented to it as good to do. Practical *good*, however, is that which determines the will by means of representations of reason, hence not from subjective causes, but objectively, i.e., from grounds that are valid for every rational being as such. It is distinguished from the *agreeable*, as that which has influence on the will only by means of sensation from merely subjective causes, those which are valid only for the senses of this or that one, and not as a principle of reason, which is valid for everyone.*

NOTES

* The dependence of the faculty of desire on sensations is called 'inclination', and this always therefore proves a *need*. But the dependence of a contingently determinable will on principles of reason is called an *interest*. This occurs, therefore, only with a dependent will, which does not always of itself accord with reason; with the divine will one cannot think of any interest. But the human will, too, can *take an interest* without therefore *acting from interest*. The former signifies the *practical* interest in the action, the second the *pathological* interest in the object of the action. The first indicates only the dependence of the will on principles of reason in itself, the second on those principles of reason on behalf of inclination, where, namely, reason furnishes only the practical rule as to how the need of inclination is to be supplied.

NOTES

A perfectly good will would thus stand just as much under objective laws (of the good), but it would not be possible to represent it as *necessitated* by them to lawful actions, because of itself, in accordance with its subjective constitution, it can be determined only through the representation of the good. Hence for the *divine* will, and in general for a *holy* will, no imperatives are valid; the *ought* is out of place here, because the *volition* is of itself already necessarily in harmony with the law. Hence imperatives are only formulas expressing the relation of objective laws of volition in general to the subjective imperfection of the will of this or that rational being, e.g., to the human being.

Now all *imperatives* command either *hypothetically* or *categorically*. The former represent the practical necessity of a possible action as a means to attain something else which one wills (or which it is possible that one might will). The categorical imperative would be that one which represented an action as objectively necessary for itself, without any reference to another end.

Because every practical law represents a possible action as good, and therefore as necessary for a subject practically determinable by reason, all imperatives are formulas of the determination of action, which is necessary in accordance with the principle of a will which is good in some way. Now if the action were good merely as a means to *something else*, then the imperative is *hypothetical*; if it is represented as good *in itself*, hence necessary, as the principle of the will, in a will that in itself accords with reason, then it is *categorical*.

The imperative thus says which action possible through me would be good, and represents the practical rule in relation to a will that does not directly do an action because it is good, in part because the subject does not always know that it is good, in part because if it did know this, its maxims could still be contrary to the objective principles of a practical reason. The hypothetical imperative thus says only that the action is good for some *possible* or *actual* aim. In the first case it is a **problematically**, in the second an **assertorically** practical principle. The categorical imperative, which declares the action for itself as objectively necessary without reference to any aim, i.e., also without any other end, is valid as an **apodictically** practical principle. [...]

The Possibility of the Categorical Imperative

Now the question arises: How are all these imperatives possible? [...]

Regarding this problem we will first try to see whether perhaps the mere concept of a categorical imperative does not also provide us with its formula, containing the proposition which alone can be a categorical

In the first case the action interests me, in the second the object of the action (insofar as it is agreeable to me). In the First Section we have seen that with an action from duty it is not the interest in an object that has to be looked to, but merely the action itself and its principle in reason (the law).

imperative; for how such an absolute command is possible, even if we know how it is stated, will still demand particular and difficult effort, which, however, we will postpone until the last section.

If I think of a *hypothetical* imperative in general, then I do not know beforehand what it will contain until the condition is given to me. But if I think of a *categorical* imperative, then I know directly what it contains. For since besides the law, the imperative contains only the necessity of the maxim,* that it should accord with this law, but the law contains condition to which it is limited, there remains nothing left over with which the maxim of the action is to be in accord, and this accordance alone is what the imperative really represents necessarily.

The categorical imperative is thus only a single one, and specifically this: *Act only in accordance with that maxim through which you can at the same time will that it become a universal law.* Now if from this one imperative all imperatives of duty can be derived as from their principle, then although we leave unsettled whether in general what one calls 'duty' is an empty concept, we can at least indicate what we are thinking in the concept of duty and what this concept means. Because the universality of the law in accordance with which effects happen constitutes that which is really called *nature* in the most general sense (in accordance with its form), i.e., the existence of things insofar as it is determined in accordance with universal laws, thus the universal imperative of duty can also be stated as follows: *So act as if the maxim of your action were to become through your will a* universal law of nature.

Now we will enumerate some duties, in accordance with their usual division into duties toward ourselves and toward other human beings, and into perfect and imperfect duties:†

(1) One person, through a series of evils that have accumulated to the point of hopelessness, feels weary of life but is still so far in possession of his reason that he can ask himself whether it might be contrary to the duty to himself to take his own life. Now he tries out whether

NOTES

* A *maxim* is the subjective principle for action, and must be distinguished from the *objective principle*, namely the practical law. The former contains the practical rule that reason determines in accord with the conditions of the subject (often its ignorance or also its inclinations), and is thus the principle in accordance with which the subject *acts*; but the law is the objective principle, valid for every rational being, and the principle in accordance with which it *ought to act*, i.e., an imperative.

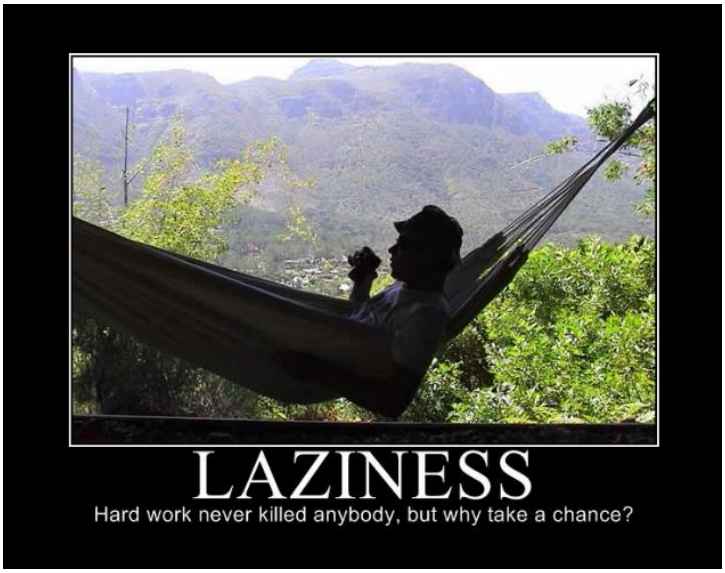
† Here one must note well that I reserve the division of duties entirely for a future *metaphysics of morals*; the division here therefore stands only as a discretionary one (to order my examples). For the rest, I understand by a perfect duty that which permits no exception to the advantage of inclination, and I do have *perfect duties* that are not merely external but also internal, which runs contrary to the use of words common in the schools; but I do not mean to defend that here, because for my aim it is all the same whether or not one concedes it to me.

NOTES

the maxim of his action could become a universal law of nature. But his maxim is: 'From self-love, I make it my principle to shorten my life when by longer term it threatens more ill than it promises agreeableness'. The question is whether this principle of self-love could become a universal law of nature. But then one soon sees that a nature whose law it was to destroy life through the same feeling whose vocation it is to impel the furtherance of life would contradict itself, and thus could not subsist as nature; hence that maxim could not possibly obtain as a universal law of nature, and consequently it entirely contradicts the supreme principle of all duty.

(2) Another sees himself pressured by distress into borrowing money. He knows very well that he will not be able to pay, but he also sees that nothing will be lent him if he does not firmly promise to pay at a determinate time. He wants to make such a promise; yet he has conscience enough to ask himself: ''Is it not impermissible and contrary to duty to get out of distress in such a way?'' Supposing he nevertheless resolved on it, his maxim would be stated as follows: 'If I believe myself to be in pecuniary distress, then I will borrow money and promise to pay it back, although I know this will never happen'. Now this principle of self-love, or of what is expedient for oneself, might perhaps be united with my entire future welfare, yet the question now is: ''Is it right?'' I thus transform this claim of self-love into a universal law and set up the question thus: ''How would it stand if my maxim became a universal law?'' Yet I see right away that it could never be valid as a universal law of nature and still agree with itself, but rather it would necessarily contradict itself. For the universality of a law that everyone who believes himself to be in distress could promise whatever occurred to him with the intention of not keeping it would make impossible the promise and the end one might have in making it, since no one would believe that anything has been promised him, but rather would laugh about every such utterance as vain pretense.

(3) A third finds in himself a talent, which could, by means of some cultivation, make him into a human being who is useful for all sorts of aims. But he sees himself as in comfortable circumstances and sooner prefers to indulge in gratification than to trouble himself with the expansion and improvement of his fortunate natural predispositions. Yet he still asks whether, apart from the agreement of his maxim of neglecting his gifts of nature with his propensity to amusement, it also agrees with what one calls 'duty'. Then he sees that, although a nature could still subsist in accordance with such a universal law, though then the human being [...] would think only of letting his talents rust and applying his life merely to idleness, amusement, procreation, in a word, to enjoyment; yet it is impossible for him to **will** that this should become a universal law of nature, or that it should be implanted in us as such by natural instinct. For as a rational being he necessarily wills that all the faculties in him should be developed, because they are serviceable and given to him for all kinds of possible aims.



(4) Yet a *fourth*—for whom it is going well, while he sees that others have to struggle with great hardships (with which he could well help them) —thinks: "What has it to do with me? Let each be as happy as heaven wills, or as he can make himself, I will not take anything from him or even envy him; only I do not want to contribute to his welfare or to his assistance in distress!" Now to be sure, if such a way of thinking were to become a universal law of nature, then the human race could well subsist, and without doubt still better than when everyone chatters about sympathetic participation and benevolence, and even on occasion exerts himself to practice them, but, on the contrary also deceives wherever he can, sells out, or otherwise infringes on the right of human beings. But although it is possible that a

NOTES

NOTES

universal law of nature could well subsist in accordance with that maxim, yet it is impossible to **will** that such a principle should be valid without exception as a natural law. For a will that resolved on this would conflict with itself, since the case could sometimes arise in which he needs the love and sympathetic participation of others, and where, through such a natural law arising from his own will, he would rob himself of all the hope of assistance that he wishes for himself.

Now these are some of the many actual duties, or at least of what we take to be duties, whose partitioning from the single principle just adduced clearly meets the eye. One must *be able to will* that a maxim of our action should become a universal law: this is the canon of the moral judgment of this action in general. Some actions are so constituted that their maxim cannot even be *thought* without contradiction as a universal law of nature, much less could one *will* that it *ought* to become one. With others, that internal impossibility is not to be encountered, but it is impossible to *will* that their maxims should be elevated to the universality of a natural law, because such a will would contradict itself. One easily sees that the first conflict with strict or narrow (unremitting) duty, the second only with wide (meritorious) duty, and thus all duties regarding the kind of obligation (not the object of their action) have been completely set forth through these examples in their dependence on the one principle.

Now if we attend to ourselves in every transgression of a duty, then we find that we do not actually will that our maxim should become a universal law, for that is impossible for us, but rather will that its opposite should remain a law generally; yet we take the liberty of making an *exception* for ourselves, or (even only for this once) for the advantage of our inclination. Consequently, if we weighed everything from one and the same point of view, namely that of reason, then we would encounter a contradiction in our own will, namely that objectively a certain

principle should be necessary as a universal law and yet subjectively that it should not be universally valid, but rather that it should admit of exceptions. But since we consider our action at one time from a point of view that accords entirely with reason, and then, however, also the same action from the point of view of a will affected by inclination, there is actually no contradiction here, but only a resistance of inclination against the precept of reason *(antagonismus)*, through which the universality of the principle *(universalitas)* is transformed into a mere general validity *(generalitas)*, so that the practical principle of reason is supposed to meet the maxim halfway. Now although this cannot be justified in our own impartially rendered judgment, it proves that we actually recognize the validity of the categorical imperative and (with every respect for it) allow ourselves only a few exceptions, which are, as it seems to us, insignificant and forced upon us.

Thus we have established at least this much: that if duty is a concept that is to contain significance and actual legislation for our actions, then this duty could be expressed only in categorical imperatives, but by no means in hypothetical ones; likewise, which is already quite a bit, we have exhibited distinctly and for every use the content of the categorical imperative which would have to contain the principle of all duty (if there is such a thing at all). But we are still not ready to prove *a priori* that there actually is such an imperative, that there is a practical law which commands for itself absolutely and without any incentives, and that it is a duty to follow this law.

The Absolute Worth of Persons

With the aim of attaining that, it is of the utmost importance to let this serve as a warning that one must not let it enter his mind to try to derive the reality of this principle from the *particular quality of human nature*. For duty ought to be the practically unconditioned necessity of action; thus it must be valid for all rational beings (for only to them can an imperative apply at all), and must *only for this reason* be a law for every human will. That which, by contrast, is derived only from what is proper to the particular natural predisposition of humanity, or from certain feelings and propensities, or indeed, if possible, from a particular direction of human reason, and would not have to be valid necessarily for the will of every rational being—that can, to be sure, be a maxim for us, but cannot yield any law; it can yield a subjective principle, in accordance with which we may have a propensity and inclination, but not an objective one, in accordance with which we would be *assigned* to act, even if it were to go directly contrary to all our propensities, inclinations, and natural adaptations; it even proves all the more the sublimity and inner dignity of the command in a duty, the less subjective causes are for it and the more they are against it, without on this account the least weakening the necessitation through the law or taking anything away from its validity.

NOTES

NOTES

Now here we see philosophy placed in fact at a perilous standpoint, which is to be made firm, regardless of anything either in heaven or on earth from which it may depend or by which it may be supported. Here it should prove its purity as self-sustainer of its own laws, not as a herald of those that an implanted sense or who knows what tutelary nature whispers to it, which, taken collectively, although they may be better than nothing at all, yet they can never yield the principles that reason dictates and that must have their source fully *a priori* and therewith at the same time their commanding authority: expecting nothing of the inclination of the human being, but everything from the supremacy of the law and the respect owed to it; or else, if that fails, condemning the human being to self-contempt and inner abhorrence.

Thus everything that is empirical is, as a contribution toward the principle of morality, not only entirely unfit for it, but even highly disadvantageous to the purity of morals themselves, in which precisely consists the sublime worth of a will absolutely good in itself and elevated above all price, that the principle of the actions is free of all influences of contingent grounds that only experience can provide. One cannot be given too many or too frequent warnings against this negligent or even base way of thinking, which seeks out the principle among empirical motivations and laws, since human reason in its weariness gladly reposes on this pillow and, in the dream of sweet illusions (which lets it embrace a cloud instead of Juno), supplants the place of morality with a bastard patched together from limbs of quite diverse ancestry, which looks similar to whatever anyone wants to see, but not to virtue, for him who has once beheld it in its true shape.*

The question is therefore this: Is it a necessary law *for all rational beings* to judge their actions always in accordance with those maxims of which they themselves can will that they should serve as universal laws? If it is, then it must be bound up (fully *a priori*) with the concept of the will of a rational being in general. But in order to discover this connection, one must, however much one may resist it, take one step beyond, namely to metaphysics, though into a domain of metaphysics that is distinguished from that of speculative philosophy, namely into the metaphysics of morals. In a practical philosophy, where what are to be established are not grounds for what *happens*, but laws for what *ought to happen*, even if it never does happen, i.e., objectively practical laws, there we do not find it necessary to institute an investigation into the grounds why something pleases or displeases, how the gratification of mere sensation is to be distinguished from taste, and whether the latter is distinct from a universal satisfaction of reason; on

* To behold virtue in its authentic shape is nothing other than to exhibit morality denuded of all admixture of the sensible and all ungenuine adornment of reward or self-love. How completely it eclipses everything else that appears charming to inclinations, everyone can easily be aware of by means of the least attempt of his reason, if it is not entirely corrupted for abstraction.

what the feelings of pleasure and displeasure rest, and how from them arise desires and inclinations, and from these, again, through the cooperation of reason, maxims arise; for all that belongs to an empirical doctrine of the soul, which constitutes the second part of the doctrine of nature, if one considers it as *philosophy of nature* insofar as it is grounded on *empirical laws*. Here, however, we are talking about objectively practical laws, hence about the relation of a will to itself insofar as it determines itself merely through reason, such that everything that has reference to the empirical falls away of itself; because if *reason for itself alone* determines conduct (the possibility of which we will investigate right now), it must necessarily do this *a priori.*

The will is thought as a faculty of determining itself to action *in accord with the representation of certain laws*. And such a faculty can be there to be encountered only in rational beings. Now that which serves the will as the objective ground of its self-determination is the *end*, and this, if it is given through mere reason, must be equally valid for all rational beings. By contrast, what contains merely the ground of the possibility of the action whose effect is the end is called the *means*. The subjective ground of desire is the *incentive*, the objective ground of volition is the *motive*; hence the distinction between subjective ends, which rest on incentives, and objective ones, which depend on motives that are valid for every rational being. Practical principles are *formal* when they abstract from all subjective ends; but they are *material* when they are grounded on these, hence on certain incentives. The ends that a rational being proposes as *effects* of its action at its discretion (material ends) are all only relative; for only their relation to a particular kind of faculty of desire of the subject gives them their worth, which therefore can provide no necessary principles valid universally for all rational beings and hence valid for every volition, i.e., practical laws. Hence all these relative ends are only the ground of hypothetical imperatives.

But suppose there were something *whose existence in itself* had an absolute worth, something that, as *end in itself*, could be a ground of determinate laws; then in it and only in it alone would lie the ground of a possible categorical imperative, i.e., of a practical law.

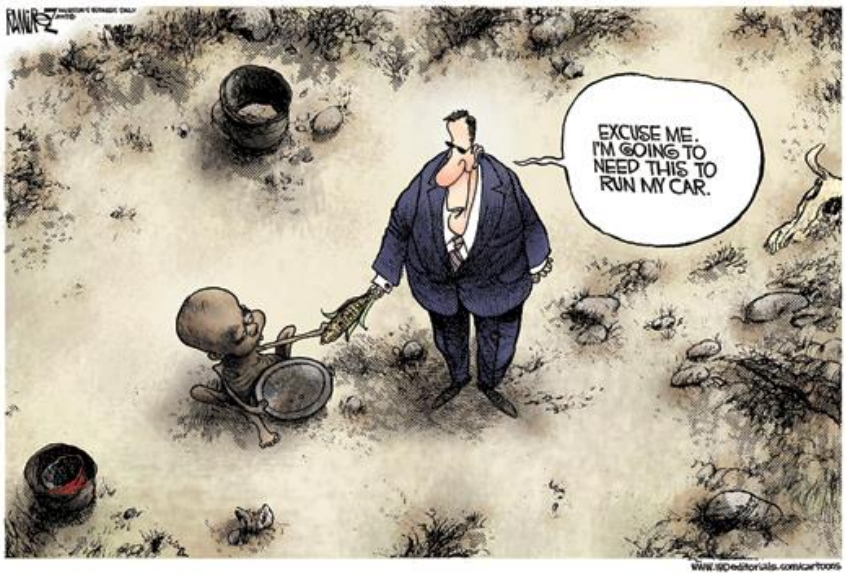
Now I say that the human being, and in general every rational being, *exists* as end in itself, *not merely as means* to the discretionary use of this or that will, but in all its actions, those directed toward itself as well as those directed toward other rational beings, it must always *at the same time* be considered as an *end*. All objects of inclinations have only a conditioned worth; for if the inclinations and the needs grounded on them did not exist, then their object would be without worth. The inclinations themselves, however, as sources of needs, are so little of absolute worth, to be wished for in themselves, that rather to be

NOTES

NOTES

entirely free of them must be the universal wish of every rational being. Thus the worth of all objects *to be acquired* through our action is always conditioned. The beings whose existence rests not on our will but on nature nevertheless have, if they are beings without reason, only a relative worth as means, and are called *things*; rational beings, by contrast, are called *persons*, because their nature already marks them out as ends in themselves, i.e., as something that may not be used merely as means, hence to that extent limits all arbitrary choice (and is an object of respect). These are not merely subjective ends whose existence as effect of our action has a worth *for us*; but rather *objective ends*, i.e., things whose existence in itself is an end, and specifically an end such that no other end can be set in place of it, to which it should do service *merely* as means, because without this nothing at all of *absolute worth* would be encountered anywhere; but if all worth were conditioned, hence contingent, then for reason no supreme practical principle could anywhere be encountered.

If, then, there is supposed to be a supreme practical principle, and in regard to the human will a categorical imperative, then it must be such from the representation of that which, being necessarily an end for everyone, because it is an *end in itself*, constitutes an *objective* principle of the will, hence can serve as a universal practical law. The ground of this principle is: *Rational nature exists as end in itself.* The human being necessarily represents his own existence in this way; thus to that extent it is a *subjective* principle of human actions. But every other rational being also represents his existence in this way as consequent on the same rational ground as is valid for me; thus it is at the same time an *objective* principle, from which, as a supreme practical ground, all laws of the will must be able to be derived. The practical imperative will thus be the following: *Act so that you use humanity, as much in your*



own person as in the person of every other, always at the same time as end and never merely as means. We will see whether this can be accomplished.

NOTES

Testing the Example Cases

In order to remain with the previous examples,

First, in accordance with the concept of the necessary duty toward oneself, the one who has suicide in mind will ask himself whether his action could subsist together with the idea of humanity *as an end in itself*. If he destroys himself in order to flee from a burdensome condition, then he makes use of a person merely as *a means*, for the preservation of a bearable condition up to the end of life. The human being, however, is not a thing, hence not something that can be used *merely* as a means, but must in all his actions always be considered as an end in itself. Thus I cannot dispose of the human being in my own person, so as to maim, corrupt, or kill him. (The nearer determination of this principle, so as to avoid all misunderstanding, e.g., the amputation of limbs in order to preserve myself, or the risk at which I put my life in order to preserve my life, etc., I must here pass over; they belong to morals proper.)

Second, as to the necessary or owed duty toward others, the one who has it in mind to make a lying promise to another will see right away that he wills to make use of another human being *merely as means*, without the end also being contained in this other. For the one I want to use for my aims through such a promise cannot possibly be in harmony with my way of conducting myself toward him and thus contain in himself the end of this action. Even more distinctly does this conflict with the principle of other human beings meet the eye if one approaches it through examples of attacks on the freedom and property of others. For then it is clearly evident that the one who transgresses the rights of human beings is disposed to make use of the person of others merely as a means, without taking into consideration that as rational beings, these persons ought always to be esteemed at the same time as ends, i.e., only as beings who have to be able to contain in themselves the end of precisely the same action.*

Third, in regard to the contingent (meritorious) duty toward oneself, it is not enough that the action does not conflict with humanity in our person as end in itself; it must also *harmonize with it*. Now in humanity there are predispositions to greater perfection, which belong to ends

* Let one not think that the trivial *quod tibi non vis fieri, etc*. [*What you do not want to be done to yourself do not do to another*] could serve here as a standard or principle. For it is only derived from that principle, though with various limitations; it cannot be a universal law, for it does not contain the ground of duties toward oneself, nor that of the duties of love toward others (for many would gladly acquiesce that others should not be beneficent to him, if only he might be relieved from showing beneficence to them), or finally of owed duties to one another, for the criminal would argue on this ground against the judge who punishes him, etc.

[Note that Kant here is explicitly pointing out how the CI is not at all the same thing as the Golden Rule.]

NOTES

of nature in regard to the humanity in our subject; to neglect these would at most be able to subsist with the *preservation* of humanity as end in itself, but not with the *furthering* of this end.

Fourth, as to the meritorious duty toward others, the natural end that all human beings have is their own happiness. Now humanity would be able to subsist if no one contributed to the happiness of others yet did not intentionally remove anything from it; only this is only a negative and not a positive agreement with *humanity as end in itself*, if everyone does not aspire, as much as he can, to further the ends of others. For regarding the subject which is an end in itself: if that representation is to have its *total* effect on me, then its ends must as far as possible also be *my* ends. This principle of humanity and of every rational nature in general *as end in itself* (which is the supreme limiting condition of the freedom of the actions of every human being) is not gotten from experience, first, on account of its universality, since it applies to all rational beings in general, and no experience is sufficient to determine anything about that; second, because in it humanity is represented not as an end of human beings (subjectively), i.e., as an object that one actually from oneself makes into an end, but as an objective end which, whatever ends we may have, is to constitute as a law the supreme limiting condition of all subjective ends, hence must arise from pure reason. The ground of all practical legislation, namely, lies *objectively in the rule* and the form of universality, which makes it capable of being a law (at least a law of nature) (in accordance with the first principle), but *subjectively* it lies in the *end*; but the subject of all ends is every rational being as end in itself (in accordance with the second principle): from this now follows the third practical principle of the will, as the supreme condition of its harmony with universal practical reason, the idea *of the will of every rational being as a will giving universal law*.

All maxims are repudiated in accordance with this principle which cannot subsist together with the will's own universal legislation. The will is thus not solely subject to the law, but is subject in such a way that it must be regarded also *as legislating to itself*, and precisely for this reason as subject to the law (of which it can consider itself as the author). Imperatives represented in the above way, namely of the lawfulness of actions generally similar to an *order of nature*, or of the universal *preference of the end* of rational beings themselves, just by being represented as categorical, excluded from their commanding authority all admixture of any interest as an incentive; but they were only *assumed* as categorical, because one had to assume such a thing if one wanted to explain the concept of duty. But that there are practical propositions which command categorically cannot be proven for itself here, just as little as this can still happen anywhere in this section; yet one thing could have happened, namely that the withdrawal of all interest in the case of volition from duty, in the imperative itself,

through any determination that it could contain, is indicated as the specific sign distinguishing the categorical from the hypothetical imperative, and this happens in the third formula of the principle, namely the idea of the will of every rational being as *a universally legislative will*.

For if we think of such a will, then although a will *that stands under laws* may be bound by means of an interest in this law, nevertheless it is impossible for a will that is itself supremely legislative to depend on any interest; for such a dependent will would need yet another law, which limited the interest of its self-love to the condition of a validity for the universal law.

Thus the *principle* of every human will as *a will legislating universally through all its maxims*, if otherwise everything were correct about it, would be quite *well suited* for the categorical imperative by the fact that precisely for the sake of the idea of universal legislation, it *grounds itself on no interest* and hence it alone among all possible imperatives can be *unconditioned*; or still better, by converting the proposition, if there is a categorical imperative (i.e., a law for every will of a rational being), then it can command only that everything be done from the maxim of its will as a will that could at the same time have as its object itself as universally legislative; for only then is the practical principle and the imperative it obeys unconditioned, because it cannot have any interest at all as its ground.

Moral Agents as Law-Givers to Themselves

Now it is no wonder, when we look back on all the previous efforts that have ever been undertaken to bring to light the principle of morality, why they all had to fail. One saw the human being bound through his duty to laws, but it did not occur to one that he was subject *only to his own* and yet *universal legislation*, and that he was obligated only to act in accord with his own will, which, however, in accordance with its natural end, is a universally legislative will. For if one thought of him only as subject to a law (whatever it might be), then this would have to bring with it some interest as a stimulus or coercion, because as a law it did not arise from *his* will, but rather this will was necessitated by *something else* to act in a certain way in conformity with the law. Through this entirely necessary consequence, however, all the labor of finding a supreme ground of duty was irretrievably lost. For from it one never got duty, but only necessity of action from a certain interest. Now this might be one's own interest or someone else's. But then the imperative always had to come out as conditioned, and could never work at all as a moral command. Thus I will call this principle the principle of the *autonomy* of the will, in contrast to every other, which on this account I count as *heteronomy*.

NOTES

NOTES

The concept of every rational being that must consider itself as giving universal law through all the maxims of its will in order to judge itself and its actions from this point of view, leads to a very fruitful concept depending on it, namely that of *a realm of ends*.

By a *realm*, however, I understand the systematic combination of various rational beings through communal laws. Now because laws determine ends in accordance with their universal validity, there comes to be, if one abstracts from the personal differences between rational beings, as likewise from every content of their private ends, a whole of all ends—(of rational beings as ends in themselves, as well as of their own ends, which each may set for himself) in systematic connection, i.e., a realm of ends—can be thought, which is possible in accordance with the above principles.

For rational beings all stand under the *law* that every one of them ought to treat itself and all others *never merely as means*, but always *at the same time as end in itself*. From this, however, arises a systematic combination of rational beings through communal objective laws, i.e., a realm that, because these laws have as their aim the reference of these beings to one another as ends and means, can be called a 'realm of ends' (obviously only an ideal). But a rational being belongs as a *member* to the realm of ends if in this realm it gives universal law but is also itself subject to these laws. It belongs to it *as supreme head*, if as giving law it is subject to no will of another. The rational being must always consider itself as giving law in a realm of ends possible through freedom of the will, whether as member or as supreme head. It can assert the place of the latter, however, not merely through the maxim of its will, but only when it is a fully independent being, without need and without limitation of faculties that are adequate to that will.

Morality thus consists in the reference of all action to that legislation through which alone a realm of ends is possible. But the legislation must be encountered in every rational being itself, and be able to arise from its will, whose principle therefore is: 'Do no action in accordance with any other maxim, except one that could subsist with its being a universal law, and hence only so *that the will could through its maxim at the same time consider itself as universally legislative*'. Now if the maxims are not through their nature already necessarily in harmony with this objective principle of the rational beings, as universally legislative, then the necessity of the action in accordance with that principle is called 'practical necessitation', i.e., *duty*. Duty does not apply to the supreme head in the realm of ends, but it does to every member, and specifically, to all in equal measure.

NOTES

The practical necessity of acting in accordance with this principle, i.e., duty, does not rest at all on feelings, impulses, or inclinations, but merely on the relation of rational beings to one another, in which the will of one rational being must always at the same time be considered as *universally legislative*, because otherwise the rational being could not think of the other rational beings as *ends in themselves*. Reason thus refers every maxim of the will as universally legislative to every other will and also to every action toward itself, and this not for the sake of any other practical motive or future advantage, but from the idea of the *dignity* of a rational being that obeys no law except that which at the same time it gives itself. In the realm of ends everything has either a **price** or a **dignity**. What has a price is such that something else can also be put in its place as its *equivalent*; by contrast, that which is elevated above all price, and admits of no equivalent, has a dignity.

That which refers to universal human inclinations and needs has a *market price*; that which, even without presupposing any need, is in accord with a certain taste, i.e., a satisfaction in the mere purposeless play of the powers of our mind, an *affective price*; but that which constitutes the condition under which alone something can be an end in itself does not have merely a relative worth, i.e., a price, but rather an inner worth, i.e., *dignity*.

Now morality is the condition under which alone a rational being can be an end in itself, because only through morality is it possible to be a legislative member in the realm of ends. Thus morality and humanity, insofar as it is capable of morality, is that alone which has dignity. Skill and industry in labor have a market price; wit, lively imagination, and moods have an affective price; by contrast, fidelity in promising, benevolence from principle (not from instinct) have an inner worth. Lacking these principles, neither nature nor art contain anything that they could put in the place of them; for the worth of these principles does not consist in effects that arise from them, in the advantage and utility that they obtain, but rather in the dispositions, i.e., the maxims

NOTES

of the will, which in this way are ready to reveal themselves in actions, even if they are not favored with success. These actions also need no recommendation from any subjective disposition or taste, regarding them with immediate favor and satisfaction, and no immediate propensity or feeling for it: they exhibit the will that carries them out as an object of an immediate respect, for which nothing but reason is required in order to *impose* them on the will, not to *cajole* them from it *by flattery*, which latter would, in any event, be a contradiction in the case of duties. This estimation thus makes the worth of such a way of thinking to be recognized as dignity, and sets it infinitely far above all price, with which it cannot at all be brought into computation or comparison without, as it were, mistaking and assailing its holiness.

And now, what is it that justifies the morally good disposition or virtue in making such high claims? It is nothing less than the *share* that it procures for the rational being *in the universal legislation*, thereby making it suitable as a member in a possible realm of ends, for which it by its own nature was already destined, as end in itself and precisely for this reason as legislative in the realm of ends, as free in regard to all natural laws, obeying only those that it gives itself and in accordance with which its maxims can belong to a universal legislation (to which it at the same time subjects itself). For nothing has a worth except that which the law determines for it. The legislation itself, however, which determines all worth, must precisely for this reason have a dignity, i.e., an unconditioned, incomparable worth; the word *respect* alone yields a becoming expression for the estimation that a rational being must assign to it. *Autonomy* is thus the ground of the dignity of the human and of every rational nature.

The three ways mentioned of representing the principle of morality are, however, fundamentally only so many formulas of precisely the same law, one of which unites the other two in itself.

WE, THE PEOPLE, RECOGNIZE THAT WE HAVE RESPONSIBILITIES AS WELL AS RIGHTS; THAT OUR DESTINIES ARE BOUND TOGETHER; THAT A FREEDOM WHICH ONLY ASKS WHAT'S IN IT FOR ME, A FREEDOM WITHOUT A COMMITMENT TO OTHERS, A FREEDOM WITHOUT LOVE OR CHARITY OR DUTY OR PATRIOTISM, IS UNWORTHY OF OUR FOUNDING IDEALS AND THOSE WHO DIED IN THEIR DEFENSE.

(BARACK OBAMA)

KANTIAN DEONTOLOGY

There is a *lot* going on in this lengthy selection from Kant! I'm going to break it down into chunks, so we can see more clearly what all he's arguing. Our discussion will begin with his argument for the *good will* being the greatest good, then we'll look at the distinction between *hypothetical* and *categorical* imperatives and look at how three formulations of the latter all logically amount to a single categorical imperative. In the process of doing this last thing, we'll look at some worries we might have about Kant's morality of duty, looking specifically at his own case study (making a deceitful promise) and two other apparent objections.

We'll then move on to the work of American philosopher Christine Korsgaard, who modifies Kant's theory by borrowing some characteristics of Rawls' politico-ethical theory.* Since Korsgaard's theory modifies Kant's, we'll treat it as a distinct approach to deontology, much like in chapter 17 we treated Hare's and Singer's as distinct approaches to utilitarianism.

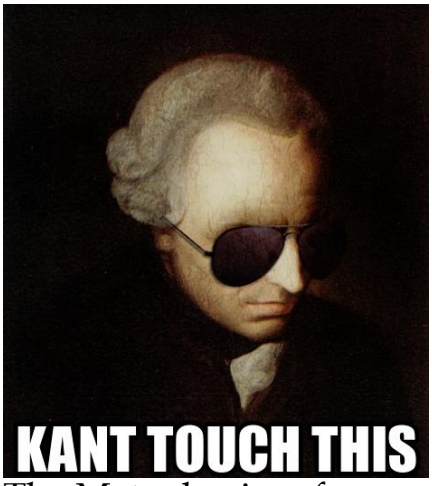
For now, on to Kant.

The Good Will

Step back. Look at utilitarianism for a moment. What does this theory presuppose regarding human nature? What sort thing is a human, must a human be? The good for a human, according to Bentham and Mill was the maximization of pleasure. To determine the good required an application of careful reasoning—the calculus. So human beings, at the most basic, are rational pleasure receptors and rational pleasure generators. That's crudely put, but it's really all that matters for morality, if you're a hedonistic utilitarian.

If you're a welfarist or preference utilitarian, it's a bit different. We're still rational, but we're a bit more than just pleasure receptors/generators. For Hare and Singer the core of morality is that humans are rational interest collectors or rational interest producers. Again, it's crude, how I've stated it here, but this is really all that matters in the determination of right and wrong actions.

If you're uncomfortable with this, you're not alone. Kant didn't like it. Of course, he didn't know anything of Mill or anyone who came along later, but hedonism and consequentialist ethics had been around since Epicurus. In the middle of the Enlightenment, Kant believed that the essence of a human being was more than pleasure generation or interest collection. He focused on the rationality part of the human equation. What in the world was this all about?



The Metaphysics of Morality

Kant, being a careful metaphysician,† seeks to build ethics on metaphysics. That is, it is meaningless to determine how something ought to behave without first knowing what the essential nature of that thing is. Imagine saying that it is an obligation for elephants to breathe freely under water. This is crazy talk nonsense simply because elephants aren't the sort of thing that even have the capacity to breathe under water. Thus, Kant begins his discussion with a consideration of the kind of thing we are.‡

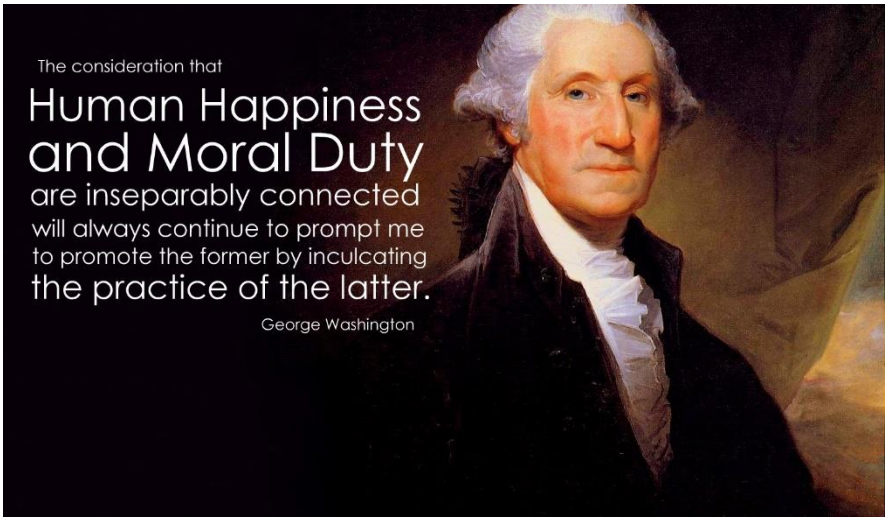
The first distinction Kant makes is between a priori and a posteriori reasoning,§ that is, between reasoning about things we can understand or determine without any reference to experience and those we cannot understand or determine except with reference to experience. The term ***a priori*** refers to those things we can know

* We'll look specifically at Rawls' theory in chapter 19.

† That is, a philosopher whose concern is primarily the nature of reality itself.

‡ Notice how this is an approach to the *is-ought* problem.

§ A distinction we discussed ourselves in chapter 11.



before any experience,* whereas ***a posteriori*** refers to those things that require experience to know or understand. You might see prior and posterior in these terms and remember the distinction as prior to experience and posterior to (behind) experience. For Kant, ethics has to date been considered a posteriori. But is it possible for us to find a standard that is a priori, a standard that is logically necessary and not dependent upon any experiences that a human might or might not have?

Such a standard would be the starting point for a ***pure moral philosophy***. Whatever the theory would be, then, would be free of cultural differences, social preferences, individual desires, and anything else that might make it seem limited. Kant is looking for something absolute, something that cannot be rejected as irrelevant because it fails to take into account some situation or other. Every principle that has been raised as the highest standard, Kant writes, was based on experience. Thus, if experience should change, the morality based upon such a standard would possibly become irrelevant. And this would not do. Kant seeks a universal standard, an unchanging standard.

Setting up such a standard is going to be tough. Whatever it is, Kant writes, it is going to have to demand that some action is moral not simply because it *happens* to agree with the standard S, but that it agrees with S *because* S is the standard. We don't say that somebody is a moral person because she just *happened* to fall upon the right action. We call that person lucky, not morally praiseworthy. We want, Kant writes, a morality that sets a standard that *determines* actions as moral or immoral much like mathematical formulae determine answers as right or wrong.

To find such a thing, we need to start with human nature. After all, it's *we* who will be the agents whose actions will be ethically weighted. And this takes us back to my question earlier. If utilitarians see us as reasoning pleasure receptors or reasoning interest producers, Kant sees us as *agents*. That is, we are things that act, and we act because we *want* or *will* things. Rationality itself demonstrates our willing to do things. A ***good will***, Kant argues, is the *only* thing that is desirable *for its own sake*.

Four Choiceworthy Things, With Qualification

What does this mean? He unpacks it by looking at other things we desire. Consider the *talents of the mind*, as he calls them. Intelligence, wit, judgment—all of these can be misused. Some of the most immoral persons who ever existed were brilliant minds. People make judgments that favor themselves and wreak horrors for millions of others. Consider *qualities of temperament* next—things like courage, resolution, perseverance. We call these *virtues*, but to be a courageous villain, to have unyielding resolve to commit some atrocity and to persevere in the carrying it out is anything but good. Thus neither the talents of the mind nor the qualities of temperament are good without qualification. These are good *if people use them well*.

Consider now *gifts of fortune* like power, or wealth, or honor, or health, or even happiness. A powerful or healthy or highly respected villain is a villain still. Do we see any of these as morally valuable in and of themselves? Contrary to what a utilitarian might say, Kant holds that any sort of prosperity—including the contentment of met interests or the satisfaction of preferences—

* And until Kant's *Critique of Pure Reason* (also called the *First Critique*), all *a priori* reasoning was considered to be *analytic*. Remember the distinction between analytic and synthetic as discussed in chapter 16, specifically when we looked at Ayer's emotivism.

can easily be twisted into arrogance and inhumanity. Imagine an ideal, impartial spectator. This person would not see such as good without qualification, rather, would say that it is *better* that these gifts of fortune go to *good* people than to *wicked* people (recognizing for now that we're speaking loosely).

We might say that *moderation* is certainly desirable. People who reason calmly and clearly, people who are self-controlled and not prone to emotional or other excesses seem better to us. But even this, Kant argues, is only good with qualification. Consider the difference between a murder of passion and one that's done with cool detachment. We find the latter all the more horrifying simply *because* it's so coolly moderate.

What all of these lack, what it is that would make all of these good, is the *good will* itself. And the good will is not good for whatever consequences it might effect, but simply for its own sake. Consider again. Suppose there's a person who means well, who does the right thing whenever possible, even though all goes to hell around him. We see that person even still as *good*, and we might even emulate him. Thus Kant's theory begins with a sharp break from utilitarianism: it's adamantly ***non-consequentialist***. Whether the good willed person effects any good consequences or whether that person receives only ill fortune,*

Kant's Function Argument

Is this intrinsic, unqualified value of the good will a "high-flown fantasy" or is it grounded in careful reasoning? What reason do we have to think that a *good will* is better than *happiness*? Elsewhere, Kant says that happiness isn't an ideal of reason, but of imagination. Hard words for the utilitarian. Kant here throws out the Greatest Happiness Principle as poorly-reasoned fluff. But can Kant justify the good will as better than happiness? To answer this, Kant offers an argument that I'll here call ***Kant's Function Argument*** (KFA):

KFA

1. In living beings, no organ (or faculty) will be found that isn't the fittest and best adapted to its purpose.
2. Instinct would be a better guide to happiness than practical reason is.
3. So, happiness cannot be the sole or primary purpose of practical reason.
4. So, there must be another and higher purpose for practical reason.

---

5. So reason's proper function (or purpose) must be to produce a will that is good in itself.

Let's break this down. Premise 1 says that every part or faculty of any living being is best suited for its unique purpose. For example, our eyes are best suited to take in light and thereby to see. Our hearts are best suited to pump blood, and so on.

Now we have this thing Kant calls ***practical reason***. Practical reason we might nowadays rather call *applied* reason. ***Pure Reason***, in Kant's terminology, refers to the reasoning faculty that enables us to make *a priori* judgments. It's logic, mathematics, and even some introspection. But practical reason refers to the use of reason in practice—how to live right, how to reason well. It's *applied* reasoning. Premise 2 considers the consequentialist claim that the aim of practical reason—of ethical reasoning—is *happiness*. Whether or not one is an egoist or a utilitarian, one will say that it's happiness that is the measure of good ethical judgements. But consider anything else capable of happiness. And consider even ourselves. If happiness were all that mattered, our unreflective instincts would get us happiness far easier than careful reasoning. In fact, those who are far more capable reasoners—great minds—are often less likely to find easy happiness than those who are less capable.† People who seek to do the right thing—that is, people who engage in practical reason—often make their lives more difficult and less pleasant by so doing. On the other hand, our instincts seem to lead us right into

* You might have in mind here the Ring of Gyges thought experiment presented in chapter 11, where Glaucon (Plato's brother) wants to see whether it is better to be wholly just and impoverished, hated, and alone in one's virtue or to be wholly unjust and wealthy, beloved, and mistakenly respected by all as virtuous. Later in the *Republic* (in fact, the whole rest of the book) Plato argues (through Socrates) that the former is indeed the better position. Kant here clearly agrees. Justice—or in this case the good will—is choiceworthy for its own sake, not for what it might get you.

† And in fact, Kant notes here that this is why those who think happiness is the end-all be-all are anti-intellectualist, or what he calls *misologic*—haters of reason. If it's all about happiness, then good reasoning seems to be useless and even destructive. But what if happiness is only a corollary or even perhaps a happy by-product of what really matters?

contentment and happiness. And our practical reason tends, at times—and even often—to steer us to paths contrary to our instincts. Thus, instincts are a better guide to happiness than practical reason.

It follows, Kant concludes in premise 3, that happiness can't be the purpose of practical reason. So what is practical reason aimed towards? It certainly influences our will. It tells us what we *ought* and *ought not* do. It tells us sometimes that we *ought* to disregard our instincts. So, since our faculty of practical reason isn't aimed to happiness, it must be aimed for something else—something better perhaps. Such is his claim in premise 4.

Since our practical reason guides our will, and since it's not the greatest at guiding us to happiness, it must be a faculty for guiding our will not as a *means* to other things, but to make the will *itself* good. Go back to all those things we think are qualifiedly good. Each of them is better with a good will. And since our practical reason aims to guide our will not towards any ends, it seems right to conclude that the purpose of practical reason is simply to make our will good. To make our will something that will make all the other things we desire better. To make something that even if we didn't have all the other things would itself be good.

Thus Kant concludes that the good will is the *highest* good, even though it's not the only good or even a *complete* good. To have a good will is good, but it's even better with happiness. But one cannot be truly happy unless one already has a good will. Happiness is a *secondary* good that complements the good will. In sum, we should have two goals: a good will, and only secondarily to this, happiness.

So yay for that. But what *is* a good will?

Four Kinds of Actions

A good will is a will that is properly directed. But what is the right direction? When Kant—or for that matter, most of us—talk about doing the right thing, we're usually talking about our moral obligation, our moral *duty*. So whatever a good will is, it will be directed towards doing one's moral duty. But whatever this is, it's certainly not yet clear. We can understand what the good will is by talking about motivations (the will) and their relationship to moral duty.

The first kind of action I might undertake, when we think about motivations, is simply *contrary* to my duty. Maybe I'm being a jerk. Maybe I'm seeing that whatever the moral obligation, I just don't wanna. Or maybe I think my obligation is too onerous, and I don't want to suffer whatever personal costs or inconveniences it would entail. Whatever it is, this motivation is clearly outright wrong. And it's clearly not going to be an indicator of a good will.

The second kind of action I might undertake is in accordance with duty, but done from some *indirect inclination*. I don't want to do my duty—don't give a rip about duty—but what I wind up doing agrees with my duty by chance.

Say I own and run a candy store, and as a consequence, it's my obligation to treat patrons fairly, especially by not overcharging them. But it happens that, on this day, the only customers I have are small children. It's a scene from "Willy Wonka and the Chocolate Factory," and I'm the Candyman. But I'm not as friendly, musical, or kind as the character in the film; instead, I'm a person intent on getting as rich as I can as fast as I can. All I want is money money money.

I *could* cheat the children, charging them double or even triple the cost of the candy. Who would know? Yeah, that's right. I have to keep the books clean if I want to get rich and *stay* rich. So, out of a selfish inclination, a desire to preserve my own skin from ugly legal consequences—including potential audits and so forth—I refrain from overcharging the children.

It's my duty to treat customers fairly, and I treat them fairly. But I don't treat them fairly *because* it's my duty to do so, rather, I treat them fairly because I'm afraid that if I don't, I'm going to suffer in the pocketbook. That's not a good will, clearly. I don't give a rip about the children, just about myself. My *direct* inclination is simply for me; only *indirectly* am I motivated to act in a fair manner.

A third kind of action comes from a *direct inclination* to do the action. This kind of action is also in accordance with duty. But again, my motivation in such a case is to do whatever the action is because I want to do it because I'm sympathetic, not because that

action is *simply the right thing to do*. Suppose now I'm a next-door neighbor to some elderly gentleman, and I feel sorry for him. It is my duty to ensure the well-being of others, but I wish to ensure his well-being because I like him. In the winter I blow his snow and keep the ice melted for him. I do whatever I can to make his life pleasant. Ain't I sweet?

Kant says that such an action is praiseworthy, but not worth anything at all *morally*. We encourage people to act sympathetically. It's a good thing. But it doesn't rise to *moral* value. It is good to be sympathetic towards others, but sympathy is not a solid grounding for morality.

Just like any emotion, it's no good indication of truth, and it cannot stand as the foundation of an unchanging moral standard. Suppose I one day found out this elderly gentleman was, before he moved next door, a war criminal? My sympathy would probably evaporate, and along with it, the inclination to ensure his well-being. But my duty towards ensuring the well-being of human beings would not evaporate.

Suppose now that he's a former war criminal, which I now know, and that he's done his time and repaid his debt as far as he can, even to the extent of giving his every penny to support the children of those whom his earlier actions had destroyed. But I don't know that he's reformed, only that he was a war criminal at one time. Were I to learn of his reformed ways, my sympathy might re-emerge, and I would probably wish to ensure his well-being again. How fickle! If the standard of morality is our sympathies, it will

DO WHAT IS RIGHT, THOUGH THE WORLD MAY PERISH.

(IMMANUEL KANT)

Schiller's Objection

To summarize so far, we can see that self-interest is not a morally reliable motive, contrary to the utilitarian or ethical egoist. In fact, sometimes—often enough to be obscenely noticeable in the daily news—we can get ahead or satisfy our personal inclinations without any actions that even remotely agree with our moral duty. Sympathy isn't a morally reliable motive either. We can be sympathetic towards the wrong people.

Suppose you were sympathetic towards somebody and always helped that person by keeping his van in good working order, but unknown to you, the only use he had for that van was to pick up and molest children? Here you are, sympathetically maintaining a pedo-van. In fact, it is sympathy that often directs us to do things *contrary* to duty, like when people sympathetically enable addicts to continue ruining their own and others' lives. Sympathy is too changeable and too easily directed. Like emotions are not a good indicator of truth, sympathy is not a reliable motive for morality. Neither self-interest nor sympathy are directly interested in one's moral duty. Thus, neither is the ground of morality or the indicator of a good will.

But it seems like, especially when you consider the example of the child-averse savior, Kant is expecting us to become *un*sympathetic. It seems like he wants us to eradicate emotion and sympathy altogether. Do we have to become cold, unfeeling agents of goodness? Is this what Kant means by the good will? Friedrich Schiller, a contemporary of Kant and a well-known German poet and philosopher thought so. He mocked Kant's theory by writing this acerbic poem:

continued…

ebb and flow like the tide on a windy day—far from the universal, unchanging standard of morality.

The point is that sympathy is good, but it is far too unreliable—too changing—to be the motivation that rises to moral worth. So a will that operates from sympathy is to be praised, but it's not good enough to be that truly moral paragon, the good will.

The final kind of action is that action that is both in accordance with duty and, more importantly, performed *because* the action is one's duty. It is *performed from duty*. It's doing the right thing not because you care for the one to whom your duty is directed, not because you're afraid of the consequences if you don't do the right thing, but because—gosh darn it—it's the right thing to do. Period. There's no sympathy in the motivation. There's no malice in the motivation. It's just doing the right thing.

This can be difficult to wrap your mind around. Suppose now I'm doing the right thing out of duty. Say it's my duty to save a drowning child, as in Singer's example in chapter 17. Now suppose further, I don't like children. I mean, I *really* don't like children. But I see that child suffering, and I realize it's *my* duty to intervene. I don't feel any sympathy for the child—maybe I even feel aversion, like I saw the child acting stupidly and creating the situation that she's now in. It's her own damn fault. But I intervene anyway, even if I don't want to, even if I'm not sympathetic, because *it is the right thing to do*.

Maybe this seems cold hearted to you. Change the situation a bit. Say you see some jerk in a life-

Schiller's Objection, *continued.*

> *Gladly I serve my friends, but alas I do it with pleasure.*
> *Thus I am plagued with doubts that I am not virtuous.*
> *To this, the answer is given:*
> *Surely your only choice is to try to loathe them entirely,*
> *And then with aversion do what your duty enjoins you.*[*]

1. Moral worth depends on not doing something by / from inclination.
2. Thus, moral worth requires I seek to despise my friends and do my duty with repugnance.
3. But 2 is just crazy.

4. So 1 must not be true.

Premise 1 and its conclusion (2) are stated in the poem; premise 3 and the final conclusion are clear from the biting sarcasm. Let's look at the objection more carefully. Premise 1 of the objection notes that the morally valuable action is only morally valuable if it is not done by inclination—by sympathy. This is right. But he concludes in 2 that therefore we must cobble together an aversion, an inclination *away from sympathy*. But Kant doesn't say this. He says that we cannot call an action that arises from sympathetic inclination morally praiseworthy, not that we cannot act from duty and also have sympathy.

Notice what Schiller does. He replaces one inclination with another! He replaces an inclination towards sympathy with an inclination towards aversion. He seems to be saying that we must act from an inclination of aversion in accordance with duty, but this isn't what Kant says at all. Kant isn't saying that the mere *presence* of sympathy renders an action as less than morally valuable. But this is what Schiller presumes Kant means. Kant says that sympathy can't be the *motivation* of an action, if that action is to have moral worth.

Remember that Kant says that actions done from sympathy are praiseworthy, that we are to encourage such behavior. It's *a good thing* to be sympathetic. Keep it coming. Sympathetic actions are good actions, just not *esteemed* actions, not *morally* good actions; thus Schiller's claim that we should loathe people entirely and act from total aversion is a ***straw man***. The straw man is a fallacy, where one misrepresents another's argument or claim and then knocks down the misrepresentation. It's like setting up a straw man and knocking it over, then crowing that you clobbered the original person.

[*] Although this epigram is widely attributed by scholars to Schiller, it isn't a full statement of his response to Kant, which is found in Schiller's 1793 essay, "On Grace and Dignity."

continued…

threatening situation. Say it's somebody you really can't stand, somebody you're careful to avoid. You're the only one around, and if you don't intervene, this person will die. If you do intervene, you're pretty sure he's not going to suddenly reform and treat you nicely. He'll probably go on his jerky way, persisting in being a total jerk, maybe even more so than before. But you know it's the right thing to do, to save his life. So you do. *That's* what Kant means by acting *from* duty.

This last kind of action is the *only* kind of action that has real moral worth. It's the only one that comes from a truly good will. Kant writes that for an action to be *morally* valuable, it must be done solely from duty. Not from an inclination of sympathy. Not for desired outcomes.

Acting from the motivation of sympathy and acting from the motivation of duty will often (but not always) have the *same* consequences. But the principle of the good will is ***respect for the moral law***. A sympathetic person might or might not have this respect, but a dutiful person will clearly have it, and it is *this* that marks the good will.

Schiller's Objection, *continued.*

Kant doesn't say that the mere presence of an inclination counts against the moral worth of an action. But it's a heck of a lot easier to *know* an action is morally praiseworthy when that inclination isn't there or when there's a disinclination towards the action. Notice that Kant points to the *role* of the inclination, not the *presence* of the inclination. If the inclination is what moves you to act, if it is what determines your will, your choosing, then your action isn't morally valuable. When you are sympathetic, it's very hard—if not impossible—to determine what is the motivation of your action. It might be duty, it might be sympathy, so the moral worth of the action is unknown. But this is not at all the same thing as saying you have to be *disinclined* to act in order to be moral in the action.

Kant does not at all say that we are to cultivate an aversion to others or a repugnance towards duty. That is, the very core of Schiller's objection misses the mark entirely. Sympathetic actions are admirable and to be encouraged. Rather, we are to cultivate a motive towards duty—not indifference. Cultivating duty does not require us to actively seek out any rogue sympathy and drown it in psychological Round-Up, thus ensuring its total demise.

We can set aside Schiller's Objection—which many others have also had upon their first, unreflective reading of Kant's argument. And with this set aside, we're ready to look at the different kinds of imperatives.

Hypothetical & Categorical Imperatives

First off, an ***imperative*** is a command.* We can understand it as a statement that contains an *ought*. But not all imperatives are created equal. Some are only conditional, dependent upon what one might want. Take these three imperatives:

If you want to bisect a line segment, draw intersecting arcs from the end points.

If you want to stay financially comfortable, be sure to save some money for the end of the quarter.

If you want to keep your job, make a habit of showing up on time and working hard.

Each of these is only relevant for those who actually *want* what's in the antecedent: to bisect a line segment, to stay financially comfortable while in college circumstances, to keep a job.

* See chapter 3.

These are imperative *only as a means to some desired goal*. Kant calls these kinds of imperatives ***hypothetical***, because they're only imperatives on the hypothesis that one desires the relevant end:

> *X is a **hypothetical imperative** iff x is an imperative that states a necessary condition for goal that not all persons will have.*

Not everyone wants to bisect a line segment, contrary to your firmly held beliefs, I'm sure. Not everyone wants to keep a job—for example, some people aren't even employed at all, hence *keeping* a job is irrelevant. Maybe they're retired. Maybe they're independently wealthy. Maybe they're only six years old. In such cases, the imperative doesn't apply to them. It only applies *on the hypothesis* that one desires a certain end.

In contrast, a ***categorical imperative*** is one that applies to *everyone* who fits in the category *human being*. This kind of *ought* statement says that some action is good and ought to be done without *any reference to any other purpose* or end. Categorical imperatives are not a means to some other end like hypothetical imperatives. Such an imperative is necessary, unconditional, and universal. No ifs, ands, or buts. A categorical imperative applies to everyone, regardless their goals, desires, life situations, or anything else that might direct their actions.

> *X is a **categorical imperative** iff x is an imperative that states a necessary condition that applies universally to all members that fall into the category of rational things (i.e., to all things that have rationality).*

It follows then, since we're looking for an *a priori* standard for morality, that moral obligation should be based on a categorical imperative, not a hypothetical one.

Again, notice the contrast with utilitarianism. We could restate the Greatest Happiness Principle thus: *if you want to maximize happiness, then do x*. It's a hypothetical imperative. Notice that *any* consequentialist maxim will be a hypothetical imperative. It thus cannot be the supreme principle of morality for Kant, since it relies on the whims of people's fickle inclinations and desires.

The Formulation of Universal Law

So what is the supreme principle of morality? I've said it before and I'll say it again: ***duty***. We can now understand it more clearly:

> *X is **duty** iff x is the necessity of an action executed from the respect for the moral law.*

That is, duty is the non-negotiableness of an action, the '*musting' to do*, and it is this necessity to do something *from the motive of respecting moral law*. What in the … ?

Okay, break it down. We have taken away inclinations like sympathy. We have taken away consequences altogether, hence removed any and all hypothetical imperatives. All that's left is the idea of a universal moral law, the idea of conforming to the idea of law. That's it. That means there's actually only ***one*** categorical imperative—which we should now properly capitalize as the ***Categorical Imperative*** or the CI.

There are three *formulations* of this one Categorical Imperative. It's like a three-legged stool or a three-sided prism (yes, I know that's physically impossible. Humor me.) It's one imperative that we can understand three ways, through three aspects. Logically, Kant believes, all three ways amount to the same thing. These formulations each look at the CI from a different angle. The first, the ***Formulation of Universal Law*** or **FUL** expresses the CI as an implementation of the universal moral law.

The FUL states that we should act in such a way that what we do can at the same time be a universal law of nature. As Kant states it,

> **FUL:** *Act only according to that maxim by which you can, at the same time, will that it be a universal law.*

The FUL looks at the Categorical Imperative strictly from the view of pure *a priori* logic. To see how it works, let's look at Kant's own example. Say you're confronted with the following moral question: *should I make a deceitful promise to get myself out of a difficult situation*? You might think, it would certainly get me helpful consequences. It might even maximize happiness, in which case it would be the moral thing to do for a utilitarian. But the CI stands outside of consequences. Applying the FUL to this question is completely different than testing possible outcomes according to the GHP.

There are *two* parts to the FUL: the *maxim* and the *universalization* of the maxim. For Kant,

What the FUL is not

The Formula of Universal Law is certainly *similar* to what your mom would say when you wanted to do something because your friends could.

"What if everyone did that?"

In my home, it was *if everyone jumped off the Monroe Street Bridge, would you need to, too?* But although the application of the FUL seems similar to this, it isn't really the same thing at all.

Moms around the world say this to point out the faulty reasoning of peer pressure, whereas Kant doesn't give a rip about what anyone thinks or does in this imagined possible world. Rather, he looks to see whether there arises a *logical contradiction* with the conjunction of M and U. He's looking to see if the conjunction creates an incoherency. The idea is that immoral acts are illegitimate because illogical, not that they are illegitimate because based on popular opinion.

continued…

> *X is a **maxim** iff x is a statement by rational agent A of an action that A wants to perform.*

Suppose your maxim is as follows: "When I am in financial difficulty, I will borrow money and promise to repay it, even though I know I won't really ever pay it back." Let's call that maxim ***M***. The question is, then, can I logically will *both* that M and that M is universalized? We can say that

> *Maxim M is **universalized** iff M is a law of nature that all rational agents always perform M.*

We'll refer to the universalization of some maxim as ***U***. So the FUL looks to the logical consequences of a theoretical state of affairs where both M and U would be true. Can I simultaneously will that *I* borrow money, making a deceitful promise to repay and that the world be such that *everyone always* borrows money, always while making deceitful promises to repay?

The test of the Formulation of Universal Law is to see whether there exists some possible world where both M and U obtain.* We are ***not*** looking to see whether such a world would be pleasant—which would be to look for consequences—but to see whether such a world is logically possible. If the conjunction of M and U creates a contradiction, then we know that the action described by our maxim is contrary to the Categorical Imperative, hence is immoral. If the universalization of a maxim makes the action in the maxim impossible, then the action is impermissible.

So let's test the maxim we stated above. Suppose it is a universal law of human nature that everyone always lies about their intention to repay a loan. Is it then possible to borrow money with a deceitful promise to repay it? Not so much. In such a world, everyone would *know* that you're lying, since that's what everyone always does. Thus, it would be impossible to borrow money. It'd still be possible that people would *give* you money, but a loan includes an expectation of repayment, which everyone in such a world would recognize as irrational because everyone's every repayment promise repay is known to be a lie. It follows that the action in the maxim—borrowing money with a deceitful promise to repay—is immoral.

Here's a task to get you into seeing how to apply the FUL. Think of *five* different actions you engage in that you think have some moral weight. These can be actions you considered from the last chapter, or they can be different ones. For each one, determine your maxim M and the universalization of it U. Then test to see whether the conjunction (M & U) creates a logical impossibility, an incoherency. Does the application of the FUL deem your actions moral or immoral? If the latter, where's the contradiction? Label this as Task 71, and have it ready to turn in when this reading is discussed.

Here's an example. Say I'm trying to determine whether I can cheat on my taxes because I just don't like having to pay. I would write

* Remember possible worlds talk from chapter 10.

What the FUL is not, *continued.*

It's also nothing like the Golden Rule. Sometimes Moms say things like *if everyone did that…* to arouse your sense of sympathy towards others. Kant notes that sympathy is good, and he encourages us to act with sympathy, but since the FUL is a test of whether an action is based on *duty*, it is really nothing like considering how you would want to be treated. Because the Golden Rule is based on desires, it cannot be a universal principle of morality. Because the Golden Rule is based on empathy, it cannot be what the FUL is exploring.

Note that when we're testing an action by the FUL, it is looking at what is logically possible, *not* at whether potential consequences are something I'd like for myself or for everyone. It is not asking

> "would I like a world where everyone did M?"

but

> "is a world where everyone did M and where I wish to do M a logically possible world?"

down my maxim as something like this:

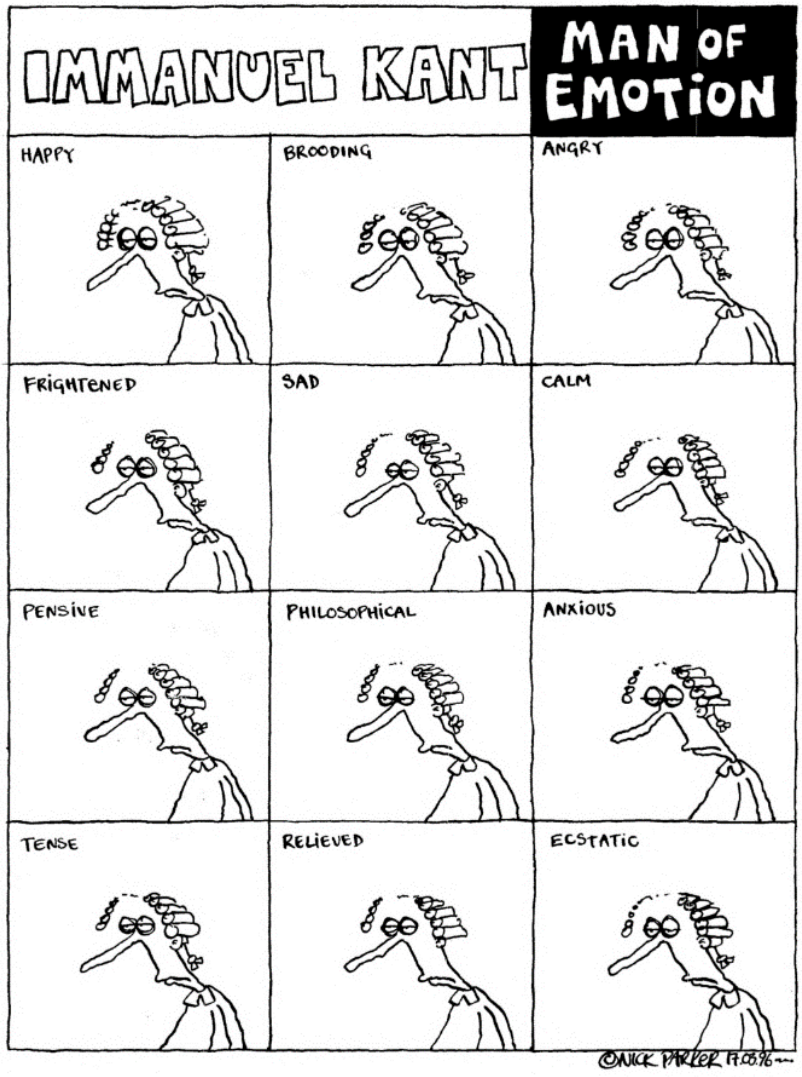
M: When I fill out my tax forms, I will cheat so that I don't have to pay.

Then I would figure out what the universalization of that would look like. I might write something like this:

U: It is a universal law of nature that every time somebody fills out their tax forms, they cheat so that they don't have to pay.

Now I test the conjunction of M and U. What would a world where everyone cheats on their taxes look like? Would it be possible for *me* to cheat on my taxes? Well, in such a world, the IRS would know everyone cheats and would expect it. But cheating *just means* that the one you're cheating doesn't *know* that you are. It requires deception, but here that's impossible. So it would be impossible for me to cheat. Contradiction.

Now you try.



Problems with the FUL?

The good will, that morally excellent, highest good thing is a will that acts always from (or at least acts in accord with) the Categorical Imperative. It's a will that guides one to live—'FUL'ly. A good will, Kant writes, is a will whose principle is *always* to act only upon *universalizable* maxims, regardless of any other inclination for or against that action.

False Negatives

The FUL is designed to test actions to see whether they are moral or immoral. The outcome is certain, without doubt. If the FUL test brings up a contradiction, the action in question is ever and always immoral—period. It always boils down to whether M + U creates a contradiction. And it should work for *every maxim* one might have.

A maxim is a statement of an action some agent wants to perform. Some actions I perform include making coffee for breakfast, sleeping in a bed that faces northwest, scratching my cats on their furry cheeks, listening to music on my computer, and facing a window while writing. One thing I love is barbecued salmon. So one maxim I might test by the FUL is "when I'm hungry, I'll have barbecued salmon for dinner." To test it, I must then see whether the universalization of this maxim creates a contradiction. The universalization of this would look like such: *it's a universal law of human nature that, when hungry, everyone always has barbecued salmon for dinner*.

And it's pretty clear pretty fast that this is impossible. First off, there's not enough salmon to feed the whole world. Secondly, there are parts of the world where salmon isn't even accessible or affordable. It's just not possible for everyone to eat salmon, let alone *barbecued* salmon. And supposing it were for a time possible, certainly the demand would outrun the supply, and salmon would go extinct. It seems then, that it is immoral for me to eat barbecued salmon for dinner. If the FUL accurately represents the supreme principle of morality, the Categorical Imperative that applies to all human beings, then I cannot in good conscience ever have salmon again.

But this seems odd. Change that to *steak*. Or *Kraft dinner*. Or *chicken*. Or *ramen noodles*. The same contradiction will invariably arise. But it seems patently obvious that it's not immoral to have salmon or steak or chicken or soybean patties or whatever else you want to name for dinner.

Or suppose my maxim is "When I need to relax, I'll hang out on my couch for a couple of hours." But it gets *really* ridiculous when I universalize that. There's not enough room for every human being to hang out on my couch. If it were a universal law of human nature that everyone always hung out on my couch when they needed to relax, there'd be no room for me. And it would be anything but relaxing.

The application of the FUL thus seems to have a problem with ***False Negatives***. that is, it marks things as immoral—as *morally blameworthy*—that are not. There are clearly *permissible* maxims that seem to be *prohibited* by the FUL. Let's spell this worry out into a full-fledged objection, which I'll call the ***Objection from False Negatives***:

FALSE NEGATIVES

1. Deontology holds the Categorical Imperative, which is presented in the Formulation of Universal Law (FUL), as the supreme principle of morality.
2. A supreme principle of morality is a principle that is universal, unchanging, and without exceptions.
3. The FUL says that an action A is immoral if one cannot consistently will both a maxim to commit A and the universalization of that maxim.
4. Actions like hanging out on my couch when I want to relax or eating barbecued salmon for dinner, according to the FUL, are immoral.
5. But clearly it isn't immoral—or isn't always immoral—to hang out on my couch when I want to relax or to eat barbecued salmon for dinner.
6. So there are at least two exceptions to the FUL.

---

7. So the FUL cannot be the supreme principle of morality.

Responding to the False Negatives Objection

How in the world might Kant be able to get out of this fix? The problem seems to arise from unnecessary specificity. That is, my maxim shouldn't be the easy, surface description of what my immediate situation is, but a more general principle that is more akin to what I'm really trying to determine. Let's look at the dinner example again. Is my maxim *really* "when I'm hungry I'll eat barbecued salmon for dinner"? Every time I'm hungry? Only at dinner time? It's far more likely that the action I am assessing is more like *eating*

something I like for a meal than eating that *exact* thing at that *exact* time. The maxim aims at the fundamental, underlying intention. Remember, deontology looks at *motivations*, so my maxim should be looking more to what my motivations are, not the precise action.

So how would this look, then, according to the FUL? If my maxim is

M: When I'm hungry, I'll eat something I like at mealtime.

Then the universalization of that will be

U: It is a law of human nature that everyone, when hungry, eats something they like at mealtime.

Does this force an impossibility? Well… no. Not a bit.

Let's look at the *relaxing on the couch* instance. Do I *really* intend that every time I want to relax, it must be in that *exact* location? Probably not. I like to relax in restaurants sometimes, too. And there's this place right by the Spokane River in Riverside State Park that I find one of the most perfect places in the world ever as a location for peaceful relaxation of rejuvenating solitude. So there are other places that I find quite lovely as relaxing spots. My maxim is better stated as

M: When I want to relax, I'll go someplace I find conducive to my need for relaxation.

So the universalization of this will be

U: It is a universal law of human nature that everyone always goes someplace they find conducive to their need for relaxation whenever they need to relax.

Again, we find no contradiction. So it turns out the *false negatives* problem only arises with too-specific, inaccurate maxims.

But notice this. If I'm inflexible—if I refuse, in a Sheldon Cooper-ish fashion to go *anywhere at all* but my specific desired location for relaxation (that's *my spot* on the couch)—*no matter what*—if I refuse to open up the possibility of other food choices or relaxation places, then it turns out my action ***is*** immoral. Having a fixation on the particularity of one's maxim might very well be morally problematic. Kant frowns upon Sheldon.

The important thing is to be sure we *don't tailor-make* our maxims to ensure they pass the FUL test. We don't want to make a maxim that passes—and then act on a totally different maxim. This is just paying lip-service to the Categorical Imperative, not acting from duty but from some other inclination, and wrapping it in a nice little blanket of hypocrisy.

False Positives

Suppose you need some money, and bad. Suppose further that you have a friend, Suzanne Smith, who has a chunk of change and a generous heart. You're going to have lunch with her the first Friday in June. You want to borrow from her, but you are pretty sure you can't pay her back, and you only want ever to borrow money this one time. Of course, she won't loan money if you tell her you won't pay her back, so you're going to … uh … *fib.* You don't see her all that often, so it's unlikely you'll ever find her available for any potential future loans. So you decide to test the following maxim:

M: On the first Friday in June this year, if I have lunch with her, I will make a false promise to Suzanne Smith to repay a loan (to get myself out of a financial jam)

Being the careful ethicist that you are, you decide to test your maxim by the FUL, universalizing it thus:

U: It is a universal law of human nature that everyone who has lunch with Suzanne Smith on the first Friday in June this year makes a false promise to repay a loan (to get out of a financial jam)

Is it impossible for there to be a world where both M and U are true? Well, everyone who has lunch with

AM I A GOOD PERSON? DEEP DOWN, DO I EVEN REALLY WANT TO BE A GOOD PERSON, OR DO I ONLY WANT TO SEEM LIKE A GOOD PERSON SO THAT PEOPLE (INCLUDING MYSELF) WILL APPROVE OF ME? IS THERE A DIFFERENCE? HOW DO I EVER ACTUALLY KNOW WHETHER I'M BULLSHITTING MYSELF, MORALLY SPEAKING?

(DAVID FOSTER WALLACE)

Suzanne on the first Friday of June could be a lot or very few people. They could all lie to her. She'd not know. She could loan everyone the money—she *does* have a generous heart, after all. And this is just about *this* June, not *every June for the rest of all time*. So no, it's not impossible. M and U don't force a contradiction.

Woo hoo! I can borrow money from Suzanne and never repay her! I'm off the hook!

But this seems pretty bad. Clearly, if it's immoral to make a lying promise to repay a debt to *anyone ever*, it would be immoral to make a lying promise to *just one person just one time*. Still, the LEM knows all, so this action is permissible. Thus we have the problem of ***false positives***, or things that seem clearly to be immoral but the LEM allows.

The ***Objection from Falls Positives*** looks much the same as that of false negatives:

FALSE POSITIVES

1. Deontology holds the Categorical Imperative, which is presented in the Formulation of Universal Law (FUL), as the supreme principle of morality.
2. A supreme principle of morality is a principle that is universal, unchanging, and without exceptions.
3. The FUL says that an action A is morally permissible if one can consistently will both a maxim to commit A and the universalization of that maxim.
4. I can consistently will both the maxim "On the first Friday in June, if I have lunch with her, I will make a false promise to Suzanne Smith to repay a loan (to get myself out of a financial jam)" and the universalization "It is a universal law of human nature that everyone who has lunch with Suzanne Smith on the first Friday in June makes a false promise to repay a loan (to get out of a financial jam)" (There is no contradiction).
5. But it is immoral to make a deceitful promise.
6. So there are some exceptions to what the FUL permits.

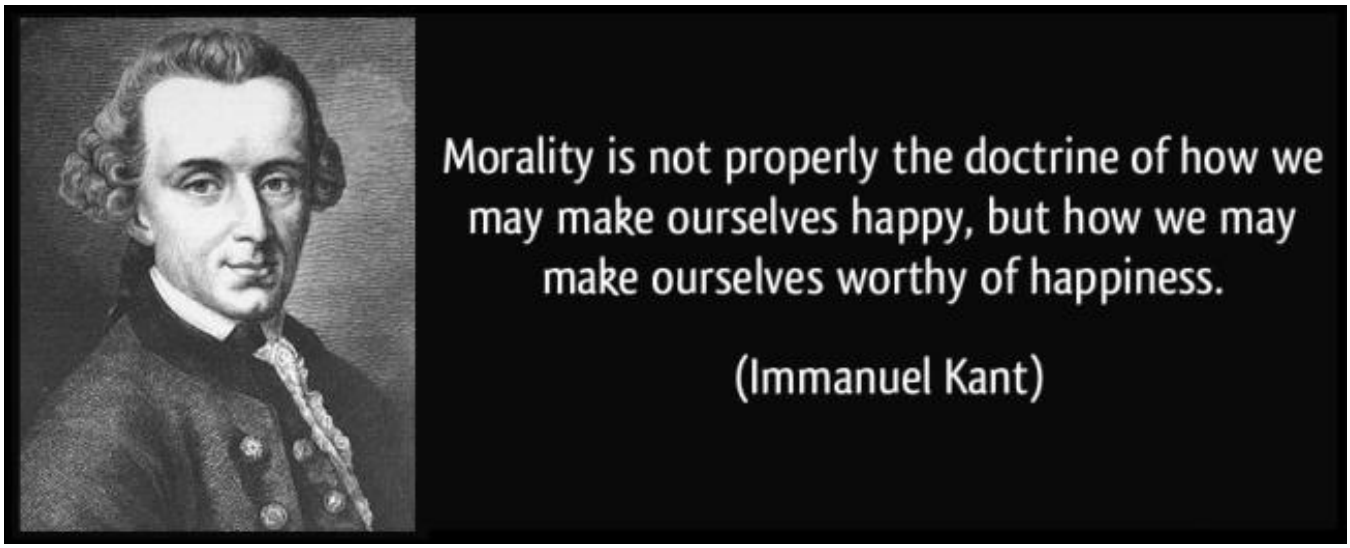
---

7. So the FUL cannot be the supreme principle of morality.

Responding to the False Positives Objection

Unsurprisingly, the response to this objection is similar to the response to the false negatives objection. The problem before was that the maxim was too specific. And we noted that we needed to be intellectually honest when making a maxim, that we don't make it unnecessarily specific. Here, we might take a leaf from the same book. What is the maxim I am *really* considering when I'm thinking about lying to Suzanne? Is it *really* that I will only make such a false promise at that precise moment? What if our lunch date moved to a dinner date? Or to a different day? What if we were to cancel the meal and simply get together for a coffee? What if the planned date were cancelled and we still happened to run into each other? Would I then suddenly not want to borrow money with a deceitful promise? Doubtful.

Kant would say that such a very carefully worded and utterly specific is not really the maxim we're looking at. Rather, it's just that we're again focused on the surface action and not the motivation itself. In fact,



this seems to be exactly that kind of tailor-made maxim we were warned against above. If we were honest with ourselves, we wouldn't try to cram something like this through the LEM-machine.

Kant wouldn't say that *if* the LEM found some *actual* maxim consistent with its universalization, even if the action seemed clearly immoral, that the action so considered would *still be immoral*. The LEM is the test of the supreme principle of morality, so if some odd person actually had such a specific and unlikely maxim, then that action would be permitted.

Sheldon gets off after all.

But this would be a tiny problem, not a huge catastrophe. Still, you might find it a bit worrisome to think that a theory that stands on pure logic can't catch all the little problems.

The Formulation of Humanity

The nagging takes us into the second formulation of the Categorical Imperative. If we think about our actions, we note that we always do something for some ultimate end, for some goal. And we remember that some of our goals or ends are *subjective*, like wanting certain consequences or meeting certain personal desires. And we further remember that an imperative that aims towards such subjective ends is a *hypothetical* imperative, whereas the supreme principle of morality is the Categorical Imperative. The justification of deceitfully borrowing money is subjective—not categorical.

How can we tease this out so that false positives never emerge again? Kant thinks about what sorts of things are ends-in-themselves, intrinsically valuable. He writes,

The Formulation of Autonomy

The FH doesn't say that we should treat people as we want to be treated. It doesn't rely on sympathy or personal preference. It says that we should respect humanity. So what is it about humanity that is intrinsically valuable? This gets unpacked in the final formulation of the Categorical Imperative. Sure, it's our rationality. But what makes rationality so darn special?

Context is helpful. The Formulation of Universal Law requires us to think of ourselves as legislating universal law: my maxim is at the same time considered a universal law. The Formulation of Humanity reminds us that all humans have equal dignity. Could it be that the latter is connected logically to the former? Sure thing, kemosabe. Kant reminds us that "autonomy is … the ground of the dignity of the human and of every rational creature."

What is this autonomy?

The word ***autonomy*** comes from the Latin *auto* (self) and *nomos* (law). That is, we are self-laws, laws unto ourselves, or moral legislators. Just like the FUL has me actually legislating morality based on pure logic, on *a priori* reason, so I can see that *every person* has this same power of reason, and that therefore every single person is a legislator of morality.

continued…

But suppose there were something *whose existence in itself* had an absolute worth, something that, as *end in itself*, could be a ground of determinate laws; then in it and only in it alone would lie the ground of a possible categorical imperative.

His reasoning is something like this:

1. If there are no objective ends, then there is nothing of absolute worth.
2. If there is no absolute worth (i.e., if all worth is conditional and contingent), then there is no Categorical Imperative.

3. So if there is a Categorical Imperative, then there must be objective ends.

This is a valid argument.* Further, Kant argues that there *are* objective ends:

> Now I say that the human being, and in general every rational being, *exists* as end in itself, *not merely as means* to the discretionary use of this or that will, but in all its actions, those directed toward itself as well as those directed toward other rational beings, it must always *at the same time* be considered as an *end*.

We are objective ends. We are of absolute worth. Because we are, there is a ground for the Categorical Imperative, and this ground gives us a second formulation, the ***Formulation of Humanity*** or FH:

> **FH:** *Act so that you treat humanity, whether in your own person or in that of another, always as an end and never as a means only.*

In short, *never* treat a person as only a mere means to some other end.

How am I treating Suzanne when I falsely promise to repay her loan? Not exactly as intrinsically valuable. In fact, I'm treating her as a piggy bank, a cash machine that is less valuable than my own financial worries. When I lie to somebody, I'm using that person to get

The Formulation of Autonomy,

continued.

When I make a moral judgment, I'm participating, taking my share, in the *universal legislation*, as a citizen in a "realm of ends." Thus the final formulation, which is a logical combination of the FUL and the FH, the ***Formulation of Autonomy*** or FA:

> **FA:** *Act only so that the will, through its maxims, could regard itself at the same time as universally lawgiving.*

Considering that lie to Suzanne, again, I can see that such a maxim is really given by my desires, and that it conflicts with the very moral principles I give—that all of us give—to ourselves independently of desire. It conflicts with our very rationality, it ignores (and tramples) the idea that Suzanne is also a lawgiver. Think about it. How can Suzanne make a moral judgment without all the facts? By lying to her, I'm preventing her from acting as a lawmaker; I'm preventing her from using her reason fully to determine the correct course of action. Thus, I am both denying her lawmaking powers which is just to say I'm devaluing her humanity, and with it, my own.

When I am acting like a universal lawgiver, I'm acting as if my every action matters. I'm saying that *doing this thing here* is the right thing *for everyone*. My judgements are saying that *this right here* is a good thing.

Heck, whether I want to admit it or not, even if I'm not acting circumspectly, my actions say that I think they're good. But when I do think carefully, I'm making sure that they're saying exactly what they should be saying, that they're rationally representing the good of all humanity.

WHEN YOU ALWAYS WANT THE NEXT THING, EVERY HUMAN BEING BECOMES A MEANS TO AN END.

(ECKHART TOLLE)

* It's a Hypothetical Syllogism with some transposition tossed in for good measure. See chapter 6.

THE ATTRACTIONS OF DEONTOLOGY

There are five characteristics that make a deontological ethics extremely attractive. First, it speaks to our intuition that motives matter. That is to say that deontology is not consequentialist.

Second, its appeal to pure reason speaks to our intuition—voiced by Rand—that logic and rationality matter, that reason should be the foundation of ethics. In this, it provides us with an absolute and inflexible obligation.

A third characteristic is that it doesn't appeal to pleasure. Its focus on reason means that it's not hedonistic, and this can be appealing when we think something is more important than good feelings.

And in contrast to hedonistic theories, deontology values respect and the dignity of every single human being. That's a fourth characteristic that draws us to deontology: each person is intrinsically valuable. Utilitarianism doesn't have that, for all its charms. One pleasure generator is as valuable as the next, with nothing giving it worth beyond its part in the calculus. Deontology gives all humans intrinsic value, as an end, not merely as a part of a sum.

And finally, a fifth characteristic that makes deontology appealing is its focus on our shared power of legislation: our autonomy.

something I want, something that the lie demonstrates I value more than the very humanity of the person to whom I lie.

In contrast, things like eating when I'm hungry—which sometimes includes barbecued salmon for dinner—or relaxing in an environment I find restful—which is sometimes on my own couch—are completely consistent with respecting the intrinsic value of other people.

The FH enables us to see the point that underscores the FUL: this is a principle aimed at respecting the dignity of human beings as rational beings. It's based on rationality. The absolute worth of persons is a value that doesn't reduce to some price or value that is relative to some other desired end. People are intrinsically valuable.

It might be that the FH is more intuitive than the FUL. But it's easier to get wrong, too. Suppose you know only the FH and want to see whether an action is correct. So you think, *I want to be sure I never use anyone as a means.* And you then find life suddenly very difficult. If you buy a coffee at an espresso stand, you're using the barista as a means to your Joe. If you take your car into the mechanic, you're using the mechanic. ARGH! Must I do everything myself? What kind of principle *is* this?

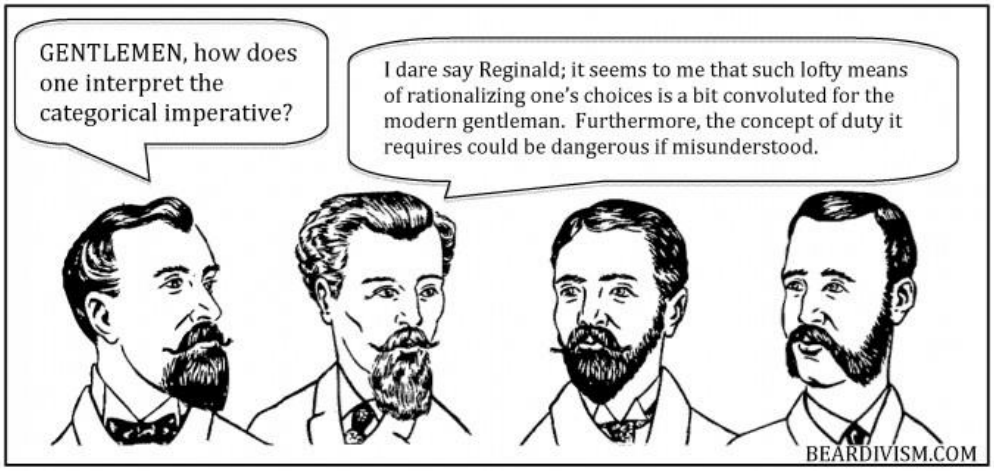
This error comes from jumping too quickly. The formulation doesn't say we should never use people as a means, but that we should never use people as *only* a means. Of *course* we use each other. Right now, you're using me as a means to knowledge or maybe to a grade. Is that immoral?

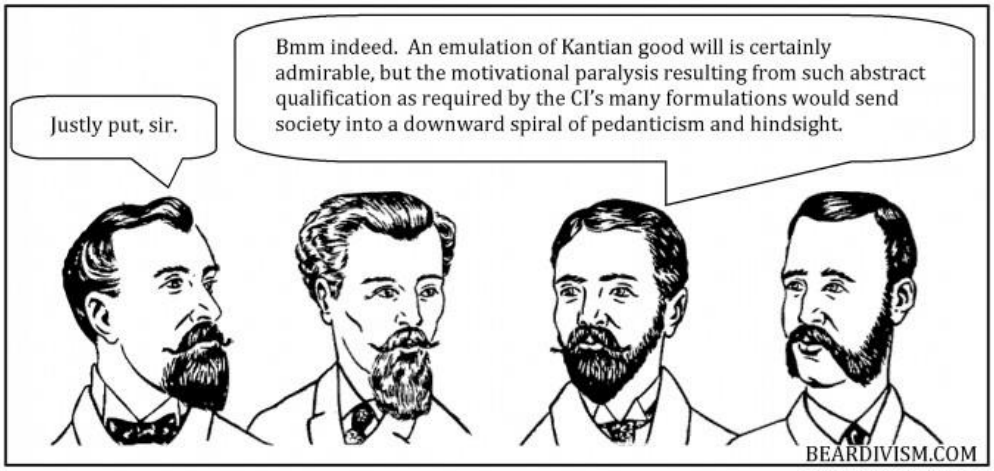
This principle sucks!

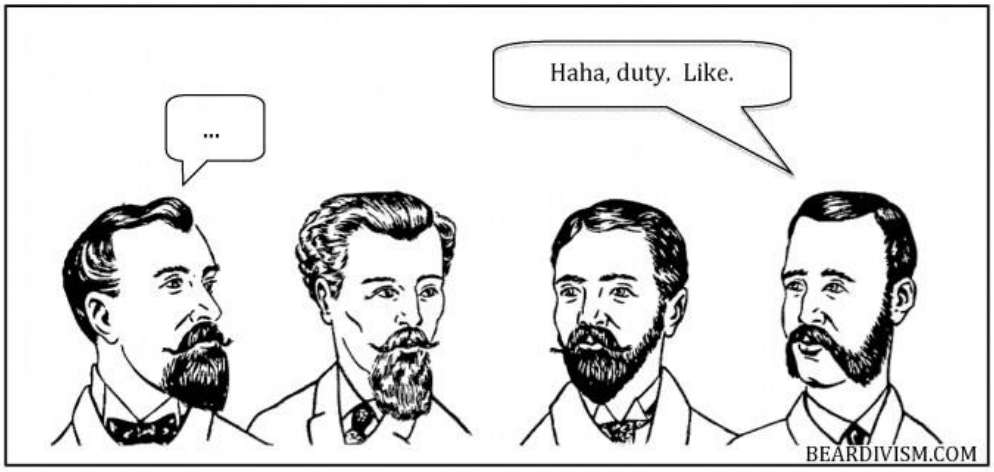
No, the FH says that we shouldn't see people as *nothing more* than a means to our ends. Of *course* we use each other. But our relationships should ever and always be *more* than just the using.

> Whatever is my right as a man is also the right of another; and it becomes my duty to guarantee as well as to possess.
>
> (Thomas Paine)

When you get your coffee, remember that the barista is a full human being, intrinsically valuable, and treat him as such. Respect his humanity, even if he mispronounces your name when calling you to the bar. Even when he atrociously misspells your name on the cup. When you drive down the interstate, remember that other drivers are full human beings, intrinsically valuable, and treat them as such. Acknowledge their humanity, even if they are driving agonizingly slow or they abruptly cut you off.







The FH says that we should *always* respect each human being as intrinsically valuable and dignified *even when* we're using that person's services. So we treat the mechanic as a well-rounded, complete human being, worthy of respect, even if she takes three more hours on your car than she said she would. Even if the repair costs more than the estimate. Treat the person who answers, when you call for IT help, with dignity and respect, even if their English isn't great or their accent is thick or they don't give you helpful information of do seem to give you the run-around. It's not about *their duty to you*, it's about yours to them.

It's not them; it's you.

TEAM PROJECT: COMPARING TWO APPROACHES

Two Supreme Principles

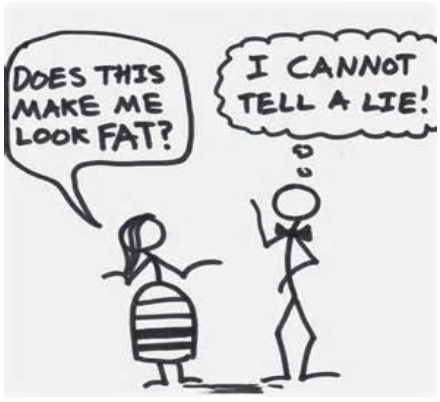
Return to that task your team chose for the first project in the last unit. As a team, reassess that action, but this time, measure it according to the Categorical Imperative.

Test it first by the FUL, making sure you carefully explain your team's maxim and the universalization of it. Then test it by the FH and the FA.

Do all three formulations agree? Does the deontological approach agree with the utilitarian approach you took before? Is something now morally blameworthy that was before praiseworthy? Is something now obligatory that was before blameworthy? How do the two theories compare?

Discuss and write down your team's findings, including which theory you as a team prefer (if any) and why. What do you find to be the strengths of each approach? What are the weaknesses of each approach? Do you think they can work together?

Your instructor will set the due date for this project. Write that date on the assignment, along with the names of all your participating team members. Turn in *one* paper for the whole team. Please write legibly.



A Third Problem: Rigorism

The very autonomy that we so value, this independence of reason that we cherish, might give us greater worries if we look more carefully. In fact, it gives us a large problem that might shake deontology to its core. It certainly caused enough worry to bring philosopher Christine Korsgaard into the conversation.

It begins with a thought experiment, and not one that's implausible, unlikely, or even ahistorical. Suppose you live in a part of the world that is ripped apart by factious violence. 1990s Rwanda. Or Nazi Germany. Or in the middle of Slobodan Milošević's massacre in former Yugoslavia. You are fortunate to be a member of the socially accepted ethnic group: you're Hutu or Aryan or Serbian.

But you're a good deontologist, so you treat everyone with human dignity. You respect the autonomy of others, and you find the genocide to be as atrocious as it truly is. You realize you must act, so you harbor refugees in your basement, perhaps behind a false wall. You're one of the good guys.

Now suppose that a leader from the Interahamwe or a Nazi SS officer or a member of one of the numerous Serbian militias comes to your door. He tells you that he is going door-to-door, looking for Tutsis or Jews or Croats, and asks you point-blank if you know where any are in the area.

If we apply *any* of the formulations, it comes out clear as can be that lying is wrong. We saw it with the lie about repayment, but what about a lie to save another's life? Kant says that truthfulness in statements that can't be avoided is the *formal duty* one has to *everyone*, however great the disadvantage that might arise from it, either to ourselves or to anyone else.* Ouch.

Let's cut to the heart of this thought experiment. As a good deontologist, you realize your maxim is

> M: I will lie to protect my innocent neighbor from a murderer

So the universalization of this would be

> U: It is a universal law of human nature that everyone always lies to protect innocent neighbors from a murderer.

Thinking as quickly as you can, you recognize that a world where U is true would be one where M would be impossible. In a world where everybody lies to murderers, no murderer would believe what anyone tells them, so there'd be no point in lying to them. It's against FUL. Strike one against lying.

You think again. Well, this guy is a horrible person, but he *is* a person. So you look to FH. You realize that if you lie to him, you're not treating him as intrinsically valuable. Your lying to him is to minimize his humanity, to use him only as a means to your protecting your harbored neighbors. To lie to him is to reject his rationality. Strike two against lying.

One last time, as you see he's getting fidgety. You remember that you should treat him as a member of the kingdom of ends. He's a lawgiver, too. You can't deny him his ability to legislate morality, and if you were to lie to him, you'd be deliberately giving

* He writes this in a 1797 essay called "On A Supposed Right to Lie because of Philanthropic Concerns." He wrote this essay as a response to Benjamin Constant's objection that "the moral principle stating that it is a duty to tell the truth would make any society impossible if that principle were taken singly and unconditionally."

him limited information, removing from him his legislative power. And strike three. You're out.

The problem we can call the problem of ***rigorism***. Kant's theory doesn't seem to allow for any wiggle during extraordinary circumstances. What options do we have? Are we really morally obligated to hand over the victims and thereby become participants in their murder?

Options?

It seems the only options Kant offers us are three. We *could* take other actions. Without lying to the fellow at the door, we might try to defend our neighbors. We might try to distract the murderer while the refugees sneak out the back way. We might try to dissuade him from his evil evil ways. We're smart, and it's true that in many—if not most—cases telling the truth is *avoidable* without outright lying. Kant doesn't tell us we have to *volunteer* information—just that we can't outright *lie*. So if the guy asks us if we know whether there are any Tutsis or Jews or Croats around, we could simply offer an "I do." That might not exactly impress the murderer at the door, but it might give you the option to change the subject. "Sure," you might say. "I heard of some who were in the marketplace yesterday. Just down the road." Thus you're truthful but not risking the lives of those in your basement.

If that seems ugly to you, we could go for option number two. Our squeamishness comes from our attachment to potential consequences. We are worried that if we do something, death will ensue. We could ask ourselves then whether lying is the only way to save our innocent neighbors. Probably not. So maybe we could—before the murderer actually shows up—set up an alternative plan. A different way to save them. Of course, we must also remember, if we're so attached to being consequentialists all of a sudden after a lifelong quest of being non-consequentialist deontologists—that it's quite possible that lying could make things even worse, not better.

But still, it seems like we should sometimes worry about the consequences of our actions. Even when we have a theory that says they cannot be the foundation of morality. There's the obvious solution behind curtain number three—tell the truth. It's just as plausible that this would be the consequence: he comes to your door, and he asks you whether there are any Tutsis, Jews, or Croats around. And you say,

Why yes, sir, there are! In fact, I happen to have three hidden in my basement right now, and if you wait a moment, I'll go fetch them for you!

His response could very well be one of shock, since he's *expecting* you to lie to him. And thus he interprets your brutal honesty as a lie. And he pulls himself up to his full height, and growls at you, "Don't mock me, you! This is serious!" And then he stomps away, leaving you with a warning and your hidden neighbors safe and sound.

This might seem too much of a gamble to take, but it isn't any less a gamble than any other option before you in such a horrifying situation. Are these our only options? Is it ever permissible just to lie and close that door?

American philosopher Christine Korsgaard is just as worried about this uncomfortable conclusion as you. She re-examines Kantian deontology and offers a two-level theory as a response, distinguishing between two kinds of circumstances: ***ideal*** (ordinary, everyday situations) and ***non-ideal*** (extraordinarily terrible situations).

Read her argument carefully, and prepare a critical question over it. Do you believe her reworking of the theory adequately answers this worry?

Lying to ourselves is more deeply ingrained than lying to others.
(Fyodor Dostoyevsky)

THE RIGHT TO LIE: KANT ON DEALING WITH EVIL

Christine Korsgaard[*†]

One of the great difficulties with Kant's moral philosophy is that it seems to imply that our moral obligations leave us powerless in the face of evil. Kant's theory sets a high ideal of conduct and tells us to live up to that ideal regardless of what other persons are doing. The results may be very bad. But Kant says that the law "remains in full force, because it commands categorically." (G 438-439/57) The most well-known example of this "rigorism", as it is sometimes called, concerns Kant's views on our duty to tell the truth.

[*] [All footnotes, unless specifically specified, are Korsgaard's.]

Where I have cited or referred to any of Kant's works more than once in this paper I have inserted the reference into the text. The following abbreviations are used:

G *Foundations of the Metaphysics of Morals*. (1785) The first page number is that of the Prussian Academy Edition Volume IV; the second is that of the translation by Lewis White Beck. Indianapolis: Bobbs-Merrill Library of Liberal Arts, 1959.

C2 *Critique of Practical Reason*. (1788) Prussian Academy Volume V; Lewis White Beck's translation. Indianapolis: Bobbs-Merrill Library of Liberal Arts, 1956.

MMV *The Metaphysical Principles of Virtue*. (1797) Prussian Academy Volume VI; James Ellington's translation in *Immanuel Kant: Ethical Philosophy*. Indianapolis: Hackett, 1983.

MMJ *The Metaphysical Elements of Justice*. (1797) Prussian Academy Volume VI; John Ladd's translation. Indianapolis: Bobbs-Merrill Library of Liberal Arts, 1965.

PP *Perpetual Peace*. (1795) Prussian Academy Volume VIII, translation by Lewis White Beck in *On History*, edited by Lewis White Beck. Indianapolis: Bobbs-Merrill Library of Liberal Arts, 1963.

SRL "On a Supposed Right to Lie from Altruistic Motives" (1797) Prussian Academy Volume VIII; translation by Lewis White Beck in *Immanuel Kant: Critique of Practical Reason and Other Writings in Moral Philosophy*. Chicago: University of Chicago Press, 1949; rpt: New York: Garland Publishing Company, 1976.

LE *Lectures on Ethics*.(1775-1780) edited by Paul Menzer from the notes of Theodor Friedrich Brauer, using the notes of Gottlieb Kutzner and Chr. Mrongovius; translated by Louis Infield. London: Methuen & Co., Ltd., 1930; rpt: New York, Harper Torchbooks, 1963; current rpt: Indianapolis, Hackett Press.

[†] This paper was delivered as the Randall Harris Lecture at Harvard in October, 1985. Versions of the paper have been presented at the University of Illinois at Urbana-Champaign, the University of Wisconsin at Milwaukee, the University of Michigan, and to the Seminar on Contemporary Social and Political Theory at Chicago. I owe a great deal to the discussions on these occasions. I want to thank the following people for their comments: Margaret Atherton, Charles Chastain, David Copp, Stephen Darwall, Michael Davis, Gerald Dworkin, Alan Gewirth, David Greenstone, John Koethe, Richard Kraut, Richard Strier, and Manley Thompson. And I owe special thanks to Peter Hylton and Andrews Reath for extensive and useful comments on the early written versions of the paper.

NOTES

In two passages in his ethical writings, Kant seems to endorse the following pair of claims about this duty: First, one must never under any circumstances or for any purpose tell a lie. Second, if one does tell a lie one is responsible for all of the consequences that ensue, even if they were completely unforeseeable.

One of the two passages occurs in the *Metaphysical Principles of Virtue.* There Kant classifies lying as a violation of a perfect duty to oneself. In one of the casuistical* questions, a servant, under instructions, tells a visitor the lie that his master is not at home. His master, meanwhile, sneaks off and commits a crime, which would have been prevented by the watchman sent to arrest him. Kant says:

> Upon whom ... does the blame fall? To be sure, also upon the servant, who here violated a duty to himself by lying, the consequence of which will now be imputed to him by his own conscience. (MMV 431/93)

The other passage is the infamous one about the murderer at the door from the essay, "On A Supposed Right to Lie From Altruistic Motives." Here Kant's claims are more extreme, for he says that the liar may be held legally as well as ethically responsible for the consequences, and the series of coincidences he imagines is even more fantastic:

> After you have honestly answered the murderer's question as to whether his intended victim is at home, it may be that he has slipped out so that he does not come in the way of the murderer, and thus that the murder may not be committed. But if you had lied and said he was not at home when he had really gone out without your knowing it, and if the murderer had then met him as he went away and murdered him, you might justly be accused as the cause of his death. For if you had told the truth as far as you knew it, perhaps the murderer might have been apprehended by the neighbors while he searched the house and thus the deed might have been prevented. (SRL 427/348)

Kant's readers differ about whether Kant's moral philosophy commits him to the claims he makes in these passages. Unsympathetic readers are inclined to take them as evidence of the horrifying conclusions to which Kant was led by his notion that the necessity in duty is rational necessity —as if Kant were clinging to a logical point in the teeth of moral decency. Such readers take these conclusions as a defeat for

* 'Casuistical' simply means 'case based,' referring to a method of reasoning or evaluating. Here Korsgaard is evaluating and modifying Kant's theory by means of case evaluation. [Kurle note]

Kant's ethics, or for ethical rationalism generally; or they take Kant to have confused principles which are merely general in their application and *prima facie* in their truth with absolute and universal laws. Sympathetic readers are likely to argue that Kant here mistook the implications of his own theory, and to try to show that, by careful construction and accurate testing of the maxim on which this liar acts, Kant's conclusions can be blocked by his own procedures.

NOTES

Sympathetic and unsympathetic readers alike have focused their attention on the implications of the first formulation of the categorical imperative, the Formula of Universal Law. The *Foundations of the Metaphysics of Morals* contains two other sets of terms in which the categorical imperative is formulated: the treatment of humanity as an end in itself, and autonomy, or legislative membership in a Kingdom of Ends. My treatment of the issue falls into three parts. First, I want to argue that Kant's defenders are right in thinking that, when the case is treated under the Formula of Universal Law, this particular lie can be shown to be permissible. Second, I want to argue that when the case is treated from the perspective provided by the Formulas of Humanity and the Kingdom of Ends, it becomes clear why Kant *is* committed to the view that lying is wrong in every case. But from this perspective we see that Kant's rigorism about lying is not the result of a misplaced love of consistency or legalistic thinking. Instead, it comes from an attractive ideal of human relations which is the basis of his ethical system. If Kant is wrong in his conclusion about lying to the murderer at the door, it is for the interesting and important reason that morality itself sometimes allows or even requires us to do something that from an ideal perspective is wrong. The case does not impugn Kant's ethics as an *ideal* system. Instead, it shows that we need special principles for dealing with evil. My third aim is to discuss the structure that an ethical system must have in order to accommodate such special principles.

Universal Law

The Formula of Universal Law tells us never to act on a maxim that we could not at the same time will to be a universal law. A maxim which cannot even be conceived as a universal law without contradiction is in violation of a strict and perfect duty, one which assigns us a particular action or omission. A maxim which cannot be willed as universal law without contradicting the will is in violation of a broad and imperfect duty, one which assigns us an end, but does not tell us what or how much we should do towards it. Maxims of lying are violations of perfect duty, and so are supposed to be the kind that cannot be conceived without contradiction when universalized.

The sense in which the universalization of an immoral maxim is supposed to "contradict" itself is a matter of controversy. On my

NOTES

reading, which I will not defend here,* the contradiction in question is a "practical" one: the universalized maxim contradicts itself when the efficacy of the action as a method of achieving its purpose would be undermined by its universal practice. So, to use Kant's example, the point against false promising as a method of getting ready cash is that if everyone attempted to use false promising as a method of getting ready cash, false promising would no longer *work* as a method of getting ready cash, since, as Kant says, "no one would believe what was promised to him but would only laugh at any such assertion as vain pretense." (G 422/40)

Thus the test question will be: could this action be the universal method of achieving this purpose? Now when we consider lying in general, it looks as if it could not be the universal method of doing anything. For lies are usually efficacious in achieving their purposes because they deceive, but if they were universally practiced they would not deceive. We believe what is said to us in a given context because most of the time people in that context say what they really think or intend. In contexts in which people usually say false things — e.g., when telling stories that are jokes — we are not deceived. If a story that is a joke and is false counts as a lie, we can say that a lie in this case is not wrong, because the universal practice of lying in the context of jokes does not interfere with the *purpose* of jokes, which is to amuse and does not depend on deception. But in most cases lying falls squarely into the category of the sort of action Kant considers wrong: actions whose efficacy depends upon the fact that most people do not engage in them, and which therefore can only be performed by someone who makes an exception of himself. (G 424/42)

When we try to apply this test to the case of the murderer at the door, however, we run into a difficulty. The difficulty derives from the fact that there is probably already deception in the case. If murderers

* I defend it in "Kant's Formula of Universal Law", forthcoming in *Pacific Philosophical Quarterly*.

standardly came to the door and said: "I wish to murder your friend — is he here in your house?" then perhaps the universal practice of lying in order to keep a murderer from his victim would not work. If everyone lied in these circumstances the murderer would be aware of that fact and would not be deceived by your answer. But the murderer is not likely to do this, or, in any event, this is not how I shall imagine the case. A murderer who expects to conduct his business by asking questions must suppose that you do not know who he is and what he has in mind.* If these are the circumstances, and we try to ascertain whether there could be a universal practice of lying in these circumstances, the answer appears to be yes. The lie will be efficacious even if universally practiced. But the reason it will be efficacious is rather odd: it is because the murderer supposes you do not know what circumstances you are in — that is, that you do not know you are addressing a murderer — and so does not conclude from the fact that people in those circumstances always lie that *you* will lie.

"Are you sure everything on your resume is accurate?"

The same point can be made readily using Kant's publicity criterion. (PP 381-383/129-131) Can we announce in advance our intention of lying to murderers without, as Kant says, vitiating our own purposes by publishing our maxims? (PP 383/131) Again the answer is yes. It does not matter if you say publicly that you will lie in such a situation,

NOTES

* I am relying on an assumption here, which is that when people ask us questions they give us some account of themselves and of the context in which the questions are asked. Or, if they don't, it is because they are relying on a context that is assumed. If someone comes to your door looking for someone, you assume that there's a family emergency or some such thing. I am prepared to count such reliance as deception if the questioner knows about it and uses it, thinking that we would refuse to answer his questions if we knew the real context to be otherwise. Sometimes people ask me, "Suppose the murderer just asks whether his friend is in your house, without saying anything about why he wants to know?" I think that, in our culture anyway, people do not *just ask* questions of each other about anything except the time of day and directions for getting places. After all, the reason why refusal to answer is an unsatisfactory way of dealing with this case is that it will almost inevitably give rise to suspicion of the truth, and this is because people normally answer such questions. Perhaps if we did live in a culture in which people regularly *just asked* questions in the way suggested, refusal to answer would be commonplace and would not give rise to suspicion; it would not even be considered odd or rude. Otherwise there would be no way to maintain privacy.

NOTES

for the murderer supposes that you do not know you are in that situation.* These reflections might lead us to believe, then, that Kant was wrong in thinking that it is never all right to lie. It is permissible to lie to deceivers in order to counteract the intended results of their deceptions, for the maxim of lying to a deceiver is universalizable. The deceiver has, so to speak, placed himself in a morally unprotected position by his own deception. He has created a situation which universalization cannot reach.

Humanity

When we apply the Formula of Humanity, however, the argument against lying that results applies to any lie whatever. The formula runs:

Act so that you treat humanity, whether in your own person or in that of another, always as an end and never as a means only. (G 429/47)

In order to use this formula for casuistical purposes, we need to specify what counts as treating humanity as an end. "Humanity" is used by Kant specifically to refer to the capacity to determine ends through rational choice. (G 437/56; MMV 392/50) Imperfect duties arise from the obligation to make the exercise, preservation, and development of this capacity itself an end. The perfect duties — that is, the duties of justice, and, in the realm of ethics, the duties of respect — arise from the obligation to make each human being's capacity for autonomous choice the condition of the value of every other end.

In his treatment of the lying promise case under the Formula of Humanity, Kant makes the following comments:

For he whom I want to use for my own purposes by means of such a promise cannot possibly assent to my mode of acting against him and cannot contain the end of this action in himself. ... he who transgresses the rights of men intends to make use of the persons of others merely as means, without considering that as rational beings, they must always be esteemed at the same time as ends, i.e. only as beings who must be able to contain in themselves the end of the very same action. (G 429-430/48)

In these passages, Kant uses two expressions that are the key to understanding the derivation of perfect duties to others from the Formula of Humanity. One is that the other person “cannot possibly

* In fact, it will now be the case that if the murderer supposes that you suspect him, he is not going to ask you, knowing that you will answer so as to deceive him. Since we must avoid the silly problem about the murderer being able to deduce the truth from his knowledge that you will speak falsely, what you announce is that you will say whatever is necessary in order to conceal the truth. There is no reason to suppose that you will be mechanical about this. You are not going to be a reliable source of information. The murderer will therefore seek some other way to locate his victim.

On the other hand, suppose that the murderer does, contrary to my supposition, announce his real intentions. Then the arguments that I have given do not apply. In this case, I believe, your only recourse is refusal to answer (whether or not the victim is in your house, or you know his whereabouts). If an answer is extorted from you by force you may lie, according to the argument I will give later in the paper.

assent to my mode of acting toward him" and the second is that the other person cannot "contain the end of this action in himself." These phrases provide us with a test for perfect duties to others: an action is contrary to perfect duty if it is not possible for the other to assent to it or to hold its end.

It is important to see that these phrases do not mean simply that the other person *does not* or *would not* assent to the transaction or that she does not happen to have the same end I do, but strictly that she *cannot* do so: that something makes it impossible. If what we cannot assent to means merely what we are likely to be annoyed by, the test will be subjective and the claim that the person does not assent to being used as a means will sometimes be false. The object you steal from me may be the gift I intended for you, and we may both have been motivated by the desire that you should have it. And I may care about you too much or too little to be annoyed by the theft. For all that this must be a clear case of your using me as a mere means.*

So it must not be merely that your victim will not like the way that you propose to act, that this is psychologically unlikely, but that something makes it impossible for her to assent to it. Similarly, it must be argued that something makes it impossible for her to hold the end of the very same action. Kant never spells out why it is impossible, but it is not difficult to see what he has in mind.

People cannot *assent* to a way of acting when they are given no chance to do so. The most obvious instance of this is when coercion is used. But it is also true of deception: the victim of the false promise cannot assent to it because he doesn't know it is what he is being offered. But even when the victim of such conduct does happen to know what is going on there is a sense in which he cannot assent to it. Suppose, for example, that you come to me and ask to borrow some money, falsely promising to pay it back next week, and suppose that by some chance I know perfectly well that your promise is a lie. Suppose also that I have the same end you do, in the sense that I want you to have the money, so that I turn the money over to you anyway. Now here I have the same end that you do, and I tolerate your attempts to deceive me to the extent that they do not prevent my giving you the money. Even in this case I cannot really assent to the transaction *you* propose. We can imagine the case in a number of different ways. If I call your bluff openly and say "never mind that nonsense, just take this money" then what I am doing is not accepting a false promise, but giving you a handout, and scorning your promise. The nature of the transaction is

NOTES

* Kant himself takes notice of this sort of problem in a footnote to this passage in which he criticizes Golden-Rule type principles for, among other things, the sort of subjectivity in question: such principles cannot establish the duty of beneficence, for instance, because "many a man would gladly consent that others should not benefit him, provided only that he might be excused from showing benevolence to them." (G 430n/48n)

NOTES

changed: now it is not a promise but a handout. If I don't call you on it, but keep my own counsel, it is still the same. I am not accepting a false promise. In this case what I am doing is *pretending* to accept your false promise. But there is all the difference in the world between actually doing something and pretending to do it. In neither of these cases can I be described as accepting a false promise, for in both cases I fix it so that it is something else that is happening. My knowledge of what is going on makes it *impossible* for me to accept the deceitful promise in the ordinary way.

The question whether another can assent to your way of acting can serve as a criterion for judging whether you are treating her as a mere means. We will say that knowledge of what is going on and some power over the proceedings are the conditions of possible assent; without these, the concept of assent does not apply. This gives us another way to formulate the test for treating someone as a mere means: Suppose it is the case that if the other person knows what you are trying to do and has the power to stop you, then what you are trying to do cannot be what is really happening. If this is the case, the action is one that by its very nature is impossible for the other to assent to. You cannot wrest from me what I freely give to you; and if I have the power to stop you from wresting something from me and do not use it, I am in a sense freely giving it to you. This is of course not intended as a legal point: the point is that any action which depends for its nature and efficacy on the other's ignorance or powerlessness fails this test. Lying clearly falls into this category of action: it only deceives when the other does not know that it is a lie.*

A similar analysis can be given of the possibility of holding the end of the very same action. In cases of violation of perfect duty, lying included, the other person is unable to hold the end of the very same action because the way that you act prevents her from *choosing* whether to contribute to the realization of that end or not. Again, this is obviously true when someone is forced to contribute to an end, but it is also true in cases of deception. If you give a lying promise to get some money, the other person is invited to think that the end she is contributing to is your temporary possession of the money: in fact, it is your permanent possession of it. It doesn't matter whether that would

* Sometimes it is objected that someone could assent to being lied to in advance of the actual occasion of the lie, and that in such a case the deception might still succeed. One can therefore agree to be deceived. I think it depends what circumstances are envisioned. I can certainly agree to remain uninformed about something, but this is not the same as agreeing to be deceived. I could say to a doctor: “don't tell me if I am fatally ill, even if I ask” for instance. But if I then do ask the doctor whether I am fatally ill, I cannot be certain whether she will answer me truly. Perhaps what's being envisioned is that I simply agree to be lied to, but not about anything in particular. Will I then trust the person with whom I've made this odd agreement?

be all right with her if she knew about it. What matters is that she never gets a chance to choose the end, not knowing that it is to be the consequence of her action.*

According to the Formula of Humanity, coercion and deception are the most fundamental forms of wrongdoing to others — the roots of all evil. Coercion and deception violate the conditions of possible assent, and all actions which depend for their nature and efficacy on their coercive or deceptive character are ones that others cannot assent to. Coercion and deception also make it impossible for others to choose to contribute to our ends. This in turn makes it impossible, according to Kant's value theory, for the ends of such actions to be good. For on Kant's view "what we call good must be, in the judgement of every reasonable man, an object of the faculty of desire." (C2 60/62-63) If your end is one that others cannot choose — not because of what they want, but because they are not in a position to choose — it cannot, as the end of that action, be good. This means that in any cooperative project — whenever you need the decisions and actions of others in order to bring about your end — everyone who is to contribute must be in a position to *choose* to contribute to the end.

The sense in which a good end is an object for everyone is that a good end is in effect one that everyone, in principle, and especially everyone who contributes to it, gets to cast a vote on. This voting, or legislation, is the prerogative of rational beings; and the ideal of a world in which this prerogative is realized is the Kingdom of Ends.

The Kingdom of Ends

The Kingdom of Ends is represented by the kingdom of nature; we determine moral laws by considering their viability as natural laws. On Kant's view, the will is a kind of causality. (G 446/64) A person, an end in itself, is a free cause, which is to say a first cause. By contrast a thing, a means, is a merely mediate cause, a link in the chain. A first cause is, obviously, the initiator of a causal chain, hence a real determiner of what will happen. The idea of deciding for yourself whether you will contribute to a given end can be represented as a decision whether to initiate that causal chain which constitutes your contribution. Any action which prevents or diverts you from making this initiating decision is one that treats you as a mediate rather than a first cause; hence as a mere

NOTES

* A similar conclusion about the way in which the Formula of Humanity makes coercion and deception wrong is reached by Onora O'Neill in "Between Consenting Adults," *Philosophy and Public Affairs* Volume 14, No. 3 (Summer, 1985), pp. 252-277.

NOTES

means, a thing, a tool. Coercion and deception both do this. And deception treats you as a mediate cause in a specific way: it treats your reason as a mediate cause. The false promiser thinks: if I tell her I will pay her back next week, then she will choose to give me the money. Your reason is worked, like a machine: the deceiver tries to determine what levers to pull to get the desired results from you. Physical coercion treats someone's person as a tool; lying treats someone's *reason* as a tool. This is why Kant finds it so horrifying; it is a direct violation of autonomy.

We may say that a tool has two essential characteristics: It is there to be used, and it does not control itself: its nature is to be directed by something else. To treat someone as a mere means is to treat her as if these things were true of her. Kant's treatment of our duties to others in the *Metaphysical Principles of Virtue* is sensitive to *both* characteristics. We are not only forbidden to use another as mere means to our private purposes. We are also forbidden to take attitudes towards her which involve regarding her as not in control of herself, which is to say, as not using her reason. This latter is the basis of the duties of respect. Respect is violated by the vices of calumny and mockery (MMV 466-468/131-133): we owe to others not only a practical generosity toward their plans and projects — a duty of aid — but also a generosity of attitude toward their thoughts and motives. To treat another with respect is to treat him as if he were using his reason and as far as possible as if he were using it well. Even in a case where someone evidently *is* wrong or mistaken, we ought to suppose he must have what he takes to be good reasons for what he believes or what he does. This is not because, as a matter of fact, he probably does have good reasons.

Rather, this attitude is something that we *owe* to him, something that is his right. And he cannot forfeit it. Kant is explicit about this:

> Hereupon is founded a duty to respect man even in the logical use of his reason: not to censure someone's errors under the name of absurdity, inept judgement, and the like, but rather to suppose that in such an inept judgment there must be something true, and to seek it out. ... Thus it is also with the reproach of vice, which must never burst out in complete contempt or deny the wrongdoer all moral worth, because on that hypothesis he could never be improved either — and this latter is incompatible with the idea of man, who as such (as a moral being) can never lose all predisposition to good. (MMV 463-464/l28-l29)

To treat others as ends in themselves is always to address and deal with them as rational beings. Every rational being gets to reason out, for herself, what she is to think or to choose or to do. So if you need someone's contribution to your end, you must put the facts before her

and ask for her contribution. If you think she is doing something wrong, you may try to convince her by argument but you may not resort to tricks or force. The Kingdom of Ends is a democratic ideal, and poor judgment does not disqualify anyone for citizenship. In the *Critique of Pure Reason*, Kant says:

> Reason depends on this freedom for its very existence. For reason has no dictatorial authority; its verdict is always simply the agreement of free citizens, of whom each one must be permitted to express, without let or hindrance, his objections or even his veto.*

This means that there cannot be a good reason for taking a decision out of someone else's hands. It is a rational being's prerogative, as a first cause, to have a share in determining the destiny of things.

This shows us in another way why lying is for Kant a paradigm case of treating someone as a mere means. Any attempt to control the actions and reactions of another by any means except an appeal to reason treats her as a mere means, because it attempts to reduce her to a mediate cause. This includes much more than the utterance of falsehoods. In the *Lectures on Ethics*, Kant says "whatever militates against frankness lowers the dignity of man." (LE 231)† It is an everyday temptation, even (or perhaps especially) in our dealings with those close to us, to withhold something, or to tidy up an anecdote, or to embellish a story, or even just to place a certain emphasis, in order to be sure of getting the reaction we want.‡ Kant holds the Socratic view that any sort of persuasion that is aimed at distracting its listener's attention from either the reasons that she ought to use or the reasons the speaker thinks she will use is wrong.§

In light of this account it is possible to explain why Kant says what he does about the liar's responsibility. In a Kantian theory our responsibility has definite boundaries: each person as a first cause

NOTES

* *Immanuel Kant's Critique of Pure Reason*, translated by Norman Kemp Smith. (New York: St. Martin's Press, 1965) A738-739/B766-767, p. 593.

† It is perhaps also relevant that in Kant's discussion of perfect moral friendship the emphasis is not on good will towards one another but on complete confidence and openness. See MMV 471-472/139-139.

‡ Some evidence that Kant is concerned with this sort of thing may be found in the fact that he identifies two meanings of the word "prudence" (Klugheit); "The former sense means the skill of a man in having an influence on others so as to use them for his own purposes. The latter is the ability to unite all these purposes to his own lasting advantage." (G 416n/33n) A similar remark is found in *Anthropology from a Pragmatic Point of View*. (1798) See the translation by Mary J. Gregor (The Hague: Martinus Nijhoff, 1974) p. 183. Prussian Academy Edition Volume VII, p.322.

§ I call this view Socratic because of Socrates's concern with the differences between reason and persuasion and, in particular, because in the *Apology,* he makes a case for the categorical duty of straightforwardness. Socrates and Plato are also concerned with a troublesome feature of this moral view that Kant neglects. An argument must come packaged in some sort of presentation, and one may well object that it is impossible to make a straightforward presentation of a case to someone who is close to or admires you, without emphasis, without style, without taking some sort of advantage of whatever it is about you that has your listener's attention in the first place. So how can we avoid the non-rational influence of others? I take it that most obviously in the *Symposium*, but also in other dialogues concerned with the relation of love and teaching such as the *Phaedrus*, Plato is at work on the question whether you can use your sex appeal to draw another's attention to the reasons he has for believing or doing things, rather than as a distraction that aids your case illicitly.

NOTES

exerts some influence on what happens, and it is your part that is up to you. If you make a straightforward appeal to the reason of another person, your responsibility ends there and the other's responsibility begins.

But the liar tries to take the consequences out of the hands of others; he, and not they, will determine what form their contribution to destiny will take. By refusing to share with others the determination of events, the liar takes the world into his own hands, and makes the events his own. The results, good or bad, are imputable to him, at least in his own conscience. It does not follow from *this*, of course, that this is a risk one will never want to take.

Humanity and Universal Law

If the foregoing casuistical analyses are correct, then applying the Formula of Universal Law and applying the Formula of Humanity lead to rather different answers in the case of lying to the murderer at the door. The former seems to say that this lie is permissible, but the latter says that coercion and deception are the most fundamental forms of wrongdoing. In a Kingdom of Ends coercive and deceptive methods can never be used.

This result impugns Kant's belief that the formulas are equivalent. But it is not necessary to conclude that the formulas flatly say different things, and are unrelated except for a wide range of coincidence in their results. For one thing, lying to the murderer at the door was not shown to be permissible in a straightforward manner: the maxim did not so much pass as evade universalization. For another, the two formulas can be shown to be expressions of the same basic theory of justification. Suppose that your maxim is in violation of the Formula of Universal Law. You are making an exception of yourself, doing something that everyone in your circumstances could not do. What this means is that you are treating the reason *you* have for the action as if it were stronger, had more justifying force, than anyone else's exactly similar reason. You are then acting as if the fact that it was in particular *your* reason, and not just the reason of a human being, gave it special weight and force. This is an obvious violation of the idea that it is your humanity — your power of rational choice — which is the condition of all value and so which gives your needs and desires the justifying force of *reasons.* Thus, any violation of the Formula of Universal Law is also a violation of the Formula of Humanity. This argument, of course, only goes in one direction: it does not show that the two formulas are equivalent. The Formula of Humanity is more strict than the Formula of Universal Law — but both are expressions of the same basic theory of value: that your rational nature is the source of justifying power of your reasons, and so of the goodness of your ends.

And although the Formula of Humanity gives us reason to think that all lies are wrong, we can still give an account in the terms it provides of what vindicates lying to a liar. The liar tries to use your reason as a means - your honesty as a tool. You do not have to passively submit to being used as a means. In the *Lectures on Ethics*, this is the line that Kant in fact takes. He says:

> if we were to be at all times punctiliously truthful we might often become victims of the wickedness of others who were ready to abuse our truthfulness. If all men were well-intentioned it would not only be a duty not to lie, but no one would do so because there would be no point in it. But as men are malicious, it cannot be denied that to be punctiliously truthful is often dangerous... if I cannot save myself by maintaining silence, then my lie is a weapon of defense. (LE 228)

The common thought that lying to a liar is a form of self-defense, that you can resist lies with lies as you can resist force with force, is according to this analysis correct.* This should not be surprising, for we have seen that deception and coercion are parallel. Lying and the use of force are attempts to undercut the two conditions of possible assent to actions and of autonomous choice of ends, namely, knowledge and power. So, although the Formula of Universal Law and the Formula of Humanity give us different results, this does not show that they simply express different moral outlooks. The relation between them is more complex than that.

Two Casuistical Problems

Before I discuss this relation, however, I must take up two casuistical problems arising from the view I have presented so far. First, I have argued that we *may* lie to the murderer at the door. But most people think something stronger, that we ought to lie to the murderer — that we will have done something wrong if we do not. Second, I have argued that it is permissible to lie to a deceiver in order to counter the deception. But what if someone lies to you for a good end, and, as it happens, you know about it?

NOTES

* Of course you may also resist force with lies, if resisting it with force is not an option for you. This gives rise to a question about whether these options are on a footing with each other. In many cases, lying will be the better option. This is because when you use coercion you risk doing injury to the person you coerce. Injuring people unnecessarily is wrong, a wrong that should be distinguished from the use of coercion. When you lie you do not risk doing this extra wrong. But Kant thinks that lying is in itself worse than coercion, because of the peculiarly direct way in which it violates autonomy. So it should follow that if you can deal with the murderer by coercion, this is a *better* option than lying. Others seem to share this intuition. Cardinal John Henry Newman, responding to Samuel Johnson's claim that he would lie to a murderer who asked which way his victim had gone, suggests that the appropriate thing to do is "to knock the man down, and to call out for the police." (*Apologia Pro Vita Sua: Being a History of His Religious Opinions*. (London: Longmans, Green & Co., 1880) p. 361. I am quoting from Sissela Bok, *Lying*. (New York: Vintage Books, 1979) p 42.) If you can do it without seriously hurting the murderer, it is, so to speak, cleaner just to kick him off the front porch than to lie. This treats the *murderer himself* more like a human being than lying to him does.

NOTES

The fact that the murderer's *end* is evil has played no direct role in the arguments I have given so far. We have a right to resist liars and those who try to use force because of their methods, not because of their purposes. In one respect this is a virtue of my argument. It does not license us to lie to or to use violence against persons *just* because we think their purposes are bad. But it looks as if it may license us to lie to liars whose purposes are good. Here is a case:* suppose someone comes to your door and pretends to be taking a survey of some sort. In fact, this person is a philanthropist who wants to give his money to people who meet certain criteria, and this is his way of discovering appropriate objects for his beneficence. As it happens, you know what is up. By lying, you could get some money, although you do not in fact meet his criteria. The argument that I derived from the Formula of Universal Law about lying to the murderer applies here. Universalizing the lie to the philanthropist will not destroy its efficacy. Even if it is a universal law that everyone will lie in these circumstances, the philanthropist thinks you do not know you are in these circumstances. By my argument, it is permissible to lie in this case. The philanthropist, like the murderer, has placed himself in a morally unprotected position by his own deception. Start with the first casuistical problem. There are two reasons to lie to the murderer at the door. First, we have a duty of mutual aid. This is an imperfect duty of virtue, since the law does not say exactly what or how much we must do along these lines. This duty gives us *a* reason to tell the lie. Whether it makes the lie imperative depends on how one understands the duty of mutual aid, on how one understands the "wideness" of imperfect duties.† It may be that on such an urgent occasion, the lie is imperative. Notice that if the lie were impermissible, this duty would have no force.

Imperfect duties are always secondary to perfect ones. But if the lie is permissible, this duty will provide a reason, whether or not an imperative one, to tell the lie. The second reason is one of self-respect. The murderer wants to make you a tool of evil; he regards your integrity as a useful sort of predictability. He is trying to use you, and your good will, as a means to an evil end. You owe it to humanity in your own person not to allow your honesty to be used as a resource for evil. I think this would be a perfect duty of virtue; Kant does not say this specifically but in his discussion of servility (the avoidance of which is a perfect duty of virtue) he says "Do not suffer your rights to be trampled underfoot by others with impunity." (MMV 436/99) Both of these reasons spring from duties of virtue. A person with a good character will tell the lie. Not to tell it is morally bad. But there is no duty of justice to tell the lie. If we do not tell it, we cannot be punished,

* I owe this example to John Koethe.

† For a discussion of this question see Barbara Herman, "Mutual Aid and Respect for Persons" *Ethics* 94 (July 1984) pp. 577-602.

or, say, treated as an accessory to the murder. Kant would insist that even if the lie ought to be told this does not mean that the punctiliously truthful person who does not tell it is somehow implicated in the murder. It is the murderer, not the truthful person, who commits this crime.

Telling the truth cannot be part of the crime. On Kant's view, persons are not supposed to be responsible for managing each other's conduct.

If the lie were a duty of justice, we would be responsible for that. These reflections will help us to think about the second casuistical problem, the lie to the philanthropist. I think it does follow from the line of argument I have taken that the lie cannot be shown to be impermissible. Although the philanthropist can hardly be called evil, he is doing something tricky and underhanded, which Kant's view disapproves. He should not use this method of getting the information he wants. This is especially true if the reason he does not use a more straightforward method is that he assumes that if he does people will lie to him. We are not supposed to base our actions on the assumption that other people will behave badly. Assuming this does not occur in an institutional context, and you have not sworn that your remarks were true,[*] the philanthropist will have no recourse to justice if you lie to him. But the reasons that favor telling the lie that exist in the first case do not exist here. According to Kant, you do not have a duty to promote your own happiness. Nor would anyone perform such an action out of self-respect. This is, in a very trivial way, a case of dealing with evil. But you can best deal with it by telling the philanthropist that you know what he is up to, perhaps even that you find it sneaky. This is *because* the ideal that makes his action a bad one is an ideal of straightforwardness in human relations. This would also be the best way to deal with the murderer, if it *were* a way to deal with a murderer. But of course it is not.

Ideal and Non-Ideal Theory

I now turn to the question of what structure an ethical theory must have in order to accommodate this way of thinking. In *A Theory of Justice*,[†] John Rawls proposes a division of moral philosophy into ideal

NOTES

[*] In the *Lectures on Ethics*, Kant takes the position that you may lie to someone who lies to or bullies you as long as you don't say specifically that your words will be true. He claims this is not lying, because such a person should not expect you to tell the truth. (LE 227,229)

[†] John Rawls, *A Theory of Justice*. Cambridge, Massachusetts: Harvard University Press, 1971. Section and page numbers referring to this work will appear in the text.

NOTES

and non-ideal theory. In that work, the task of ideal theory is to determine "what a perfectly just society would be like," while non-ideal theory deals with punishment, war, opposition to unjust regimes, and compensatory justice. (§2,p. 8-9) Since I wish to use this feature of Rawls's theory for a model, I am going to sketch his strategy for what I will call a double-level theory.

Rawls identifies two conceptions of justice, which he calls the general conception and the special conception. (§§11,26,39,46) The general conception tells us that all goods distributed by society, including liberty and opportunity, are to be distributed equally unless an unequal distribution is to the advantage of everyone, and especially those who fall on the low side of the inequality. (§13) Injustice, according to the general conception, occurs whenever there are inequalities that are not to the benefit of everyone.(§11, p. 62) The special conception in its most developed form removes liberty and opportunity from the scope of this principle and says they must be distributed equally, forbidding trade-offs of these goods for economic gains. It also introduces a number of priority rules, for example, the priority of liberty over all other considerations, and the priority of equal opportunity over economic considerations. (§§11,46,82)

Ideal theory is worked out under certain assumptions. One is strict compliance: it is assumed that everyone will act justly. The other, a little harder to specify, is that historical, economic, and natural conditions are such that realization of the ideal is feasible. Our conduct towards those who do not comply, or in circumstances which make the immediate realization of a just state of affairs impossible, is governed by the principles of non-ideal theory. Certain ongoing natural conditions which may always prevent the full realization of the ideal state of affairs also belong to non-ideal theory: the problems of dealing with the seriously ill or mentally disturbed, for instance, belong in this category. For purposes of constructing ideal theory, we assume that everyone is "rational and able to manage their own affairs." (§39, p. 248) We also assume in ideal theory that there are no massive historic injustices, such as the oppression of blacks and women, to be corrected. The point is to work out our ideal view of justice on the assumption that people, nature, and history will behave themselves so that the ideal can be realized, and then to determine — in light of that ideal — what is to be done in actual circumstances, when they do not. The special conception is not applied without regard to circumstances. Special principles will be used in non-ideal conditions.

Non-ideal conditions exist when, or to the extent that, the special conception of justice cannot be realized effectively. In these circumstances our conduct is to be determined in the following way: the special conception becomes a goal, rather than an ideal to live up to: we are to work towards the conditions in which it is feasible. For

instance, suppose there is a case like this: widespread poverty or ignorance due to the level of economic development is such that the legal establishment of the equal liberties makes no real difference to lot of the disadvantaged members of society. It's an empty formality. On the other hand, some inequality, temporarily instituted, would actually tend to foster conditions in which equal liberty could become a reality for everyone. In these circumstances, Rawls's double-level theory allows for the temporary inequality. (§§ 11,39) The priority rules give us guidance as to which features of the special conception are most urgent. These are the ones that we should be striving to achieve as soon as possible. For example, if formal equal opportunity for blacks and women is ineffective, affirmative action measures may be in order. If some people claim that this causes inefficiency at first, it is neither here nor there, since equality of opportunity has priority over efficiency. The special conception may also tell us which of our non-ideal options is least bad, closest to ideal conduct. For instance, civil disobedience is better than a resort to violence not only because violence is bad in itself, but because of the way in which civil disobedience expresses the democratic principles of the just society it aspires to bring about. (§ 59) Finally, the general conception of justice commands categorically. In sufficiently bad circumstances none of the characteristic features of the special conception may be realizable. But there is no excuse, *ever*, for violation of the general conception. If inequalities are not benefiting those on the lower end of them in some way, they are simply oppression. The general conception, then, represents the point at which justice becomes uncompromising.*

A double-level theory can be contrasted to two types of single-level theory, both of which in a sense fail to distinguish the way we should behave in ideal and in nonideal conditions, but which are at opposite extremes. A consequentialist theory such as utilitarianism does not really distinguish ideal from non-ideal conditions. Of course, the utilitarian can see the difference between a state of affairs in which everyone can be made reasonably happy and a state of affairs in which the utilitarian choice must be for the “lesser of evils”, but it is still really a matter of degree. In principle we do not know what counts as a state in which everyone is “as happy as possible” absolutely. Instead, the utilitarian wants to make everyone as happy as possible relative to the circumstances, and pursues this goal holds regardless of how friendly

NOTES

* In a non-ideal case, one's actions may be guided by a more instrumental style of reasoning than in ideal theory. But non-ideal theory is not a form of consequentialism. There are two reasons for this. One is that the goal set by the ideal is not just one of good consequences, but of a just state of affairs. If a consequentialist view is one that defines right action entirely in terms of good consequences (which are not themselves defined in terms of considerations of rightness or justice) then non-ideal theory is not consequentialist. The second reason is that the ideal will also guide our choice among non-ideal alternatives, importing criteria for this choice other than effectiveness. I would like to thank Alan Gewirth for prompting me to clarify my thoughts on this matter, and David Greenstone for helping me to do so.

NOTES

the circumstances are to human happiness. The difference is not between ideal and nonideal states of affairs but simply between better and worse states of affairs.

Kant's theory as he understood it represents the other extreme of single-level theory. The standard of conduct he sets for us is designed for an ideal state of affairs: we are always to act as if we were living in a Kingdom of Ends, regardless of possible disastrous results. Kant is by no means dismissive towards the distressing problems caused by the evil conduct of other human beings and the unfriendliness of nature to human ideals, but his solution to these problems is different. He finds in them grounds for a morally motivated religious faith in God.* Our rational motive for belief in a moral author of the world derives from our rational need for grounds for hope that these problems will be resolved. Such an author would have designed the laws of nature so that, in ways that are not apparent to us, our moral actions and efforts do tend to further the realization of an actual Kingdom of Ends. With faith in God, we can trust that a Kingdom of Ends will be the consequence of our actions as well as the ideal that guides them.

In his *A Critique of Utilitarianism*,† Bernard Williams spells out some of the unfortunate consequences of what I am calling single-level theories. According to Williams, the consequentialist's commitment to doing whatever is necessary to secure the best outcome may lead to violations of what we would ordinarily think of as integrity. There is no kind of action that is so mean or so savage that it can *never* lead to a better outcome than the alternatives. A commitment to always securing the best outcome never allows you to say “bad consequences or not, this is not the sort of thing I do; I am not that sort of person.” And no matter how mean or how savage the act required to secure the best outcome is, the utilitarian thinks that you will be irrational to regret that you did it, for you will have done what is in the straightforward sense the right thing.‡ A Kantian approach, by defining a determinate *ideal* of conduct to live up to rather than setting a *goal* of action to strive for, solves the problem about integrity, but with a high price. The advantage of the Kantian approach is the definite sphere of responsibility. Your share of the responsibility for the way the world is well-defined and limited, and if you act as you ought, bad outcomes are not your responsibility. The trouble is that in cases such as that of the

* See the "Dialectic of Pure Practical Reason" of the *Critique of Practical Reason*, and the *Critique of Teleological Judgment*, §87.

† Bernard Williams, in *Utilitarianism For and Against*, by J.J.C. Smart and Bernard Williams (Cambridge: Cambridge University Press, 1973), pp. 75-150.

‡ Williams also takes this issue up in "Ethical Consistency" originally published in the Supplementary Volumes to the *Proceedings of the Aristotelian Society* XXXIX, 1965, and reprinted in his collection *Problems of the Self* (Cambridge: Cambridge University Press, 1973), pp. 166-186.

murderer at the door it seems grotesque simply to say that I have done my part by telling the truth and the bad results are not my responsibility.

The point of a double-level theory is to give us both a definite and well-defined sphere of responsibility for everyday life and some guidance, at least, about when we may or must take the responsibility of violating ideal standards. The common sense approach to this problem uses an intuitive quantitative measure: we depart from our ordinary rules and standards of conduct when the consequences of following them would be "very bad." This is unhelpful for two reasons. First, it leaves us on our own about determining *how* bad. Second, the attempt to justify it leads down a familiar consequentialist slippery slope: if very bad consequences justify a departure from ordinary norms, why do not slightly bad consequences justify such a departure? A double-level theory substitutes something better than this rough quantitative measure.

In Rawls's theory, for example, a departure from equal liberty cannot be justified by the fact that the consequences of liberty are "very bad" in terms of mere efficiency. This does not mean that an endless amount of inefficiency will be tolerated, because presumably at some point the inefficiency may interfere with the effectiveness of liberty. One might put the point this way: the measure of "very bad" is not entirely intuitive but rather, bad enough to interfere with the reality of liberty. Of course this is not an algorithmic criterion and cannot be applied without judgment, but it is not as inexact as a wholly intuitive quantitative measure, and, importantly, does not lead to a consequentialist slippery slope.

Another advantage of a double-level theory is the explanation it offers of the other phenomenon which Williams is concerned about: that of regret for doing a certain kind of action even if in the circumstances it was the "right" thing. A double-level theory offers an account of at least some of the occasions for this kind of regret. We will regret having to depart from the ideal standard of conduct, for we identify with this standard and think of our autonomy in terms of it. Regret for an action we would not do under ideal circumstances seems appropriate even if we have done what is clearly the right thing.*

NOTES

* It is important here to distinguish two kinds of exceptions. As Rawls points out in "Two Conceptions of Rules" (*The Philosophical Review*, Volume 64 (January 1965)), a practice such as promising may have certain exceptions built into it. Everyone who has learned the practice understands that the obligation to keep the promise is cancelled if one of these obtains. When one breaks a promise because this sort of exception obtains, regret would be inappropriate and obsessive. And these sorts of exceptions may occur even in "ideal" circumstances. The kind of exception one makes when dealing with evil should be distinguished from exceptions built into practices.

NOTES

Kantian Non-Ideal Theory

Rawls's special conception of justice is a stricter version of the egalitarian idea embodied in his general conception. In the same way, it can be argued that the Formula of Universal Law and the Formula of Humanity are expressions of the same idea — that humanity is the source of value, and of the justifying force of reason. But the Formula of Humanity is stricter, and gives implausible answers when we are dealing with the misconduct of others and the recalcitrance of nature. This comparison gives rise to the idea of using the two formulas and the relation between them to construct a Kantian double-level theory of individual morality, with the advantages of that sort of account. The Formulas of Humanity and the Kingdom of Ends will provide the ideals which govern our daily conduct. When dealing with evil circumstances we may depart from this ideal. In such cases, we can say that the Formula of Humanity is inapplicable because it is not designed for use when dealing with evil. But it can still guide our conduct. It defines the goal towards which we are working, and if we can generate priority rules we will know which features of it are most important. It gives us guidance about which of the measures we may take is the least objectionable.

Lying to deceivers is not the only case in which the Formula of Humanity seems to set us a more ideal standard than the Formula of Universal Law. The arguments made about lying can all be made about the use of coercion to deal with evil-doers. Another, very difficult case in which the two formulas give different results, as I think, is the case of suicide. Kant gives an argument against suicide under the Formula of Universal Law, but that argument does not work.* Yet under the Formula of Humanity we can give a clear and compelling argument against suicide: nothing is of any value unless the human person is so, and it is a great crime, as well as a kind of incoherence, to act in a way that denies and eradicates the source of all value. Thus it might be possible to say that suicide is wrong from an ideal point of view, though justifiable in circumstances of very great natural or moral evil.

* Kant's argument depends on a teleological claim: that the instinct whose office is to impel the improvement of life cannot universally be used to destroy life without contradiction. (G 422/40) But as I understand the contradiction in conception test, teleological claims have no real place in it. What matters is not whether nature assigns a certain purpose to a certain motive or instinct, but whether everyone with the same motive or instinct could act in the way proposed and still achieve their purpose. There is simply no argument to show that everyone suffering from acute misery could not commit suicide and still achieve the purpose of ending that misery.

There is also another, rather different sense of "rigorism" in which the Formula of Humanity seems to be more rigorous than that of Universal Law. It concerns the question whether Kant's theory allows for the category of merely permissible ends and actions, or whether we must always be doing something that is morally worthy: that is, whether we should *always* pursue the obligatory ends of our own perfection and the happiness of others, when no other duty is in the case. The Formula of Universal Law clearly allows for the category of the permissible. Indeed, the first contradiction test is a test of permissibility. But in the *Metaphysical Principles of Virtue*, there are passages which have sometimes been taken to imply that Kant holds the view that our conduct should always be informed by morally worthy ends. (MMV 390/48) The textual evidence is not decisive. But the tendency in Kant's thought is certainly there: for complete moral worth is only realized when our actions are not merely in accordance with duty but from duty, or, to say the same thing a different way, perfect autonomy is only realized when our actions and ends are completely determined by reason, and this seems to be the case only when our ends are chosen as instantiations of the obligatory ends.

Using the Formula of Humanity it is possible to argue for the more "rigorous" interpretation. First, the obligatory ends can be derived more straightforwardly from Humanity than from Universal Law. Kant does derive the obligatory ends from the Formula of Universal Law, but he does it by a curiously round-about procedure in which someone is imagined formulating a maxim of rejecting them and then finding it to be impermissible. This argument does not show that there would be a moral failing if the agent merely unthinkingly neglected rather than rejecting these ends. The point about the pervasiveness of these ends in the moral life is a more complicated one, one that follows from their adoption by this route: Among the obligatory ends is our own moral perfection. Pursuing ends that are determined by reason, rather than merely acceptable to it, cultivates one's moral perfection in the required way. (MMV 380-381/37-38; 444-447/108-111)

It is important to point out that even if this is the correct way to understand Kant's ideal theory, it does not imply that Kantian ethics commands a life of conventional moral "good deeds." The obligatory ends are one's own perfection and the happiness of others; to be governed by them is to choose instantiations of these larger categories as the aim of your vocation and other everyday activities. It is worth keeping in mind that natural perfection is a large category, including all the activities that cultivate body and mind. Kant's point is not to introduce a strenuous moralism but to find a place for the values of perfectionism in his theory. But this perfectionism will be a part of ideal theory if the argument for it is based on the Formula of Humanity and cannot be derived from that of Universal Law. This seems to me to

NOTES

NOTES

be a desirable outcome. People in stultifying economic or educational conditions cannot really be expected to devote all their spare time to the cultivation of perfectionist values. But they can be expected not to do what is impermissible, not to violate the Formula of Universal Law. Here again, the Formula of Humanity sheds light on the situation even if it is not directly applied: it tells us why it is morally as well as in other ways regrettable that people should be in such conditions.

Conclusion

If the account that I have given is correct, the resources of a double-level theory may be available to the Kantian. The Formula of Humanity and its corollary, the vision of a Kingdom of Ends, provide an ideal to live up to in daily life as well as a long term political and moral goal for humanity. But it is not feasible always to live up to this ideal, and where the attempt to live up to it would make you a tool of evil, you should not do so. In evil circumstances, but only then, the Kingdom of Ends can become a goal to seek rather than an ideal to live up to, and this will provide us with some guidance. The Kantian priorities — of justice over the pursuit of obligatory ends, and of respect over benevolence — still help us to see what matters most. And even in the worst circumstances, there is always the Formula of Universal Law, telling us what we must in not in any case do. For whatever bad circumstances may drive us to do, we cannot possibly be justified in doing something which others in those same circumstances could not also do. The Formula of Universal Law provides the point at which morality becomes uncompromising.

Let me close with some reflections about the extent to which Kant himself might have agreed with this modification of his views. Throughout this paper, I have portrayed Kant as an uncompromising idealist, and there is much to support this view. But in the historical and political writings, as well as in the *Lectures on Ethics*, we find a somewhat different attitude. This seems to me to be especially important: Kant believes that the Kingdom of Ends on earth, the highest political good, can only be realized in a condition of peace. (MMJ 354-355/127-129) But he does not think that this commits a nation to a simple pacifism that would make it the easy victim of its enemies. Instead, he draws up laws of war in which peace functions not as an uncompromising ideal to be lived up to in the present but as a long range goal which guides our conduct even when war is necessary. (PP 343-348/85-91; MMJ 343-351/114-125) If a Kantian can hold such a view for the conduct of nations, why not for that of individuals? If this is right, the task of Kantian moral philosophy is to draw up for individuals something analogous to Kant's laws of war: special principles to use when dealing with evil.

SPLIT-LEVEL DEONTOLOGY

Korsgaard accomplishes something quite spectacular in her reorganization of deontological ethics. Like Hare reworks utilitarianism to account for slavery, she restructures deontology to account for those times when the world is upside down. The two levels are still both governed by a deontological ethic and the single Categorical Imperative, but we can tease apart the different formulations to account for abnormal times.

Ideal and Non-Ideal Levels

The first level she calls the ***Ideal Theory*** level. Don't be confused by the term 'ideal,' however, since by this she does not mean pie-in-the-sky utopian dreams but down-to-earth everyday normalcy. Remember that deontology is the way we should, as everyday human beings, live our lives. So the *ideal* is simply those times and places where we're just being everyday human beings—rational, fallible, normal people. The ideal circumstance, then, is one where most people are most of the time doing roughly what they're supposed to be doing.

The second level indicates that *absolute*, unyielding standard. Face it: life isn't always ideal. When circumstances elevate, we still need to have a moral standard to guide our actions. In these situations, there remains the standard that marks out those things that we should never—not even in extreme situations—do. This level Korsgaard refers to as the ***Non-Ideal Theory*** level. In non-ideal circumstances, morality still stands. The ideal provides guidance on how we should act, by pointing to a world we seek to re-establish. We don't always live in an ideal situation, but we can always act in such a way that we can allow such to re-assert itself.

In ideal (or ordinary) circumstances, we follow the Formula of Humanity (FH). Since most everyone is trying also to respect the humanity in each other, this makes sense. We don't ordinarily have murderers coming to our door asking if we can donate victims to their cause. We don't ordinarily confront people whose mission is to dehumanize others or violate the Categorical Imperative. In such times, we don't need to worry about the horrifying consequences of obeying the CI by not lying or some other action that is normally obviously correct. The FH identifies the ultimate good: a world of rational lawgivers, of moral legislators, each attempting to share in that co-establishment of justice.

But life often slips out of the ideal into the less-than ideal. When the norms collapse, and it becomes commonplace for the intrinsic value of persons to be trampled upon, then respecting the intrinsic humanity in everyone is muddled. The murderer at the door is *not* acting rationally, but from an emotional or hypothetical imperative. In such conflicting times to obey the FH would seem to entail contradicting the Categorical Imperative itself. Thus whatever we do should be aimed *towards* the re-creation of the ideal, where most people act mostly rationally. But we can't just chuck the CI, so in these circumstances we still act according to the Formulation of Universal Law—we ourselves are, after all, still moral legislators, so to preserve our own humanity we need to persist in rational action, because it is our rational action that will pave the way back to normalcy.

A Duty to Lie?

So how does this work? You're at your door, in a non-ideal world where the murderer is asking you if you

I'M A UTILITARIAN, SO I DON'T SEE THE RULE AGAINST LYING AS ABSOLUTE; IT'S ALWAYS SUBJECT TO SOME OVERRIDING UTILITY WHICH MAY PREVENT ITS EXERCISE.

(PETER SINGER)

know of any Tutsis, Jews, or Croats are around. The FUL says you cannot lie. We surely seem to be stuck in the same spot we were before we read Korsgaard! Maybe. Let's look deeper. Korsgaard tells us that the *aim* of moral action is to *treat humanity always as an end*, as intrinsically valuable. In such situations as our thought experiment, humanity is not being respected by either the murderer at the door or by our mandate to tell the truth: if we do, we violate the trust of our shelter-seeking neighbors, and we legislate poorly, given what we know of the circumstances.

Interestingly, we might in such circumstances have a *duty* to lie. But we only have this duty in situations where these two criteria are met:

1. the lie is permissible (it doesn't violate FUL), and
2. the lie involves a "duty of mutual aid" and self-respect (i.e., something that promotes human dignity and respect)

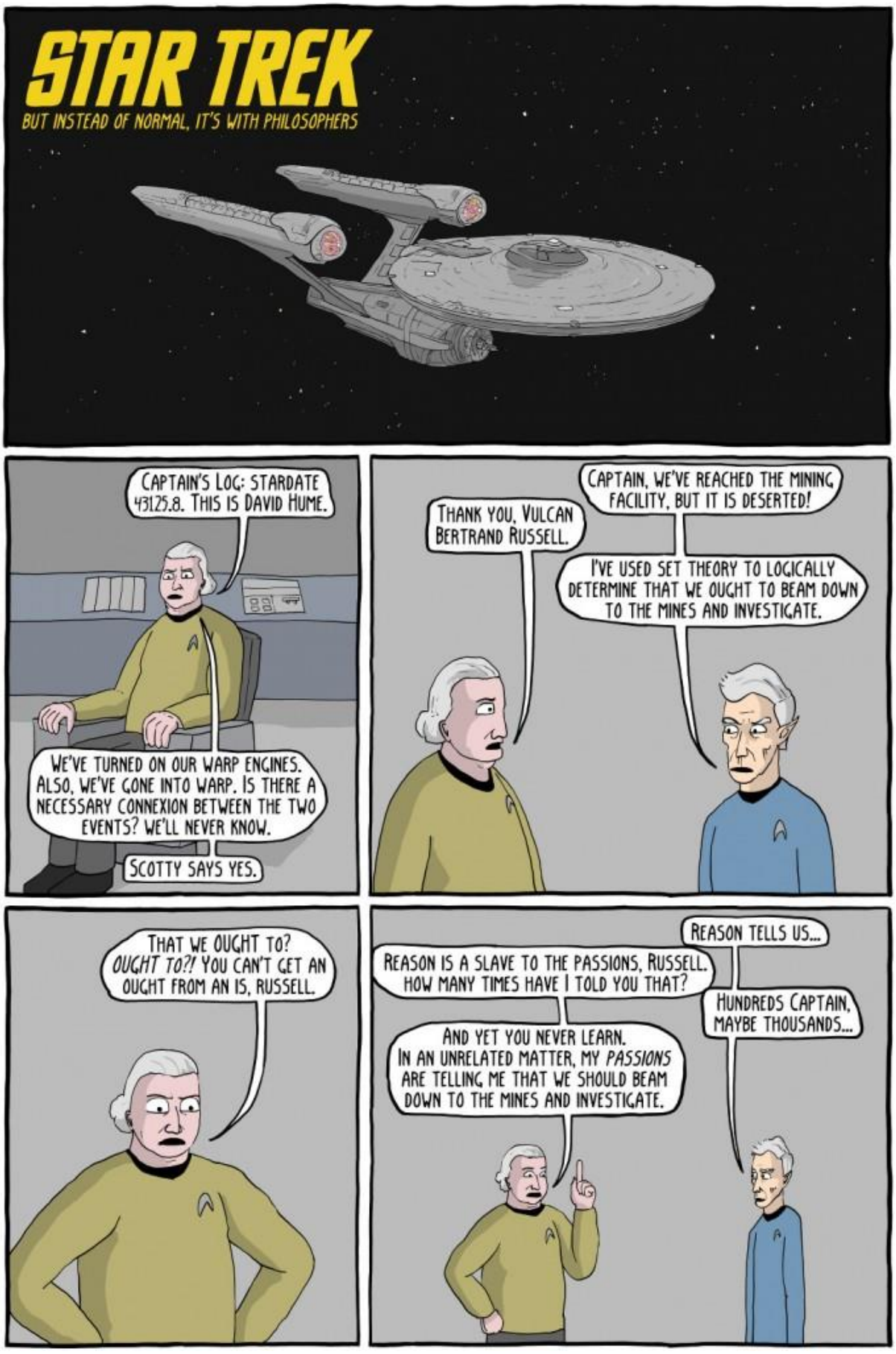
Here's how this would cash out. If lying is necessary to *promote* the intrinsic value of humanity (for example, by saving a life), then it would be our *moral obligation* to lie.

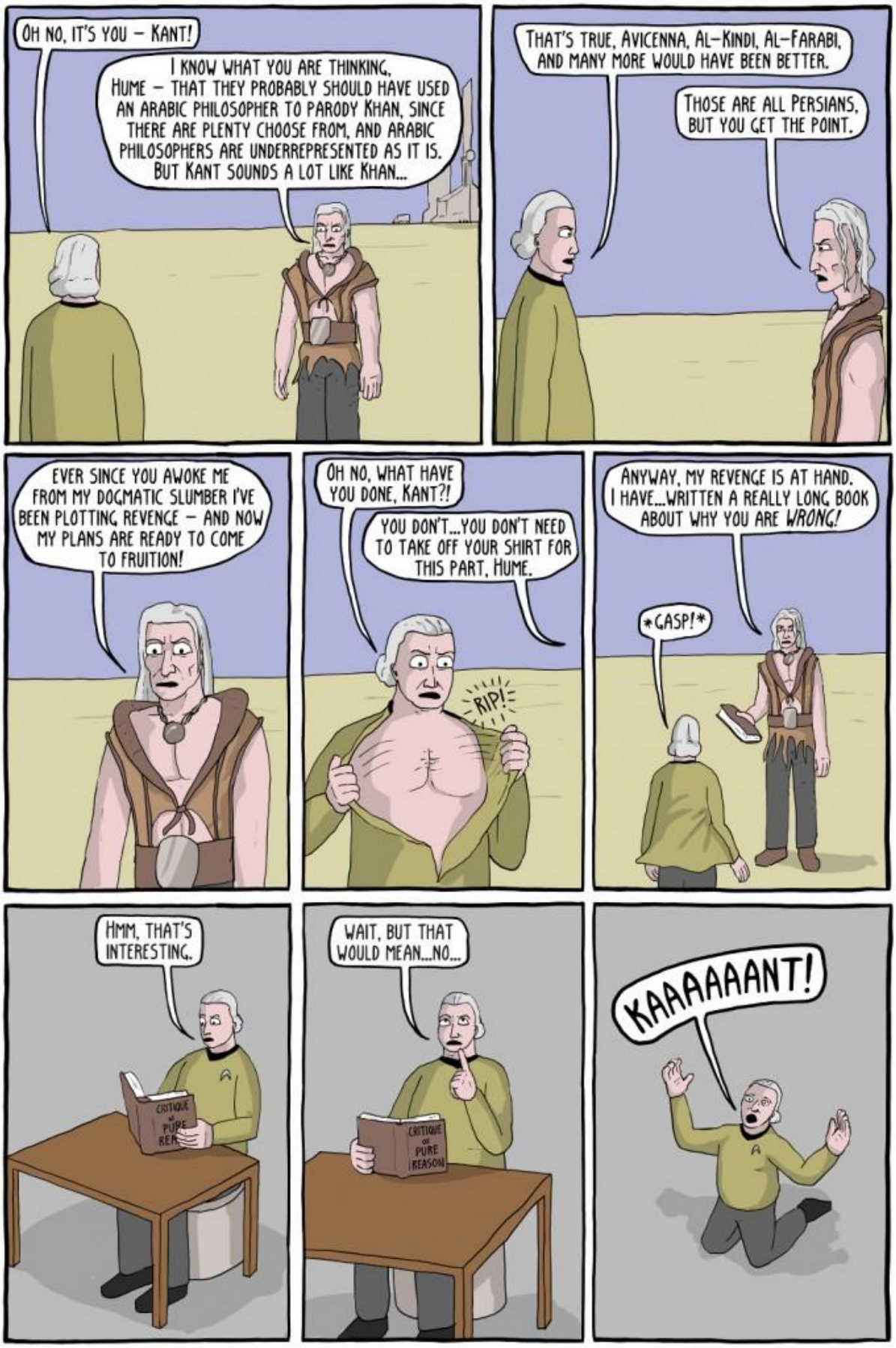
How can we lie to the murderer without violating FUL? To be frank, it's not generally the case, even in non-ideal situations that the murderer at the door will make explicit his intentions or even his identity. We often *assume* we know the motives of somebody. It could be the case, for example, that the uniformed fellow at your door is himself one of the good guys, trying to smuggle Jews away in his van (that did happen, after all). You don't know. And often the murderer himself is lying to you, or trying to—wanting to deceive you into thinking he's got some legit motive, like relocation or ghettoizing the 'undesirables.'

So for the logical contradiction to arise there needs to be some intricate dance of knowledge regarding who knows what about whom. For the murderer to know that you're lying about what you know, he has to know that you know he's a murderer. He has to first know that *it's a universal law of human nature that everybody always lies to murderers to protect an innocent neighbor*, and he has to know also that *you* know that he is one such murderer. That is, he needs two key bits of knowledge in order for the contradiction to emerge. But it seems that in such a world, he would wish to hide his identity so that you wouldn't lie to him.

In fact, this is often how non-ideal situations look: murderers hide their true identities *as* murderers, to ensure access to their victims. So unless the murderer identifies himself clearly *as* a murderer, no contradiction emerges. You are permitted to lie. And more carefully, you are *obligated* to lie if the intention behind your lie is to promote human dignity. Notice how the FH is still at work even in non-ideal circumstances. It is the *goal*, though not strictly the immediate motivation.

This still leaves us with a bit of a tension regarding deontology. The *only* thing that makes lying permissible is if the one being lied to doesn't know what I know, that he doesn't know I know he's also being deceitful. Is this gap in knowledge a solid foundation for morality, strong enough to meet the universal, unchanging, absolute criteria Kant set out to establish? In logic, we call an argument based on lack of evidence a fallacy of ignorance. Is the acceptability of deceit here a case of fallacious reasoning?


STAR TREK
BUT INSTEAD OF NORMAL, IT'S WITH PHILOSOPHERS
CAPTAIN'S LOG: STARDATE 43125.8. THIS IS DAVID HUME.
WE'VE TURNED ON OUR WARP ENGINES. ALSO, WE'VE GONE INTO WARP. IS THERE A NECESSARY CONNEXION BETWEEN THE TWO EVENTS? WE'LL NEVER KNOW.
SCOTTY SAYS YES.
CAPTAIN, WE'VE REACHED THE MINING FACILITY, BUT IT IS DESERTED!
THANK YOU, VULCAN BERTRAND RUSSELL.
I'VE USED SET THEORY TO LOGICALLY DETERMINE THAT WE OUGHT TO BEAM DOWN TO THE MINES AND INVESTIGATE.
THAT WE OUGHT TO? *OUGHT TO?!* YOU CAN'T GET AN OUGHT FROM AN IS, RUSSELL.
REASON TELLS US...
REASON IS A SLAVE TO THE PASSIONS, RUSSELL. HOW MANY TIMES HAVE I TOLD YOU THAT?
HUNDREDS CAPTAIN, MAYBE THOUSANDS...
AND YET YOU NEVER LEARN. IN AN UNRELATED MATTER, MY *PASSIONS* ARE TELLING ME THAT WE SHOULD BEAM DOWN TO THE MINES AND INVESTIGATE.


OH NO, IT'S YOU – KANT!
I KNOW WHAT YOU ARE THINKING, HUME – THAT THEY PROBABLY SHOULD HAVE USED AN ARABIC PHILOSOPHER TO PARODY KHAN, SINCE THERE ARE PLENTY CHOOSE FROM, AND ARABIC PHILOSOPHERS ARE UNDERREPRESENTED AS IT IS. BUT KANT SOUNDS A LOT LIKE KHAN...
THAT'S TRUE, AVICENNA, AL-KINDI, AL-FARABI, AND MANY MORE WOULD HAVE BEEN BETTER.
THOSE ARE ALL PERSIANS, BUT YOU GET THE POINT.
EVER SINCE YOU AWOKE ME FROM MY DOGMATIC SLUMBER I'VE BEEN PLOTTING REVENGE – AND NOW MY PLANS ARE READY TO COME TO FRUITION!
OH NO, WHAT HAVE YOU DONE, KANT?!
YOU DON'T...YOU DON'T NEED TO TAKE OFF YOUR SHIRT FOR THIS PART, HUME.
RIP!
ANYWAY, MY REVENGE IS AT HAND. I HAVE...WRITTEN A REALLY LONG BOOK ABOUT WHY YOU ARE WRONG!
GASP!
HMM, THAT'S INTERESTING.
CRITIQUE OF PURE REA
WAIT, BUT THAT WOULD MEAN...NO...
CRITIQUE OF PURE REASON
KAAAAAANT!