CHAPTER 6

Rightness as Fairness

We are now in a position to derive moral principles from the Categorical-Instrumental Imperative. Chapter 5 showed that the Categorical-Instrumental Imperative entails the following method for determining which actions satisfy it:

> **The Moral Original Position Formulation:** voluntarily aim for its own sake, in every relevant action, to act on interests it is instrumentally rational to act upon from the standpoint of a 'Moral Original Position' in which you assume that your voluntary, involuntary, and semivoluntary interests could turn out to be identical to those of any human or non-human sentient being(s), where relevant actions are defined recursively as those it is instrumentally rational to treat as such from the standpoint of the Moral Original Position.

§1 of this chapter shows that we can derive the following four principles of fairness from the Moral Original Position:

> **Four Principles of Fairness**
>
> **The Principle of Negative Fairness:** all of our morally relevant actions should have as a guiding ideal, setting all costs aside, avoiding and minimizing coercion in all its forms (coercion resulting from intentional acts, natural forces, false beliefs, and so on), for all human and non-human sentient beings, for its own sake.

The Principle of Positive Fairness: all of our morally relevant actions should have as a guiding ideal, setting all costs aside, assisting all human and nonhuman sentient beings in achieving interests they cannot best achieve on their own and want assistance in achieving, for its own sake.

The Principle of Fair Negotiation: whether an action is morally relevant, and how the Principles of Negative and Positive Fairness and Virtues of Fairness (see below) should be applied factoring in costs, should be settled through an actual process of fair negotiation guided by the Principles of Negative Fairness, Positive Fairness and Virtues of Fairness, where all human and non-human sentient beings affected by the action are afforded equal bargaining power to the extent that such a process can be approximated, and to the extent that cannot be, through a hypothetical process approximating the same, for its own sake.

The Principle of Virtues of Fairness: all of our morally relevant actions should aim to develop and express stable character traits to act in accordance with the first three principles of fairness, for its own sake.

§2 then combines these four principles into the following analysis of moral rightness:

Rightness as Fairness: an action is morally right if and only if it satisfies the Principles of Rightness as Fairness, that is, if and only if it is (A) is morally relevant, (B) has coercion-avoidance and minimization, assisting human and non-human sentient beings to achieve interests they cannot best achieve on their own and want assistance in achieving, and the development and expression of settled dispositions

to have these ends, as at least tacit ideals, and (C) is in conformity with the outcome of an actual process of fair negotiation approximating all human and sentient beings affected by the action being motivated by the above ideals and having equal bargaining power over how those ideals should be applied factoring in costs, or, if such a process is impossible, the outcome of a hypothetical process approximating the same, where moral relevance is determined recursively, by applying (B) and (C) to the question of whether the action is morally relevant.

Next, §2 argues that Rightness as Fairness reconciles several traditionally opposed conceptions of morality: deontology, consequentialism, virtue ethics, and contractualism. Additionally, §2 contends that Rightness as Fairness requires abandoning a common but problematic conception of moral problem-solving—a conception according to which we can arrive at sound answers to moral problems through principled thought or debate alone—in favor of an alternative method of 'principled fair negotiation' that requires merging principled thought and debate with actual, real-world negotiation.

Finally, §3 applies Rightness as Fairness to a small but representative variety of applied moral issues: (§3.1) Kant's famous four cases from *The Groundwork of the Metaphysics of Morals* (making a false promise, suicide, helping those in need, and developing one's natural talents), (§3.2) the question of whether, and if so when, morality permits or requires sacrificing the few for the many, (§3.3) world poverty, (§3.4) the distribution of scarce medical resources (e.g. transplantable organs), and (§3.5) the ethical treatment of non-human animals.

1 Derivation of Four Principles of Fairness

Our first task is to determine which interests it is instrumentally rational to act on from the standpoint of the Moral Original Position: a standpoint behind an Absolute Veil of Ignorance where one assumes one's interests could turn out to be identical to those of any possible human being or non-human sentient being(s). Any action that is instrumentally rational from this standpoint is one that is instrumentally rational for one's present and all possible selves to universally agree to act on for their own sake, thus satisfying the Categorical-Instrumental Imperative. We will now see that the Moral Original Position generates four principles of fairness.

1.1 The Principle of Negative Fairness

In order to determine which interests it is instrumentally rational to act upon from the standpoint of the Moral Original Position, we must reflect carefully upon one's deliberative situation within it.

First, one knows one is a *moral agent*: someone who experiences themselves first-personally as having capacity to voluntarily motivate themselves to act on principles of one's own choice (qua the Kantian and Hybrid Models of first-personal deliberation examined in Chapter 2).

Second, one is behind an *Absolute Veil of Ignorance*: one is to assume that one's interests could turn out to be identical with the interests of any possible human or non-human sentient being(s)—one's mother's interests, a stranger's interests, some animal's interests, the interests of many different people or animals, and so on.

Third, for reasons defended in Chapter 3, one ought not to deliberate on the basis of probabilities. For although it may be more likely that you will have some interests as

opposed to others—say, your interests in your own well-being over the well-being of others—the Categorical-Instrumental Imperative requires one not to bet on 'likely future selves', but instead forge and uphold a universal agreement with all of one's possible future selves: an agreement that satisfies the interests of every possible future person you could turn out to be, no matter how unlikely. Since you are to deliberate behind the veil as though you could have any possible set of interests—including different interests in risk-taking—you cannot deliberate on the basis of expectations about which risks are rational.

Fourth, because the Categorical-Instrumental Imperative requires one to seek and uphold such a universal agreement, one should assume that every possible future self one could turn out to be is voluntarily committed to acting on whatever interests turn out to be rational from one's standpoint behind the veil. This point is critical, because although one is to assume behind the Absolute Veil of Ignorance that one could turn out to have any possible set of interests—including the interests of people who violate moral norms, such as criminals, liars, and cheats—the Categorical-Instrumental Imperative requires one not to 'bet' on having those interests, but rather to forge and uphold an instrumentally rational agreement with all of one's possible selves. For, following Chapters 2 and 3, we are approaching these issues in Problem of Possible Future Selves cases: cases that we all experience at least sometimes in our lives, and in which we do not want to 'bet' on likely outcomes.

It is also critical to clarify that the above assumption—that one will voluntarily act on or 'uphold' whatever interests are instrumentally rational from the standpoint of the Moral Original Position, no matter which 'self' you turn out to be—is not an assumption that other people will also act on those same interests (namely, the Four Principles of

Fairness that emerge from the Moral Original Position). The assumption here is not a Rawlsian assumption of 'strict-compliance', one which assumes that everyone else will comply with whichever principles one selects.[1] Instead, it merely assumes that *you* will be voluntarily committed to acting on those interests, no matter which possible future 'self' you turn out to be. It is an assumption that oneself is committed to doing what is rational in Problem-cases (vis-à-vis one's possible future selves), whatever other people might do. And this, I believe is exactly what we intuitively want a moral theory to do: we want it to tell us what we should do, regardless of what others do.

Fifth, although some of the beings one could turn out to identify one's interests with—nonhuman animals in particular—may not have capacities to voluntarily tailor their own interests to the outcome of one's deliberations (the Four Principles of Fairness), one should understand one's deliberations as including their interests by proxy. In other words, one should voluntarily uphold the Four Principles of Fairness on nonhuman animals' behalf because their interests are possibly yours. The Moral Original Position thus models two notions regarding nonhuman animals: first, how we should treat them (treating their interests as possibly our own), and second, how they should hypothetically treat each other if (contrary to actual fact) they were capable of behaving morally.

Finally, one should know in the Moral Original Position that the different human and nonhuman sentient beings one could turn out to identify one's interests with can have several different types of interests: voluntary, involuntary, and semivoluntary interests. This is critical for the following reason. We saw in Chapter 2 that Michael Smith responds to something like the Problem of Possible Future Selves by affirming a rational requirement to render one's motivations consistent across time. This is not unlike Kant's

notion of a 'kingdom of ends': a systematic union of the ends of all rational agents, without any conflicts between them. However, I argued in Chapter 2 that seeking such consistency is not straightforwardly rational because it fails to account for the fact that we have involuntary interests which we cannot avoid or change (for instance, we may find ourselves angry at someone whether we like it or not), as well as semivoluntary interests that we can voluntarily alter, but only within certain bounds and at some cost to oneself (as there are many possible costs to treating others fairly).

The six assumptions just discussed enable us to fully specify one's deliberative situation in the Moral Original Position. One knows, behind its Absolute Veil of Ignorance, that (i) one might turn out to have any possible set of interests, but (ii) one does not know which, and one cannot 'bet' probabilistically on any set of them over any others. For all one knows, once the Absolute Veil of Ignorance is 'raised', one may simply be self-interested, caring about no one's interests but one's own. At the same time, however, for all one knows the opposite will be true once the veil is 'raised': one may end up caring about the interests of some, or all, other human and nonhuman sentient beings. Consequently, there is only one instrumentally rational way to proceed. Instrumental rationality requires one to adopt the best means (or instrument) for satisfying one's interests. As we have just seen, however, one cannot rationally 'bet' on any particular set of interests in the Moral Original Position (since one is 'in' the Moral Original Position precisely as a consequence of encountering the Problem of Possible Future Selves, an instance where one does not want to bet on likelihoods). What one does know is this: no matter which interests one turns out to have, one wants to advance those interests (since they are, from your perspective, possible interests of your own). Consequently, given one's ignorance from the standpoint of

the Moral Original Position, one has a higher-order interest: an interest in advancing the interests of *all* the possible 'selves' one could turn out to be, without betting on any particular selves' interests (including a mere majority of selves one could be). If there are any principles of action (or interests one can voluntarily choose) that would be optimal means for advancing this higher-order interest—optimal means, that is, for enabling every possible self you could turn out to be to advance their ends from the standpoint of the Moral Original Position—then those are the principles that instrumental rationality requires from this standpoint, given the Problem of Possible Future Selves. Those are the principles that satisfy the Categorical-Instrumental Imperative.

The question then is whether there are any such principles. I submit there are. To see how, let us focus again on the three types of interests one can have: voluntary, involuntary, and semivoluntary. Insofar as we experience our voluntary interests as under our control (as interests we can choose), and our semivoluntary interests as partly under our control, one should not assume (behind the Absolute Veil of Ignorance) that these interests are 'fixed'. One should assume instead that one can choose which voluntary interests one's possible selves (on the other side of the veil) will have, and, within certain bounds, which semivoluntary interests they will have (though, again, altering one's semivoluntary interests can be a costly affair). Consequently, we can determine that one particular strategy for advancing our higher-order interest is instrumentally rational from the Moral Original Position: namely, that all things being equal, and setting all costs and conflicts between one's possible involuntary and semivoluntary interests aside, it is rational to bring all of one's possible voluntary and semivoluntary interests into the greatest coherence possible (we will turn to costs later). For, setting costs aside, greater

coherence among one's possible interests improves the capacity of every possible self to satisfy their interests. Allow me to explain.

Consider two worlds: one in which two persons (A and B) voluntarily choose interests that are incompatible with each other's, and another in which they voluntarily render their interests consistent with each other. Let us say that in the first world, A and B voluntarily choose to make it their interest in owning a particular home, but only one of them can own it, and that in the second world, A voluntarily chooses to let B have the home and seek a different home for herself. All things being equal, the first world is suboptimal from the standpoint of the Moral Original Position: since A and B voluntarily set their interests against each other's, someone's voluntary interests will necessarily be thwarted (only one of them can satisfy their interest). From the standpoint of the Moral Original Position, costs aside, one should prefer the second world instead. For although A might have to settle for her second or third-choice home (or, worse yet, there may not be another home for A to own, all costs we will take into account later), A and B have nevertheless removed an obstacle to them both satisfying their interests: namely, the obstacle of their interests directly contradicting each other's.

Thus, setting aside all costs and conflicts between different possible selves' involuntary and semivoluntary interests, one should aim in the Moral Original Position to bring one's possible voluntary interests into unity, or the greatest coherence possible. Doing so minimizes the number of possible future selves whose interests one's present actions might contradict, thus maximizing the probability that every possible future self will be able to successfully satisfy their interests. Notice that this is consistent with the argument in Chapter 3 explaining why moral behavior pursued for its own sake, in

Problem-cases, has infinite expected value, whereas immoral behavior pursued for its own sake does not. We can now begin to see why this is. All things being equal, motivational consistency across all of one's possible future selves permits all of one's possible selves—who are, in principle, infinite in number—to satisfy their interests. In contrast, motivational inconsistency entails that only some of one's possible future selves can satisfy their interests. There are *infinitely* more possible future selves who can successfully satisfy their interests in the former case than the latter. However, as we will see shortly, because there are potential costs associated with motivational consistency, this cannot be the end of the story.

For now, let us return to the question of what motivational consistency across one's possible future selves—which we have just seen is rational to want in the Moral Original Position, all things being equal—involves. If we focus on what it is to have a motivational interest, we can see that any interest possessed by a human or nonhuman sentient being entails (setting all costs aside) a higher-order rational interest in not being coercively prevented from obtaining the interest's object. Indeed, it is clearly true that if one wants X, then—to be instrumentally rational vis-à-vis obtaining that end—one must also want not to be coercively prevented from obtaining X (again, setting all costs aside). Given that one cannot possibly achieve X if one is coercively prevented from achieving it, if one wants X, instrumental rationality—in requiring one to adopt the best means for achieving X—requires one to want to not be coercively prevented from X.

We can see this clearly though an example. Suppose I have eating a scoop of ice cream as my dominant, voluntary interest: I want a scoop of ice cream more than anything else. What is it for me to have this as an end? The following is obviously true: if I want a

scoop of ice cream more than anything I else, I instrumentally ought to want no person or thing to coercively prevent me from getting one. If you coerce me out of getting a scoop—either by stealing my ice cream or by telling me the lie that there is no ice cream in the fridge when there actually is—then you have directly contravened my end, coercively preventing me from obtaining the object of my end. Further, coercion by other agents—force, theft, deception, etc.—is not the only way one can be 'coerced', at least, as we will now see, relative to how it is instrumentally rational to understand coercion in the Moral Original Position. Indeed, from one's standpoint behind the Absolute Veil of Ignorance, any 'preventer' of one's satisfying one's interests—not just other human beings, but forces of nature—is equally problematic. For instance, if I have eating a scoop of ice cream as an end but I am paralyzed—if, that is, impersonal forces of nature prevent me from obtaining it—then that too (i.e., the coercive force of nature) contravenes my end. To have something 'X' as an end, then, is to also have not being coercively prevented—either by other agents, or by forces of nature—from achieving X.

There are two critical points here to clarify. Coercion in the relevant sense (being prevented from obtaining your goals) need not merely be 'active' coercion by other intentional agents (other human or nonhuman beings). To the extent that one has an interest in something, X, one has an interest in avoiding both 'natural' and *self*-coercion just as much as coercion by other human beings. Allow me to each. First, forces of nature can clearly be coercive in a sense relevant to the argument above. Suppose I want a scoop of ice cream, but the hot sun melts it before I can eat it. This contravenes my interest no less than a person stealing my ice cream. Similarly, suppose I want to live but am swept into the ocean by a rip-current and am in danger of drowning. Here too a force of nature

contravenes my interest, forcibly preventing me from obtaining what I want. Second, a person can—through irrational impulse or mistakes of reasoning—contravene their own ends. If one has an interest in living but (falsely) believes that jumping off a cliff without a parachute will enable one to survive, and one jumps off a cliff on that basis (falling to one's death), one's own mistake leads to the forcible contravention of one's ends (the hard concrete ending one's life). Similarly, if one has an interest in living but finds oneself impelled by addiction to take an overdose of heroin, one's irrational impulse may contravene one's own end.

Indeed, consider the case where I am about to cross a bridge that, unbeknownst to me, is broken and will collapse if I try to cross it, hurtling me towards a gruesome death on the rocks below. If you coercively prevented me from crossing—by, say, physically tackling me—this would be a case of coercion. I might cry out and protest, 'What are you doing?' But despite the fact that you are coercing me, if I wish to live more than I want to cross the bridge, and your coercing me is the only way to stop me from trying to cross the bridge, then your coercing me does not contravene my dominant end (of staying alive). On the contrary, it satisfies my end. Similarly, if the dog wishes to eat poisonous meat, but I know that the dog has its continued living as its ultimate end—the dog, presumably, wishes to live more than it wishes to eat poisonous meat—then it is rational for the dog to be coercively prevented from eating the meat as its end. Of course, the dog might not be capable of recognizing the rationality of this—but the important thing is that *we* can recognize the dog's interest in avoiding poisoning, and see their interest as possibly our own. As such, the Moral Original Position equally requires us to care about possible ways in which animals can be coerce—which is highly intuitive, as most of us think morality

requires taking steps to protect our pets against things they have interests in avoiding (we will return to the topic of the ethical treatment of animals in general in §3.5).

In short, whenever any human or nonhuman sentient being has an interest in something, X, they thereby have an instrumentally rational interest in avoiding all types of coercion: coercion resulting from the intentional actions of other human or nonhuman beings, natural coercion (resulting from nonintentional forces of nature), and self-coercion (through mistakes of reasoning, irresistible impulses, and so on). Obviously, I have been using 'coercion' as a term of art here. But what exactly is coercion? Here I must simply defer to coercion theorists—philosophers who theorize about the nature of coercion.[2] Although the precise nature of coercion is an unsettled issue, we need not investigate it here for two reasons. First, we still have a relatively good working conception of coercion, including knowledge of paradigm cases (lying and fraud are coercive, murder is coercive, and so on). Second, this book's arguments enable us to demarcate a certain sense of 'coercion' as morally relevant: namely, unwanted direct contravention of one's interests (other agents, forces of nature, and one's own mistakes can all directly contradict one's ends, making those things *obstacles* to overcome in pursuit of one's ends), as this is what one has a higher-order interest in avoiding from the Moral Original Position.

Now, it might be suggested that there are clear counterexamples to the argument given above: the argument that whenever one has an end, one thereby has instrumentally rational grounds to want not to be coercively prevented from realizing that end. For instance, suppose I want a cigarette. Does it follow that it is instrumentally rational for me not to want anyone to coercively prevent me from smoking? One might think not: that however much one might want a cigarette, cigarettes are harmful; thus, perhaps one ought

to want to be prevented from smoking. This, however, is not a genuine counterexample. For notice: when we say that one ought to want to be prevented from smoking, we are assuming that they have a stronger interest in avoiding the negative health effects of smoking. This brings us back to the issue of costs. Clearly, setting aside all costs (to their health), a person who wants to smoke clearly does have an instrumentally rational interest in no one coercively preventing them from smoking. It is only once costs are factored into the equation—and we consider other interests the person might have that might make smoking costly for them—that it may be instrumentally rational to desire coercion against one's ends (for the sake of, say, higher ends in longevity, health, and freedom from lung cancer).

Here, then, is what every moral agent in the Moral Original Position knows: (A) they have an interest in rendering their possible voluntary and semivoluntary interests into greater coherence, all things being equal, all costs aside, and (B) that no matter which such interests they have, they necessarily have a higher-order interest in being free from coercion in pursuing their ends, setting all costs aside. These two interests entail that it is instrumentally rational in the Moral Original Position to voluntarily choose a principle of coercion-avoidance and minimization, namely:

> **The Principle of Negative Fairness:** all of our morally relevant actions should have as a guiding ideal, setting all costs aside, avoiding and minimizing coercion in the world in all its forms (coercion resulting from natural forces, intentional acts, and false beliefs), for all human and non-human sentient beings, for its own sake.

Allow me to explain why.

Consider a situation in which the interests of two human beings conflict. I am drowning in a shallow pond, and you do not wish to help me. Coercing you to help me—for instance, by imprisoning you for not helping—would hamper your ability to achieve your interests (strolling by on your merry way). But not coercing you would leave me prey to coercion by natural forces (I am drowning, after all). Although one might suggest that it is rational to favor the worst off person in this position, including from the standpoint of the Moral Original Position (as drowning is surely worse than being forced to help someone), this is still suboptimal: it coerces one person for the sake of the other. If there is no better option—and there might be no better option, if the person observing refuses to help (as the person drowning may not be capable of voluntarily 'choosing to want to drown')—then, indeed, siding with the person in the worse off position is rational. For of the two worlds available, one in which one person drowns and another in which one person is forced to help the other, the latter world is less coercive. However, this is not necessarily the only option. A better option still—one we tend to favor in the real world, on grounds of moral reciprocity—is one in which both parties negotiate with one another to render their interests consistent, so that no one has to be coerced at all. For instance, if the person walking by voluntarily chooses to save the drowning person, and then is rewarded either by the person saved or in some other way (for example, by social recognition as a 'hero') in a manner that satisfies other, stronger interests of the rescuer's, then both parties can achieve things they want—continued living (for the drowning individual) and personal satisfaction (for the person who was initially inclined to simply walk by without helping).

Indeed, there is an interesting feature of negotiating during interactions that social scientists have found in recent years: namely, that intentional agents (i.e. you, I, or even a

nonhuman animal) may not even have a stable set of interests prior to acting (that they want to be free from coercion to pursue), but rather construct their interests in arriving at a decision. In traditional decision-theory, an agent's preferences or ends are understood in terms of a person's revealed behavior.[3] For instance, if I choose to eat a piece of pizza over ice cream, decision-theorists have traditionally taken this to indicate that I preferred to eat the pizza over the ice cream. This is known as 'Revealed Preference Theory.' According to Revealed Preference Theory, intentional agents (1) have preexisting preferences, and (2) those preferences are displayed by the agent's intentional behavior. In recent years, however, social scientists have argued that there is evidence that agents instead construct their preferences in the very process of arriving at decisions.[4-5] Call this 'Constructed Preference Theory.' According to Constructed Preference Theory, an agent may have no stable preference function—or ends—prior to deciding how to act. Rather, the very act of choice is a matter of arriving at a set of preferences or ends. If this is true, then although every sentient being has their own freedom from coercion as a higher-order end, some of the possible human and nonhuman beings whose interests one might identify as one's own may have negotiating their preferences—and, by extension negotiating what comprises coercing them (since what coercion involves depends on one's preferences)—as their ends.

Indeed, negotiating our interests in order to accommodate the interests of others pervade human life, and for obvious reasons: one may not be able to obtain the things one desires (or not be able to obtain them very effectively) without negotiating. For instance, in the workplace, one's supervisor may have a certain amount of power of you. In order to get things that you want—say, a day off work to care for your sick child—you may have to negotiate to do things you previously did not want to do (for instance, work an extra day on

the weekend). Similar forms of interest-negotiation are ubiquitous in human interaction: in marriages, friendships, politics, and so on. In a marriage or friendship, one often recognizes that in order for the relationship to be happy or productive—for the relationship to be satisfying for oneself—one has to negotiate one's interests with the other partner so that both parties can also obtain things they want out of the relationship. If, for instance, one does not want to invest all their discretionary income for retirement but one's spouse feels very strongly that one should, 'sticking to one's guns' rather than negotiating a mutually acceptable set of interests (say, investing some money, but leaving enough aside for enjoyment today) can result in conflict: an unhappy situation which advances neither person's interests in the relationship. Generally speaking, in human relationships, unless one is willing to negotiate one's interests with another—unless one is willing to engage in 'give and take', seeking mutually acceptable interests—the other party is unlikely to continue the relationship, at least not in the friendly, constructive manner one may wish. As such, we often (if not always) have interests in negotiating our interests with others, and by extension, interests in what coercing us involves (I may not initially want to do the dishes, but if my spouse convinces me it is fair for me to do so, I do not regard myself as coerced but rather as persuaded to change my interests).

A further important point here is that we cannot simply assume that a person has a preexisting 'optimal negotiation point' prior to negotiating—one that we could simply settle through reflection or debate. For indeed, the point to which we are willing to negotiate can depend on contextual details, including the particular situation in which we find ourselves, the particular individuals we are negotiating with, and so on. To see how, consider a simple case in which one is negotiating with another person on how many

cookies to share from a box of cookies. I may initially have eating all ten cookies as my interest (suppose I am very hungry). However, if we bought the cookies together, I may be willing to negotiate to an even distribution, giving you half and taking the other half for me. At the same time, if you do not feel very strongly about it, you may be willing to allow me to have more than that, saying to me, 'I know we bought the cookies together, but you are hungry: go ahead and have more.' And things could become even more complicated than this. For instance, if you were to add, 'But I want more next time', I may or may not decide to take a larger share, no matter how much I want more. Similarly, things might be very different with other individuals. If the situation involves my child and I instead, I might be willing to give the child many cookies simply out of love for them (even though, all things being equal, I would like to eat them all myself). Alternatively, if I think it is bad for my child to eat too many sweets, I may choose to provide them with just a couple and eat just a couple myself in order to 'be a good role-model.' As all of these complexities illustrate, we cannot typically settle, ex ante—before an organic process of 'negotiation' occurs—where any given party to a negotiation may be willing to negotiate to, in terms of altering their interests. The negotiation itself creates its 'negotiation-end-point' in an organic fashion.

Finally, although children, the mentally disabled, and nonhuman animals may not possess the same robust negotiation capacities, they often do have them in some lesser degree. Our pets, for instance, often appear to learn our preferences, adapting their behavior to ours. My dog Tex, for instance, only tends to bring me toys to play with in the evening, as he appears to recognize that this is a particular time of the day I am happy to play with him. Although he only accomplished this by initially 'bugging me' to play in the evenings—he initially brought me toys when I did not want to play—the simple fact is that

he ultimately got me to go along with it, and we now have a kind of implicit 'arrangement' to play in the evenings when he brings a toy. Indeed, perhaps more interesting still, he even seems to tailor his interests in how long we play to my decisions, as he does not continue to bring me toys after I have played with him for a bit and chosen to put his toy in an open box. Similarly, our children often 'test' us, seeing which kinds of behaviors they can 'get away with without upsetting mom and dad.' Although these are admittedly crude and highly implicit forms of 'interest negotiation', they still seem to be just that. The dog is looking to determine when it is in their interest to bring a toy to their owner to play with, by (if only tacitly) reading when their owner is interested in playing. Similarly, the child is looking to discover which of their interests they can advance without upsetting their parents (and, of course, when they do so poorly, or attempt to pursue interests regardless of what mom and dad think—engaging in actions their parents take to be misbehaving—they often face consequences, such as a 'time out' in the corner, thereby incentivizing them to 'better negotiate' with mom and dad in the future).

Here is why all of this is important. Consider again what the parties to the Moral Original Position know. First, they know that every sentient being has their own freedom from coercion as a higher-order end (an end applying to all of their first-order ends). Second, they know that they have an all-things-equal higher-order interest in rendering their ends more consistent with those of others, to the extent that doing so is in their voluntary control—avoiding coercion. Third, they know in some cases, coercion is unavoidable (if, for instance, a murderer wants to take my life, I cannot simply 'voluntarily' decide I want to die: I will find myself impelled to want to live). These three claims directly entail that, from the standpoint of the Moral Original Position, all costs aside, one should

voluntarily choose the Principle of Negative Fairness: one should have the avoidance and minimization of coercion, in all of its forms, as an ideal.

1.2 The Principle of Positive Fairness

Now let us consider the flip side of coercion: assistance in pursuing one's ends. I just argued that whenever a being has an interest in something, X, they thereby have an instrumentally rational interest in not being coercively prevented from achieving its object, X. Does the flip-side follow: namely, that anytime someone has an interest in something, they thereby have an instrumentally rational interest in other people or beings helping them achieve it? The answer is no. If I want X and can best achieve X without any assistance, then it is instrumentally rational for me to want not to be assisted by anyone in achieving X. The only time a being has an instrumentally rational interest in being assisted in their ends is when assistance would better enable them to pursue their ends than they can without assistance.

As such, every being in the Moral Original Position shares this higher order rational interest: an interest in being assisted in achieving their ends when, and only when, assistance would better enable them to do so than they can do on their own. Taken all by itself, one might think that this shared interest makes it instrumentally rational for the parties to the Moral Original Position to agree to an assistance-maximizing principle: namely, a principle of maximize the total amount of assistance in the world afforded to human and nonhuman beings in the achievement of ends they cannot best achieve on their own. However, there are several complications that undermine the rationality of such a principle.

First, agreeing to assist others in achieving their ends may impose costs on us. For instance, suppose you want affordable health care, and a publicly-funded system of health

care would better enable you to achieve that end than you would on your own (say, in a free-market system). Although my helping you (for instance, by paying additional taxes to contribute to the funding of such a system) might indeed better enable you to achieve your end, it might cost me in terms of my involuntary and semivoluntary interests. I may find myself not wanting to fund a public health care system, wanting instead to keep my money to myself. Although I could choose to voluntarily change my interests, doing so might be far from cost-free for me. The parties to the Moral Original Position—if they are to choose rationally—need to be sensitive to these costs. Their task is to arrive at a universal agreement on principles of action given the assumption that their interests could be anything, including my interest in not helping you achieve what you want.

Second, agreeing to an assistance-maximizing principle is inherently incompatible with another kind of interest many of us have: an interest in negotiating our terms of interactions with others, or the extent to which we have an interest in helping others. For instance, although some of us may come to the table with interests in not helping others (one person may not want to fund the public health care system you want), others of us encounter this very issue with initial ambivalence: we do not have any clear ends prior to negotiation or dialogue over the extent to which we want to assist others in their ends. Indeed, this may be the case on 'both ends' of the issue. For instance, if you tell me you want publicly funded health care, I may not initially have any interests one way or the other. I may instead listen to what you have to say before I form an opinion or interest one way or the other (if you convince me that I should help you, I may form an interest in helping you; but if you fail to convince me, I may form an interest in not doing so). Similarly, the person putatively in need of assistance may form higher-interests regarding

their own assistance as a result of human interaction. For instance, suppose you initially want health care, and it turns out that a publicly-funded system—one assisting you—would be the best instrument for you to get it. However, in conversation with me, I convince you that this would impose undue costs on me: that I would have to pay 'too much' for your health care, which would cost me suffering. This might lead you to revise your interests: you may still want health care, but now want *not* to be assisted in obtaining it due to the costs that assistance would impose upon others. Since this higher-order interest you develop in interacting with me modifies your first-order interest—whereas before you just wanted health care however you could best get it (with or without assistance), now, following our conversation, you may only want health care subject to a further motivational interest: the interest of others not assisting you to get it.

For these reasons, the parties to the Moral Original Position should agree to a principle of assistance, but one qualified by the nuances just discussed, namely:

> **The Principle of Positive Fairness:** all of our morally relevant actions should have as a guiding ideal, setting all costs aside, assisting all human and nonhuman sentient beings in achieving interests they cannot best achieve on their own and want assistance in achieving, for its own sake.

1.3 The Principle of Fair Negotiation

The first two principles we have arrived at from the Moral Original Position—the Principles of Negative and Positive Fairness—have been based on the assumption that non-coercion and assistance of a certain sort are things that every human and nonhuman sentient being has interests in, setting all costs aside. Accordingly, these two principles should have the status of what we might call 'regulative ideals': they are principles that

every one of us should have, setting all costs aside, in any morally relevant action. However, by restricting the argument in this way—by setting all costs aside—we have set aside two questions: (1) which types of actions these ideals should factor into, once costs are brought into the picture, and (2) how the two principles are to be balanced or weighed against one another, once costs are brought into the picture. Allow me to more fully explain each of these questions before resolving them.

Consider one central clause in the Categorical-Instrumental Imperative, which is then rephrased in new terms in its Moral Original Position formulation: the clause that whether an action is 'relevant'—whether it is one that we should try reach a universal agreement on with all of our possible future selves—is itself a higher-order question that we rationally ought to settle with our future selves through a higher-order universal agreement. We saw in Chapter 3 that it is far from obvious that all of our actions should be considered 'relevant' (or more precisely, now that we have seen the Categorical-Instrumental Imperative to be a moral principle, morally relevant). For as we saw in Chapter 3, it may actually turn out in certain cases where we confront the Problem of Possible Future Selves—wanting to know our future selves' interests—that some of the future selves we might turn out to be have interests in us not solving the problem, due to the costs that our confronting and solving it might have on them. For instance, consider again Stocker's case of visiting a sick friend in the hospital. Suppose that, unlike most people—who simply rush off to the hospital to see the sick friend—I pause to consider whether rushing off to the hospital would satisfy my future self (that is, I want to know whether going to the hospital will satisfy my future self's interest). My doing this very thing—pausing to question whether I should go to the hospital (because I feel uncertain

about what my future self will want)—may itself impose costs on my future self that he does not want to face. First, my future self may be disappointed in me: in the fact that I am the kind of person who would even pause to think about whether I should rush to the hospital. Second, suppose my friend is gravely ill, and the few moments I spend contemplating whether going to the hospital would satisfy my future self cause me to get to the hospital when it is too late, just minutes after my friend has passed away. This might cause my future self immense regret. Consequently, given the character of our arguments so far—the fact that the Principles of Negative and Positive Fairness have been derived by setting all costs aside—we cannot validly assume that these principles should in fact motivate us in all cases. We need to arrive at a universal agreement from the Moral Original Position on the higher-order question of when, and to what extent, these two principles should motivate us at all, costs included. In other words, we need to arrive at a universal agreement on which actions are morally relevant—which actions the Principles of Negative and Positive Fairness apply to, given the costs of being motivated by them at all. Thus, on my account, morality comprises a higher-order regulative ideal of using the Moral Original Position to determine which of our actions we should subject to first-order moral deliberation and moral ideals.

Second, insofar as our arguments for the Principles of Negative and Positive Fairness set all costs aside, we presently have no analysis of how these two regulative ideals should be balanced or weighed against one another or costs when cases are morally relevant. For instance, the Principles of Negative and Positive Fairness may conflict with one another in such a way that it is impossible to pursue one without imposing costs vis-à-vis the other. For instance, it may turn out that that the best way to assist people or

nonhuman sentient beings (in line with the Principle of Positive Fairness) is to coerce individuals. An example here may be universal, government-funded health care. If many citizens cannot afford health care for themselves and want assistance in being able to afford it, and the best method to provide these citizens with health care is to coercively tax all citizens in order to provide it to them, the Principles of Negative and Positive Fairness come into a kind of conflict. The Principle of Negative Fairness says we should aim to reduce coercion in the world, setting all costs aside, yet the Principle of Positive Fairness says we should aim to assist others (in a certain way), setting all costs aside. In this case, however, we cannot do either without some cost to the other: if we fail to coerce people (through taxation), we fail to assist others (those who desire government-funded health-care); yet if we do assist others, we coerce people (those who wish not to be taxed). Because pursuing either principle in this case imposes costs on people vis-à-vis the other, we need an analysis of whether, and how, to pursue the principles—balancing or weighing them against each other—given such conflicts and the costs of resolving such conflicts one way rather than another.

Let us examine, then, how the parties to the Moral Original Position should deliberate about costs. As we saw in Chapter 3, there are several possible types of costs that human and nonhuman sentient beings can have. First, involuntary interests generate possible costs: if I find myself angry at someone, wanting to lash out at them and say mean-spirited things, then not allowing myself to indulge my anger is a cost to me—one that, if my aim is thwarted, I cannot avoid or mitigate (I may find myself frustrated). Second, involuntary interests also generate a different type of cost: costs that can be mitigated. For instance, if I find myself angry and wanting to lash out at someone but actively work to

control my anger, making myself less angry and wanting to lash out at them less than before, not allowing myself to lash out will still impose a cost on me, but a lesser one—due to my own choice to control my anger (within the constraints allowed by my psychology)—than in the involuntary case. Finally, there are costs with respect to fully voluntary interests, ones we first-personally experience when we ourselves make choices. If I choose not to want to help pay for other people's health care—if I judge to myself, 'I ought not to have to pay for other people's health', and will myself to act on this judgment—then, if I am forced to do what I do not want, I face a cost: a cost that is the partial result of that particular voluntary choice.

Recall, as we saw in Chapter 2, that although typical adult human beings often (if not always) appear to have all three types of interests, other types of beings—nonhuman animals, psychopaths, and children—are arguably incapable of the same full range of interests. Most animals, in particular, do not appear to share our first-personal capacities for voluntary choice: the capacity to experience oneself as choosing to act on a normative judgment ('I ought to tell the truth'). Instead, most (if not all) animals appear to be impelled by involuntary and perhaps semi-voluntary motivations. If the dog wants to go outside, he will sit by the door; if he wants to come inside, he will sit outside looking in—but it does not appear that he ever thinks about whether he should want to go out or come in.

Since the Moral Original Position requires us to treat ourselves—moral agents—as though we could turn out to have the interests of any human being(s) and any nonhuman sentient being(s), our deliberations concerning costs should be sensitive to these differences. Insofar as the Moral Original Position's veil of ignorance requires the parties to it (you, I, and any other moral agent) to treat ourselves as though they could 'turn out to be

anyone', including nonmoral agents (such as animals), it is instrumentally rational for any universal agreement on which actions are 'morally relevant' given costs—and how the Principles of Negative and Positive Fairness should be weighed or balanced given costs—to be based on the assumption that (1) involuntary interests of human and nonhuman sentient beings entail given costs (costs which automatically and unavoidably accrue given certain states of affairs, particularly those that are contrary to the involuntary interest in question), (2) semivoluntary interests entail partially modifiable costs (costs that can be altered within some given psychological bounds), and (3) fully voluntary interests entail fully modifiable costs (costs determined by the agent's own choices).

Therefore, although some of our actions involve beings (animals) who may only have involuntary interests, insofar as we are moral agents engaging in actions that have effects on them, all of our actions involve at least some beings (namely, we who are acting) who are capable of semivoluntary interests and voluntary interests. In other words, all of our actions that fall under the Moral Original Position involve some beings (us) for whom the costs of their actions may not be fully given before acting: we can make voluntary and semivoluntary choices concerning what comprises costs for us, and how costly we experience different actions and events. For instance, if one initially wants some number of cookies from a box, one can nevertheless choose to give more to another person, *deciding* not to consider it such a big sacrifice. Similarly, if someone does something to make one angry, one typically has some amount of (semivoluntary) control over how angry one gets. One can let one's anger rage out of control, increasing the costs of the person's behavior on oneself, or one can work to control one's anger, controlling how much the person's

behavior upsets you (thus mitigating, at least to some extent, the extent to which one suffers from their actions).

For reasons just given, the parties to the Moral Original Position—moral agents such as you and I, considering how to treat ourselves and other possible moral and nonmoral agents (agents whose interests our possible future selves could identify as their own)—cannot assume, in attempting to reach a universal agreement on how to treat costs, that costs for many of the possible agents they could be (agents with voluntary and semivoluntary interests) are necessarily fixed prior to those agents' actions. Interestingly, this also means that the parties to the Moral Original Position cannot treat the costs that nonmoral agents (animals) may experience due to only having involuntary interests as settled either: for the costs that an agent with only involuntary interests (an animal) experience depends on the actions of moral agents who do have voluntary and semi-voluntary control over their own interests. For instance, although the chicken or cow may both have involuntary interests in living, and living without suffering—interests they cannot choose not to have—we may choose not to impose certain costs on them at some cost to ourselves, choosing not to kill or harm them, given our recognition of their involuntary interests. As such, the parties to the Moral Original Position—moral agents deliberating to principles of how to treat other moral and nonmoral agents—cannot regard the costs that anyone (moral or nonmoral agents alike) will face as a result of their actions as necessarily settled in advance, prior to anyone acting voluntarily or modifying their semivoluntary interests.

This has a very important implication. Given that we experience our voluntary and semivoluntary interests as unsettled prior to our acting, the parties to the Moral Original

Position cannot agree to any distributive principle to decide—'*a priori*' as it were—which costs moral agents should take to determine whether an action is 'morally relevant' (vis-à-vis the Categorical-Instrumental Imperative or Principles of Negative and Positive Fairness), nor which costs should pertain to the application of the Principles of Negative and Positive Fairness in cases that are morally relevant. The parties to the Moral Original Position cannot rationally agree, for instance, that the costs of determining which actions are 'morally relevant' or the costs of weighing the Principles of Negative and Positive Fairness should be spread equally across human and nonhuman sentient beings, or unequally but to the maximum advantage of the worst-off individuals (as in Rawls' theory of justice), and so on. Again, this is for the simple reason that many of the individuals the parties to the Moral Original Position could turn out to be—moral agents with semi-voluntary and fully voluntary interests—may have interests in how costs are distributed which, because they are semi-voluntary and voluntary, are not decided before their choices have been made. One cannot agree to any particular distribution of costs if, given one's deliberative situation, there is no fact of the matter on how the individuals you are reasoning about would like those costs to be distributed. It could well turn out that every individual with voluntary interests would want the costs to be distributed equally. However, it could also turn out (in principle) that they would all like costs to be distributed in some or indeed any other possible way (to the maximum advantage of those facing the most costs, to the advantage of the rich, and so on).

There is a simpler way to put this. Insofar as we (moral agents) have voluntary and semi-voluntary capacities—capacities to choose which costs we have an interest in facing (voluntary interests), or to modify the costs we are willing to face (semivoluntary

interests)—and, by definition, have motivational interests in exercising these capacities (to make choices is to be motivated to make choices), the parties to the Moral Original Position have instrumentally rational grounds to agree to a principle of fair negotiation. Such a principle enables moral agents to negotiate with one another, and (in a manner of speaking) with nonmoral agents (more on this shortly), the costs they have interests in facing for the sake of defining 'morally relevant actions', and—in cases of actions deemed morally relevant—to weigh the Principles of Negative and Positive Fairness against one another. Furthermore, note that the parties to the Moral Original Position have no clear grounds for favoring any being in such a negotiation: the parties must treat themselves as though they could turn out to be 'anyone', and the very question of whether anyone's interests in costs should be favored over anyone else's is something that they can negotiate. The parties to the Moral Original Position therefore have instrumentally rational grounds for wanting every individual they could turn out to be to have equal bargaining power over the negotiation, so that they have an 'equal shot' of realizing their favored distribution of costs, whatever interests they may turn out to have (even interests in distributing costs one way rather than another).

Of course, in the real world, organic negotiations over costs that afford all parties equal bargaining power are profoundly difficult—if not impossible—to achieve. Even in small groups, some parties to negotiations typically have greater bargaining power than others (due to things like intimidation, confidence, money, and so on). The parties to the Moral Original Position should surely know that potentially unequal negotiating power is a fact of life to be grappled with. Since the parties cannot rationally agree to any particular distribution of costs (given our interests in exercising our voluntary and semivoluntary

interests), and should rationally favor a process of negotiation (one that enables us to exercise our voluntary and semivoluntary interests concerning the costs we are willing to take on in our actions, given their effects on other human and nonhuman sentient beings), the parties should agree that such a negotiation process should aim to approximate one that affords equal bargaining power to all those affected, as far as it is possible to do so. This is of course an 'imperfect' solution—but I hold that it is the only one the parties to the Moral Original Position can rationally agree to, given their situation and knowledge that fully equal bargaining power is difficult (and often impossible) to achieve. And though admittedly imperfect, I believe that it comports well with commonsense moral convictions about how negotiations should be. For instance, it sits well with the conviction—common in liberal-democracies today—that although democracy is far from perfect, the more equal people's negotiating power is in a democracy (the less, say, the rich determine policy or who is elected, and the more the people do), the morally better it is. Given that we prefer negotiations to be fair, the best that the parties to the Moral Original Position can do in light of real-world differences in negotiating power is to aim to approximate as fair of a negotiation process as possible.

Next, the parties to the Moral Original Position should know that some affected by our actions—some 'parties to the negotiation' in terms of experiencing costs—cannot actually negotiate. Nonhuman animals, for instance, cannot negotiate solutions with us: we can only 'include them' in the negotiations by proxy (by attempting to discern their interests and give their interests equal bargaining power in the process). Since the parties to the Moral Original Position are concerned with these types of beings (the interests of nonhuman animals can turn out to be our own interests, even if unlikely), the parties

should agree upon a principle of fair negotiation that affords these beings' interests equal bargaining power in the process as well.

Finally, as we will see in more detail in §1.4, there is another principle the parties to the Moral Original Position should want to incorporate in their negotiation: a Principle of the Virtues of Fairness, which requires developing and expressing dispositions that facilitate pursuit of the Principles of Negative and Positive Fairness, and the Principle of Fair Negotiation, in the process of negotiation itself. For it is only to the extent that such a negotiation process is based on dispositions consistent with the principles it embodies—the Principles of Negative and Positive Fairness, and the Principle of Fair Negotiation—that entire negotiation process is truly motivated by the principles it is intended to be motivated by.

For these reasons, it is rational for the parties to the Moral Original Position to agree to the following principle:

> **The Principle of Fair Negotiation:** whether an action is morally relevant, and how the Principles of Negative and Positive Fairness and Virtues of Fairness (see below) should be applied factoring in costs, should be settled through an actual process of fair negotiation guided by the Principles of Negative and Positive Fairness and Virtues of Fairness, where all human and non-human sentient beings affected by the action are afforded equal bargaining power to the extent that such a process can be approximated, and to the extent that cannot be, through a hypothetical process approximating the same, for its own sake.

1.4 The Principle of Virtues of Fairness

There is a fourth and straightforward principle that is rational for the parties in the Moral Original Position to agree upon. Given that it is rational to agree to the Principles of Negative and Positive Fairness and the Principle of Fair Negotiation in the Moral Original Position, it is also rational to develop and express stable character traits—or psychobehavioral dispositions—to apply and act in conformity with the first three principles of fairness. After all, such traits simply are the disposition of being motivated to apply and act in accordance with three principles of fairness—which, as we have just seen, are rational to agree upon from the standpoint of the Moral Original Position. It is therefore, by definition, instrumentally rational to prefer oneself to be disposed to apply and act according to the principles that one should be motivated by. Thus, we have:

> **The Principle of Virtues of Fairness:** all of our morally relevant actions should aim to develop and express stable character traits to act in accordance with and on the outcomes generated by the first three principles of fairness, for its own sake.

This principle enables us to resolve a question that I suspect has been in the back of many readers' minds for some time. In developing the Problem of Possible Future Selves in Chapter 2—the problem I later argued is solved by the Categorical-Instrumental Imperative—I began with the observation that we arguably only encounter the Problem on some, but not all, occasions. I held that in some cases, we simply act without thinking, and it may only be in cases of uncertainty about the future (including moral uncertainty) that we encounter the Problem at all (which, again, is wanting to know our future selves' interests). One thing that may have troubled readers about this argument, and my contention that the Categorical-Instrumental Imperative is a solution to the Problem, is that it makes morality

seem arbitrary in a certain sense: that moral questions only arise, and moral principles only apply, when we in fact encounter the Problem of Possible Selves. What if, one may ask, one encounters that problem only rarely, or in different instances than other individuals? Does this not make morality itself completely relative regarding whether, and when, each individual encounters the Problem of Possible Selves?

The Principle of the Virtues of Fairness enables us to resolve this concern in an intuitive fashion. Insofar as we all encounter the Problem of Possible Future Selves at least sometimes in our lives (as Chapter 2 argued), the Four Principles of Fairness entail that we should apply the Four Principles in those cases—negotiating what is fair with others—in ways that lead us to develop dispositions to behave fairly in a similar way in the future. In other words, the Four Principles of Fairness require us, in all of our actions, to develop dispositions to be fair to present and future selves and others, encountering the Problem of Possible Future Selves, and solving it, *when it is fair to ourselves and others to do so*. But this is a commonsense idea. It simply means that morality it itself a matter of negotiating with others what kind of people we should become, and which of our actions we should consider to be morally relevant—which, I would argue, is exactly what we do in relationships, in the workplace, and in society at large.

Consider, for instance, changing social mores concerning sensitivity. Several decades ago, certain uses of language and ways of speaking—use of racially insensitive language (referring to people of certain racial/ethnic backgrounds as 'colored'), gender stereotypical language (using 'he' as a default pronoun in written language), and casual use of crude language concerning the physically and mentally disabled (words such as 'retarded') were not considered moral issues, and a person who engaged in these types of

behaviors was not considered to lack moral virtue. This was almost certainly because, given social inequalities at the time, members of the affected populations (those who find the above language hurtful or demeaning) had not yet negotiated standards of language sensitivity with the rest of society. Insofar as the Principle of Virtues of Fairness draws on the first three principles—including the Principle of Fair Negotiation—it enables us to understand moral virtue and moral relevance (the kinds of cases we should be disposed to apply the first principles to) as being determined in an ongoing, organic fashion by social negotiation: something which is intuitive, since it is in fact what we do.

Finally, the Principle of Virtues of Fairness enables us to explain how and why moral relevance and virtue can be context sensitive, and indeed, relative to individuals and relationships within certain bounds. Since my spouse is directly affected by household habits and other actions concerning her—and we both bear different costs as a result of different types of behavior on the part of the other (she desires me to do certain things that I may find irksome, and vice versa)—the first three principles of Rightness as Fairness, and by extension the Principle of Virtues of Fairness, entail that moral virtue and moral relevance in our relationship are to be defined by us, in fair negotiation with one another. Rightness as Fairness thus enables us to make sense of the widely (if only tacitly) recognized fact that what is 'morally relevant' or virtuous in one relationship may not necessary be in another.

2 Rightness as Fairness: A Unified Standard of Right and Wrong

Given that it is instrumentally rational for the parties to the Moral Original Position (you, I, and every other moral agent) to universally agree to the Four Principles of Fairness, it is

instrumentally rational for the parties to universally agree to analyze moral rightness in terms of their conjunction:

> **Rightness as Fairness:** an action is morally right if and only if it satisfies the Principles of Rightness as Fairness, that is, if and only if it is (A) is morally relevant, (B) has coercion-avoidance and minimization, assisting human and non-human sentient beings to achieve interests they cannot best achieve on their own and want assistance in achieving, and the development and expression of settled dispositions to have these ends, as at least tacit ideals, and (C) is in conformity with the outcome of an actual process of fair negotiation approximating all human and sentient beings affected by the action being motivated by the above ideals and having equal bargaining power over how those ideals should be applied factoring in costs, or, if such a process is impossible, the outcome of a hypothetical process approximating the same, where moral relevance is determined recursively, by applying (B) and (C) to the question of whether the action is morally relevant.

We can then define other deontic notions—such as moral wrongness, permissibility, indeterminacy, and the supererogatory—in a similar fashion. An action is morally wrong if and only if it is morally relevant but violates conditions (B) and/or (C) above. An action is morally permissible—that is, neither morally required nor forbidden—if and only if it is not 'morally relevant' (since morally irrelevant actions are neither required nor prohibited by morality) or is morally relevant but negotiated to be not required. An action is supererogatory (or 'above and beyond what it is required') if and only if it is morally right to perform at some cost to oneself (whatever costs are negotiated *qua* Rightness as Fairness), but one performs it at *greater* cost to oneself than required. Finally, an action has

indeterminate moral status—there is no fact of the matter of its being right, wrong, or permissible—if and only if negotiation about its moral relevance and/or costs has not occurred, and there are multiple possible, conflicting outcomes of fair negotiation consistent with the ideals of Negative and Positive Fairness.

Before applying Rightness as Fairness to several cases to illustrate its analysis of moral rightness and moral problem-solving, I want to pause to reflect on some of its unique features.

First, Rightness as Fairness is unique in holding that morality itself is partly a matter of negotiating with other people and nonhuman sentient beings which of our actions are morally relevant. I believe this to be a very important implication, as there are several related concerns that modern moral philosophy 'overmoralizes' life, wrongly turning all of our actions in moral issues. First, Michael Stocker, Bernard Williams, and others have argued that modern moral philosophy requires 'one thought too many', requiring us to always act (at least implicitly) for moral reasons when, intuitively, many of our actions should be motivated by nothing more than love, friendship, or sympathy.[6-9] A second, related, critique is that modern moral theories require us to subsume all of our life projects to morality, requiring us to be 'moral saints', concerned with morality above all else.[6,9-10] As Susan Wolf writes, 'One attractive ideal of love would prohibit the lover not only from thinking about morality all the time, but also from being unconditionally committed to acting according to morality all the time.'[11] A third, related critique—raised typically in relation to utilitarianism, but arguably applicable to other theories as well—is that modern moral philosophy requires too much of us, demanding extreme forms of impartial concern for others.[6,9-13] For instance, classical act-utilitarianism holds that morality requires all of

our actions to maximize happiness in the aggregate, rule-utilitarianism holds that all of our actions should conform to rules that maximize happiness, and so on.[14] Yet, as many utilitarians (such as Peter Singer) have argued, maximizing happiness—either by act or by rule—may require an incredible amount of us, including giving up most of our wealth to alleviate world poverty and killing handicapped infants.[15-16] Similarly, traditional Kantian ethics requires us to always act on maxims we could will to be universal laws of nature. Yet, as Stocker points out, visiting a loved one in the hospital 'because it can be willed as a universal law of nature' seems like an overly moralized reason for acting: one should visit loved ones in the hospital simply because one loves them.[6]

While utilitarians[17], Kantians[18-19], and moral philosophers of other persuasions[20] have responded to these types of concerns, I believe Rightness as Fairness provides a more intuitive solution, holding that morality itself is fundamentally a matter of negotiating with others, in a manner guided by moral principles (the Principles of Negative and Positive Fairness), which of our actions are morally relevant, and as such, how 'demanding' morality is. This is a compelling implication for a couple of related reasons. First, real-life moral practice strongly suggests that this is exactly what we do: we negotiate, in relationships, in society, and the world more broadly, which things count as moral issues, and how demanding morality is. For instance, in marriages, one typically 'works out' with one's spouse a mutual understanding of which actions are moral issues in the context of the marriage. For instance, whereas neither my spouse nor I regards what time we eat lunch as a moral issue in the marriage—neither of us has much of an interest in what the other does—we have negotiated other things as moral issues, such as what time we go to bed. This became a moral issue for us because going to bed early is important to me and going to

bed later is important for her (my wife is a night owl and prefers to work late), and we found that we disturbed each other's sleep when we went to bed and woke up at different times. We thus experienced a conflict of interests, and came to see bedtime as a question of what is fair between us. Second, insofar as morality has costs (as we have already seen)—insofar as helping you (in line with the Principle of Positive Fairness) may impose costs upon me—Rightness as Fairness provides an elegant explanation for something that has puzzled moral philosophers. The puzzle is this: why, although we commonly recognize that there may be some sense in which we 'should' be moral saints (putting morality first, in a 'Christ-like' manner), there is also a sense in which most of us are content (morally speaking) with not being moral saints. As Eric Schwitzgabel puts it, 'it's generally true that we aim for [moral] goodness only by relative, rather than absolute standards'—that we aim, as it were, for only a grade of 'B+ on the great moral curve' rather than the 'A' grade of the moral saint (such as Buddha, Gandhi, Jesus Christ, and so on).[21] So should we be 'moral saints', or not? Rightness as Fairness provides a nuanced answer. Insofar as the Principles of Negative and Positive Fairness affirm certain ideals—coercion-minimization and assisting people who would benefit from and desire help—as moral ideals to be pursued, all costs aside, Rightness as Fairness entails that it may be right for someone to be a 'moral saint' such as Buddha, Gandhi, or Christ. Thus, it may be right for someone to be willing to endure immense personal costs for those ideals, provided they are also sensitive to and fairly negotiate with others the costs of their doing so. However, Rightness as Fairness also entails that those of us who are not willing to endure the costs of the moral saint have every right to negotiate with others the costs that we should have to face in pursuing the same moral ideals—and, if we fairly negotiate 'less saintly' moral standards, Rightness as

Fairness entails that it is fair and right for us not to be moral saints. But this is precisely what critics of existing moral theories have long suggested: that what is 'right for the moral saint' need not be right for everyone. We will see the attractiveness of this line of thought in greater detail in §3, when we apply Rightness as Fairness to specific cases.

Second, Rightness as Fairness introduces a novel method of moral problem-solving that requires us to at least partially abandon a common, seductive, and (I believe) problematic conception of how to approach applied moral issues. Many people (including philosophers) are naturally drawn to the notion that moral issues can be properly addressed through thought and debate: that we can 'think through' sound answers to applied moral questions. To illustrate, there are countless books and articles arguing for and against the notion that it is morally right for the rights or interests of the many to outweigh the rights or interests of the few (and if so, when)[22-5], whether it is right to direct a trolley to kill one person in order to save five others[25-7], whether it can ever be right to torture a person[28-33], and so on. At the same time, however, the idea that applied moral issues can be settled through thought and debate is problematic. First, as we see in the applied ethics literature on the topics just listed, people on different sides of the issues find different argumentative premises attractive, and different moral theories often lead to quite different conclusions (what produces the most utility, *qua* utilitarianism, may not respect human autonomy, *qua* Kantianism, and so on). Consequently, 'principled debate' all too often results in argumentative 'stand-offs': situations in which people fundamentally differ on the premises they find attractive and arguments they find compelling. We see this, for instance, in the debates mentioned above. When it comes to whether, and when, the rights or interests of the many outweigh the few—of whether it is morally permissible to

push a person in front of a trolley to save more lives, or torture suspected terrorists to protect large numbers of people from possible terrorist attacks—there are usually plausible arguments on 'all sides' of the issue. And though some arguments may be better than others, the issue of disputed premises often remains. Whereas some people may find utilitarian analyses of the moral permissibility of torture attractive, others may be staunchly Kantian in outlook, finding utilitarian premises flawed (and vice versa). This is a deep problem indeed. For when people disagree over premises, it is unclear how a productive argument can proceed (if you and I cannot even agree on 'moral starting points', how can either of us say anything likely to convince the other?).[34] Second, this problem often becomes particularly acute in public debate. When it comes to just about any contentious moral issue—abortion, gay marriage, and so on—there is often a pronounced unwillingness of opposing sides to engage in 'debate' with the opposing side, precisely because of apparent 'fundamental differences' over premises. This is not only a practical problem: it is arguably a moral one, as an unwillingness of people to listen to one another often (if not always) seems to result in greater conflict, fomenting divisiveness rather than leading to productive resolution of the relevant issues.

According to Rightness as Fairness, the idea that applied moral issues can be soundly addressed through principled thought and debate alone is fundamentally in error. Although Rightness as Fairness stipulates that morality is partly a principled affair (specifically, that we can and should debate which sorts of actions or policies are most in line with moral ideals of coercion-avoidance and minimization, as well as helping others), it maintains that morality is also something that cannot be wholly settled 'on principle' or through mere debate. Instead, Rightness as Fairness holds that morality is fundamentally a

matter of negotiating with others the costs that we, and they, should face for the sake of the aforementioned ideals. Rightness as Fairness thus entails that while there is indeed value in debating whether abortion, the use of torture in the 'war on terror', and gun-control are more consistent with the ideals of Negative and Positive Fairness than their opposite, the ultimate answer to these questions cannot be settled on principled grounds alone. Rather, since whichever 'answer' we arrive at will impose costs on people—pro- and anti- abortion, gun-control, and torture policies all impose different costs on people—Rightness as Fairness holds that morality requires us to negotiate those costs with one another: negotiate, that is, a fair balance of moral ideals against the costs of pursuing them in one way rather than another. Therefore, once we have debated ideals—which sorts of policies are the most consistent with the moral ideals expressed by the Principles of Negative and Positive Fairness—there can be no 'principled answer' as to what the right moral answer is in the case at hand (abortion policy, gun control, torture, and so on) is. Rather, Rightness as Fairness holds that the right answer must be created by fair negotiation: by an actual, organic process that enables all affected to weigh moral ideals (Negative and Positive Fairness) against the costs of different modes of implementation.

In one respect, this is entirely intuitive. When we have conflicts, say, during a project at work—where not everyone can 'get their way'—we tend to think that there is no 'principled answer' as to the 'right way' to resolve the conflict. Rather, we typically think it is right to resolve the conflict through a process that gives everyone a fair chance to speak and vote for their favored solution. Indeed, this very notion seems to underlie the project of modern democracies: namely, that when we (legitimately) disagree over 'what's right' (and I will say more about how to understand 'legitimacy' here shortly)—when opposing parties

both have legitimate (in their view) principles in mind, but disagree over how they should be balanced against each other, and against costs—the answer is to forge a fair solution, where citizens negotiate on an ongoing basis the right answer to the issue. Rightness as Fairness entails that this democratic notion is a fundamental part of morality itself: that morality is not a matter of 'finding out' what maximizes utility, or respects human autonomy—things that can be written in books or articles, or debated in words—but rather a matter of real, live people affected by actions on moral issues (people whose lives are at issue when it comes to abortion, torture, and so on) (1) being motivated by certain ideals (the Principles of Negative and Positive Fairness), (2) negotiating the proper balance of those ideals, and balance against costs, with others who are similarly affected and motivated, and (3) forging fair resolutions together not through mere words or debate, but through negotiation, or fair bargaining. I believe this is intuitive, since it is commonsense that 'conflicts require fair resolutions.' Furthermore, only Rightness as Fairness puts this notion center-stage, holding that morality is fundamentally a matter of negotiating how certain ideals (the Principles of Negative and Positive Fairness) should be applied to cases given costs and conflicts thereof. I therefore believe that Rightness as Fairness promises a new, productive vision of how to relate to each other than many moral debates presuppose. For although people have a certain tendency to 'stand on principle', both in philosophy and in real life—asserting, for instance, that abortion or torture is 'right' or 'wrong', *simpliciter*, without any willingness to negotiate—Rightness as Fairness holds that an unwillingness to negotiate is itself morally wrong among people who share the Principles of Negative and Positive Fairness as ideals (since it is contrary to the Principle of Fair Negotiation). Rightness as Fairness holds that there is only one situation in which it is morally right to

stand on principle: cases of morally *illegitimate* disagreement, where one's moral 'opponent' is motivated by incorrect moral ideals (such as the slaveowner or racist, who are unwilling to extend the Principles of Negative and Positive Fairness, or Fair Negotiation, to entire classes of people).

A third (and related) notable feature of Rightness as Fairness is that it merges the insights of several leading moral frameworks—deontology, consequentialism, virtue ethics, and contractualism. The Categorical-Instrumental Imperative is broadly deontological, requiring all of our morally relevant actions to conform to a universal agreement with all of our possible selves for its own sake. The Principles of Negative and Positive Fairness, which we are to pursue for their own sake, are broadly consequentialist in content, requiring us to aim to bring about certain consequences (all things being equal, setting costs aside): namely, minimizing coercion (Negative Fairness) and assisting human and nonhuman beings achieve their ends under certain conditions (Positive Fairness). Next, the Principle of Fair Negotiation is heavily contractualist, holding that we must apply the first two principles via fair negotiation with others. And finally, the Principle of Virtues of Fairness is of course virtue ethical in nature, requiring us to develop and express certain stable character traits.

A final important property of Rightness as Fairness is that it provides a unique and (I believe) compelling analysis of why it is rational to obey moral norms—an analysis that, insofar as it engages with our motivational interests, can actually motivate people to behave morally. As we saw in Chapter 3, Rightness as Fairness is based on concerns that we all have about our future from time to time—concerns that require us to be fair to all of our possible future selves. Furthermore, as we have seen, many empirical results appear to

broadly confirm this account, strongly linking imprudent and immoral behavior to failure to be concerned for one's future[35-8], and improved moral behavior to stimulation of concern for one's future self.[39-41]

I believe that all of these are compelling features in favor of Rightness as Fairness, and we can see their practical utility by briefly applying the theory to some controversial moral issues.

3 Rightness as Fairness in Practice: Principled Fair Negotiation

As we saw in Chapter 1, a compelling theory should be fruitful, solving theoretical and practical problems better than alternatives.

Existing moral theories, by and large, arguably run into one of two problems. On the one hand, 'monistic' moral theories—such as utilitarianism or Kantianism—are often criticized for the fact that they attempt to reduce morality to a simple 'formula': a formula of maximizing utility, respecting autonomy, and so on. For instance, ordinary act-utilitarianism is often alleged to entail overly simplistic, implausible analyses of applied cases, requiring us to simply 'add up' utility and pursue whichever action produces the best consequences.[42] Conversely, Kant's moral theory entails that morality is fundamentally a matter of determining which of one's maxims are 'universalizable' or 'respect humanity'—something which, at least according to Kant, has nothing to do with an action's consequences.[43] And theories that attempt to reduce all of morality to 'one thing,' such as consequences (per utilitarianism) or principled intentions (per Kant's theory), seem too simplistic. After all, in real-life we tend to think that morality is a matter of weighing competing considerations against one another—that consequences should matter in some cases, but perhaps not in others. Indeed, many alternative moral frameworks—W.D. Ross'

theory of *prima facie* moral duties[44], virtue-ethics[45], moral particularism[46], and so on—have been developed to avoid charges of 'oversimplifying' the moral domain. Yet these types of theories have been alleged to run into the exact opposite problem: that of not providing enough moral guidance. In Ross' case, it is unclear how we should weigh different duties against one another (something Ross concedes when he writes that we can never know what we morally ought to do, all-things-considered, but can only form 'probable opinions'[47]). There is a similar concern about virtue ethics. While it may be clear what the honest, kind, or helpful thing to do is, in cases where honesty, kindness, or helpfulness conflict with one another, virtue ethics provides no real guidance on how to proceed besides invoking vague (and perhaps circular) notions of 'practical wisdom' on 'what the fully virtuous agent would do'—thus providing no clear analysis on how to weigh or compare the virtues.[48-9] Lastly, moral particularism provides no general principles for moral deliberation, merely instructing us to reason about particular situations on a case-by-case basis. Although proponents of these theories have come to their defense, typically arguing that their theories are appropriately action-guiding[50-2], these worries have not gone away.

Rightness as Fairness, I believe, provides an attractive level of action-guidance. On the one hand, it holds that morality is a matter of pursuing specific principles as ideals—the Principles of Negative and Positive Fairness. On the other hand, the Principle of Fair Negotiation entails that the correct application of these ideals, costs and all, must be settled through organic processes of fair negotiation—or, failing that (if fair negotiation processes are unavailable), by approximating such a process through hypothetical reflection. And as we will now see, this is a picture that fits well with moral practice.

3.1 Kant's Four Cases

In his *Groundwork of the Metaphysics of Morals*, Kant uses both the Universal Law and Humanity Formulations of his Categorical Imperative to argue that it is wrong to make false promises for one's own advantage, commit suicide, never help those in need, and neglect to develop one's natural talents.[53] Because these are four famous examples—ones that Kant's arguments in the *Groundwork* appear to run into famous problems with[54], and which ordinary people have differing pretheoretic intuitions about (some think it is always wrong to lie, others that there are exceptions, and so on)—I believe it is useful to examine them using Rightness as Fairness.

Let us begin with the case of telling a lie (or intentionally 'making a false promise') for one's own advantage. Rightness as Fairness entails that such an action is generally wrong, since lies tend to coerce people (contrary to the Principle of Negative Fairness). However, Rightness as Fairness also holds that precisely when lying is right, wrong, or permissible is something that we need to negotiate with other people, since sometimes lying for one's own advantage can have important benefits for oneself and others. And this is something that we in fact do. Consider, for instance, the social practice (in many cultures or, even more contextually, in certain relationships) of 'making excuses' to avoid uttering an impolite truth, such as saying one 'cannot' meet a friend who invites them to lunch (because one is 'ill' or 'has an appointment') even though the real reason is simply that one does not want to. We have arrived at such norms—in some cases culturally, and in other cases within specific relationships (each relationship, we say, involves its 'own expectations')—because we recognize that in some cases it is fair to lie, given the costs and benefits to everyone involved. We lie to our friend, for instance, because we do not think it

is fair to ourselves or to our friend to tell the truth that we don't feel like getting off the couch and seeing her.

Now consider suicide. Although many of us are unwilling to side with Kant that suicide is always wrong, there are intuitively two dangers with suicide: unfairness to oneself and unfairness to others. On the one hand, if someone commits suicide when their future self might wish they hadn't (if they were to live), then their committing suicide seems unfair to themselves. This, intuitively, is why many people think it is morally permissible to commit suicide only in cases of a terminal disease or some other form of unmitigated suffering which has no reasonable prospect of resolving itself. Further, even in cases where ending one's life might be 'fair to oneself'—if, that is, a person's life contains such little prospect for happiness that their (profoundly unhappy) future self would want them to die—many of us are uncomfortable with suicide on the grounds that it is 'unfair to others': namely, family and friends left behind to suffer the aftermath of the person's act. Here, as with lying, Rightness as Fairness provides a nuanced analysis. Because suicide both runs the risk of depriving a person's future self of a potentially enjoyable future, and also runs the risk of imposing immense costs on family members and others, Rightness as Fairness entails that whether it is permissible or right for a person to commit suicide should be determined through some manner of fair negotiation among all those affected: for instance, by (A) the person informing their family and friends of their thoughts of suicide, (B) allowing them to reason with the suicidal person (giving the family and friends a fair opportunity to convince the person not to go through with it), and finally (C) pursuing counseling for a time, as such counseling might enable the person to more clearly see whether it is possible that they would be better off continuing to live. And these are

things we already tend to think are appropriate when people are suicidal. We think our family members or friends should 'come to us' before going through with such an irreversible act, giving us a fair shot to convince them otherwise, not just for our sake, but because we are also concerned about them being fair to themselves.

Finally, consider Kant's final two cases: the cases of helping people in need and developing one's natural talents. On the one hand, the Principle of Positive Fairness holds that helping people is presumptively right—but the Principle of Fair Negotiation holds that we need to negotiate with others when, to what extent, and at what costs (to ourselves and them), we should help them. Similarly, the Principles of Negative and Positive Fairness both hold that morality requires developing one's talents, as failing to do so is unfair to oneself, putting one's future self in a worse position to successfully pursue their interests (this, broadly speaking, is why we think people 'owe it to themselves' to work hard, study hard, and so on). However, since developing one's talents has costs (as when we say, 'All work and no play makes for a dull life'), Rightness as Fairness holds that we must negotiate a fair balance, with ourselves and others, on the costs we (and they) should bear for developing our talents. This is commonsense as well. We often say that people who never stop to enjoy life are unfair to themselves, and that people who work so hard that they neglect their family, friends, or children are unfair to them.

In sum, Rightness as Fairness coheres with our commonsense beliefs and practices concerning Kant's four examples. It provides nuanced explanations of when, and why, lying and suicide are wrong and when they are not wrong, and when, and to what extent, we have duties to help others and develop our talents.

3.2 How Numbers Should Count: Trolleys, Torture, and Organ Donors

One of the most longstanding problems in all of moral philosophy is whether 'numbers should count'—that is, whether the interests or autonomy of many people should outweigh those of the few—and if so, how.[22-7]

Here is a famous example: a doctor on a transplant ward can save five patients dying from organ failure, but only by covertly killing one innocent, relatively healthy patient.[42] If the doctor could accomplish this action without getting caught, the doctor would save more lives and produce more happiness in the aggregate than by not doing it. Yet although the doctor would save more lives and produce more happiness this way, almost everyone seems to agree that it would be wrong. Doctors, we say, should not kill patients, even to save a larger number of people. In this case, most of us want to say that 'numbers should not count.'

In other cases, however, many of us are inclined to say that numbers should count. Consider so-called 'Trolley-Cases.'[25-7] In one version of the case ('Pull the Switch'), we are to imagine that there is a trolley hurtling down a track, and that it will run over and kill five innocent people unless a switch is pulled to divert it to a second track, where it will kill only one other innocent person. However, in a second version ('Push the Man'), we are to imagine that instead of pulling a switch that will kill one innocent person to save five, the only way to save the five lives is to push a single innocent person in front of the trolley, killing them. Although the numbers in these two cases are exactly the same—either one innocent person will die, or five will die—many of us judge it to be morally right or permissible to pull the switch, but morally wrong to push a person their death.[55]

Finally, consider another, very pressing ethical issue regarding 'whether numbers should count': whether it is ethical to torture suspected terrorists to potentially prevent future terrorist attacks. While some philosophers argue that torture is always morally wrong[28-9], many people argue that it depends on whether torture is likely to save innocent lives, how many lives it is likely to save, and so on.[30-2]

When applied to these types of cases, existing moral theories tend to experience the two problems discussed earlier. On the one hand, some theories seem to give overly simplistic answers. Act-utilitarianism, for instance, entails that torture is ethical if and only if it maximizes utility. Although there is of course debate about whether, and if so when, torture does this[30-2], the relevant point is that act-utilitarianism requires the issue to be settled through the mere calculation of utility. Similarly, Kantianism has been used to argue that torture is wrong because the tortured individual is treated mere means for the good of others, thus (once again) reducing the question to a single issue (does torture 'respect humanity' in a Kantian sense, or not?).[28-9,33] Conversely, when it comes to 'numbers cases' such as torture, other theories appear to provide too little guidance. For instance, Ross' theory of *prima facie* duties includes a duty to promote a maximum of aggregate good[56] as well as a duty of non-maleficence, or duty not to harm.[57] Since torture is harmful but might produce maximum aggregate good, Ross' theory provides no clear guidance. Similarly, while virtue-theoretic analyses of right action broadly instruct one to act as the virtuous individual would act[51,58], there seem to be plausible virtue-ethical arguments both in favor of torture (it is the responsible, virtuous thing to do in response to modern terrorism) and against it (it is cruel and unnecessary). And so on.

I believe that Rightness as Fairness provides a balanced and more nuanced analysis. Since all of the cases just discussed (the Organ Donor Case, Trolley Cases, and torture) involve coercion and assistance, the Principle of Fair Negotiation entails that they are cases in which the Principle of Negative and Positive Fairness apply. Next, the Principle of Negative Fairness tells us that we should aim to minimize coercion in the world, setting costs aside, and the Principle of Positive Fairness instructs us to assist others in achieving ends in which assistance is helpful and desired, (once again) setting costs aside. Thus, in 'numbers' cases, setting all costs aside, we should aim to minimize the number of people coerced and assist as many as we can who would benefit from and desire assistance. However, in order to determine whether these principles apply, and if so, how they are to be weighed against one another and against costs, Rightness Fairness holds that we must do so through the Principle of Fair Negotiation. Therefore, let us do so, working through some of the cases summarized above.

Begin with the Organ Donor Case. In order to properly apply the Principle of Fair Negotiation, we need to specify who is to be included in the negotiation—as the Principle of Fair Negotiation states that 'all beings affected' should be included. Here, however, we face a problem: one that I believe illuminates much of what is wrong with thought-experiments such as the Organ Donor Case. Traditionally, philosophers who present such cases suggest that by considering the case in isolation—stipulating who will be affected by action, and how—we can 'isolate the case's morally relevant features.' One problem, however, is that it is unclear whether considering the cases as formulated actually does this. Rather, considering the cases in isolation may abstract away from morally relevant facts, such as the broader social effects of actions in similar real-world cases. Indeed, as we will now see,

I believe this concern is brought to light by using Rightness as Fairness to analyze abstract versus real-world cases.

Consider first the classic Organ Donor case in isolation, where one knows for certain that one can kill one healthy person to save five lives, with no further effects beyond the case at hand. If we apply Rightness as Fairness to this case so stated, it may at first appear right to kill the person for their organs. After all, if we imagine everyone involved motivated by the Principles of Negative and Positive Fairness as ideals, and then give everyone equal bargaining power in how to apply these principles in light of costs involved, as required by the Principle of Fair Negotiation, then it might seem as though the one person will be outbargained by the many in favor of the conclusion that they should be killed for their organs. However, this is too quick a conclusion 'in the real world.' The Principle of Virtues of Fairness holds that we should have stable dispositions to conform our actions to the Principles of Negative and Positive Fairness, and the Principle of Fair Negotiation—dispositions that we should have cultivated prior to facing such a situation, given real-world living. And when we consider the real world, Rightness as Fairness instructs us to develop dispositions to apply the principles of Rightness as Fairness in a manner that supports *not* killing the one person for their organs. Here is how. In the real-world, neither physicians, patients, patients' families, nor people in society more broadly are all-knowing. In particular, we cannot typically know in advance—in a specific 'numbers' case—precisely who our actions will affect, and in what way. What we do know, however, is that a moral norm permitting doctors to kill one person to save five would create fear among patients—giving doctors immense power over life and death—and incite outrage in society among the victims of such behavior. Further, given such fear, many

people would likely avoid medical treatment unless absolutely necessary, potentially leading to disastrous results (individuals dying from preventable injuries or illnesses, failing to detect health problems until they are critical, performing poorly at home and at work due to having an ongoing, untreated illness, and so on). We therefore could reasonably judge that the costs incurred by the practice of harvesting organs from a healthy individual in order to save a greater number of sick individuals would be greater than the costs of not doing so. Consequently, guided (at least implicitly) by the Principles of Negative and Positive Fairness (medical ethics is, after all, guided by the principles of 'doing no harm' and beneficence[59]), we have negotiated together laws and norms against killing healthy people to save the sick (in line with the Principle of Fair Negotiation). As such, according to Rightness as Fairness, killing one healthy patient to save five is wrong in the real world. Moreover, because Rightness as Fairness requires us to develop stable dispositions to conform to the above principles, it follows that if any of us were to find ourselves in the situation (the 'isolated' organ donor case), we should be disposed to apply the principles in the same way. In other words, if we were a physician in such an isolated case, we should be strongly disposed not to want to kill one to save the five. Similarly, if we were one of the five dying patients, we should be disposed to think it would be wrong—contrary to fair standards of medical ethics—for one healthy person to be killed for their sake. And so, if people had the dispositions Rightness as Fairness entails they should have, then even in the isolated case, a fair procedure of negotiation should still lead to the conclusion that the one should not be sacrificed for the many.

As such, Rightness as Fairness gives an intuitive analysis of the Organ Donor Case. When we think of such a case, we intuitively think those involved should find the prospect

of killing one person to save five abhorrent. And we have such strong visceral reactions to the case—strong dispositions to favor the one healthy person over the many—for more or less the reasons I have outlined. In the real world, unless doctors, patients, and people in society more broadly found killing patients abhorrent, our lives would be clouded by fear and outrage: fear of our own lives or the lives of those we care about being sacrificed for others, and outrage in response to such cases occurring. Because this is how we, real-life human beings, respond to such cases in the real world—and because we have negotiated standards of medical ethics against it—we considering killing one person to save five wrong, just as Rightness as Fairness does.

Now let us consider the Trolley Cases. Rightness as Fairness also has compelling implications here as well. In particular, it explains why many of our intuitions differ for different Trolley Cases, and suggests that there is no determinate answer as to what is right in those cases. Allow me to explain. Rightness as Fairness holds, once again, that we should approach applied cases motivated by Virtues of Fairness: that is, through settled dispositions to apply the Principles of Negative Fairness, Positive Fairness, and Principle of Fair Negotiation. On the one hand, the Principle of Negative Fairness holds that we should aim to minimize coercion in the world, setting costs aside—which supports killing one person in a Trolley Case to save the many. On the other hand, however, as we have just seen, Rightness as Fairness requires us in other cases (in medical ethics) to develop dispositions against sacrificing the few for the many. Consequently, Rightness as Fairness holds that when we encounter Trolley Cases—cases which we virtually never encounter, unlike medical ethics cases, which physicians and patients do face at times—we should be pulled in two directions (the Principle of Negative Fairness pulling us in the direction of

killing one to save many, our Virtues of Fairness pulling us in the direction of not doing so). And indeed, notice that this is how we encounter such cases. We are pulled in two directions, wanting to minimize the number of people killed while at the same time feeling an aversion to doing so. Consequently, when we imagine Trolley Cases, it is unclear how a fair negotiation (in conformity with the Principle of Fair Negotiation) might go. Since, on the one hand, a negotiation in such an awful situation might lead to five people 'outvoting' the one in favor of their lives, Rightness as Fairness allows that killing one to save five could be right. At the same time, however, since such a negotiation has not occurred—since society has never negotiated clear norms about what to do in such cases—Rightness as Fairness implies that the opposite could be true as well. As such, Rightness as Fairness generates an indeterminate result for Trolley Cases: it does not give a firm answer as to whether killing one person to save five is right in a Trolley Case, because this is not something human beings have negotiated. Finally, Rightness as Fairness explains why our intuitions are more strongly against killing one person in the 'Push the Man' case than in the 'Pull the Switch' case: because physically assaulting people has a distinct tendency to be unfair (assaulting people rarely minimizes coercion in the world), human beings have negotiated norms against assault, and should internalize those norms as virtues of fairness, according to Rightness as Fairness. Consequently, when we imagine the 'Push the Man' case, Rightness as Fairness entails that we should be more opposed to it than the 'Pull the Switch' case—while still holding that it is indeterminate whether one should push the man (since again, clear norms for this case have never been negotiated).

As such, Rightness as Fairness provides an analysis of Trolley Cases that coheres with our initial reactions to them (namely, that there doesn't appear to be a good option in

either case, but that pushing the man in front of the trolley seems particularly abhorrent). In addition to holding that it is presently morally indeterminate what should be done in such cases (which, again, sits well with our present judgments), Rightness as Fairness also provides a method for potentially resolving such cases. It entails that if we want to know what is right in various Trolley Cases, we need to collectively negotiate norms for them, much as how we have done in medical ethics and other areas of law. We need to settle and codify norms through a fair procedure if we want a determinate answer to Trolley Cases. Importantly, however, Rightness as Fairness allows that it may be right for us never to actually negotiate this. Since Trolley Cases are virtually never encountered, and collectively deliberating to codified norms would be costly (requiring us to spend time, energy, and other resources to deliberate and reach an agreement), Rightness as Fairness allows us to collectively negotiate never settling Trolley Cases and leaving their moral status indeterminate—which, essentially, is what we have done (we are uncertain about Trolley Cases because we have never negotiated norms for them, and we have never negotiated norms because they are so rare). Of course, this would not be the case if real-world circumstances were to arise that (Heaven forbid) made encountering Trolley Cases more likely.

Finally, consider the case of torture. Rightness as Fairness produces compelling results here as well. Much of the applied ethics literature on torture focuses on 'ticking bomb' cases where the likely costs and benefits of torture are clear (in the standard case, the only way to prevent a bombing is to torture the would-be bomber), and in which the long-term costs on society and the world are completely abstracted away from. In such highly artificial cases, where just about all costs are set aside beyond the results of the

action, Rightness as Fairness provides no determinate result, but rather allows that we should be pulled in two directions: in the direction of torture (vis-à-vis the Principle of Negative Fairness, which would have us minimize the total amount of coercion in the world, which torture might do), and against torture (vis-à-vis the Principle of Virtues of Fairness, since according to this principle we should develop standing dispositions against being unfair to people, which assaulting them usually does). Further, Rightness as Fairness entails that we cannot simply 'read off' whether torture is right in real-world conditions by counting up costs and benefits. Rather, it holds that in order to settle whether torture is ever right in the real world, we must negotiate an answer to the question through a fair procedure involving all those affected (namely, all people in the world as a whole, since the costs and benefits of torture are vast and wide-reaching), and then obey the results of those negotiations, whatever they may be. But this is not only an intuitively compelling answer: it is the kind of answer the world has already pursued, albeit very imperfectly. We have set up international law-making organizations, such as the United Nations, in order to represent citizens around the world and negotiate standards of international laws and norms, including norms concerning torture. Now, of course, two caveats are necessary here. First, international institutions are proxy methods through which we can only approximate fair negotiation between all individuals in the world. Recall that Rightness as Fairness requires us to approximate fair negotiation as closely as we can. Since it is of course impossible to give every person in the world equal bargaining power over international norms for torture, we can only attempt to come as close to that as possible—which is what international institutions (at least ideally) aim to do. Second, I am not suggesting that institutions such as the United Nations are, as they currently exist, come

anywhere close to being fair and equitably responsible to all—and for this very reason, Rightness as Fairness would suggest that we should be wary of simply accepting UN norms. At the same time, given that the UN is arguably the fairest available international lawmaking mechanism in place, Rightness as Fairness suggests that we should provisionally accept its norms and attempt to make the organization fairer in the future. But these too are intuitively compelling results. Many, if not all, of us think UN norms should be provisionally accepted and obeyed for these kinds of reasons, and that the UN should be made as fairer, and indeed, as fair as possible.

3.3 World Poverty

Rightness as Fairness also provides a persuasive account of how we should think about world poverty. In his famous article, 'Famine, Affluence, and Morality', Peter Singer argued that each of us in wealthy, developed nations has a duty to give up our luxuries—luxurious food, large houses, nice cars, etc.—to alleviate world poverty. Singer's argument is based on a general moral principle that he develops through a simple thought-experiment. The thought-experiment is this: you are walking by a shallow pond and witness a person drowning—a person whose life you could save at little cost to yourself. Singer contends that it is obvious that one has a moral duty to help: that this simply reflects the intuitive moral principle that if one can stop something very bad from happening without sacrificing anything of 'comparable moral significance', they morally ought to do so.[60] Next, Singer contends that this principle establishes (1) that distance does not matter, so it does not matter if the harm one could prevent is nearby or halfway across the world, and also (2) that if other people do not do their fair share to stop something very bad from happening, one has a duty to do more than one's fair share.[61] Finally, Singer argues that since world

poverty is very bad, we have the power to take action to prevent at least of some of it, and giving up our luxuries is not of comparable moral significance to the lives we might save, we all have a duty to give up our luxuries to alleviate world poverty.

Many philosophers have resisted Singer's argument on various grounds, two of which Rightness as Fairness verifies. First, some have objected to Singer's notion of 'comparative moral significance.' For although Singer might not think that giving up most of our luxuries is of comparable moral significance to alleviating world poverty, this does not seem obvious to critics. In particular, it seems to many that our ability to simply live our lives—our ability to enjoy the fruits of our hard work, among other things—is of great moral significance, perhaps even more than saving people from world poverty, since our perspective of being entitled to the money and luxuries we earn arguably plays a critical role in incentivizing economic production: the very kind of production that has given us wealth to give to charity.[62] Second, some have argued that Singer's argument involves a pernicious form of 'rampant moralism'—assuming, very implausibly, that we have a duty to prevent bad things from happening regardless of context (Kekes, for instance, argues that moral commonsense strongly suggests that context is critical: it matters whether people are responsible for their bad situations, whether they can take action to alleviate their own misfortunes, whether they want outside assistance, and so on—none of which Singer addresses[63]).

Rightness as Fairness corroborates both critiques. According to Rightness as Fairness, we cannot determine whether we have obligations to alleviate world poverty—and, if so, what the scope of those obligations are—through mere reflection on what is 'of comparable moral significance.' Indeed, Rightness as Fairness holds that Singer errs, just as

Trolley Case theorists err, precisely by focusing on isolated 'test cases' (namely, what we should do when walking by a person drowning in a shallow pond). First, whereas a person drowning in a shallow pond presumably wants help, it is not at all clear that members of impoverished nations want help from outsiders (particularly given that outside 'help' may change their lives in ways they do not want, resulting in significant changes such as moving them from rural farms into urban environments—which may be serious costs[64]). Second, whereas the costs of helping someone from a pond are simple and limited (getting one's clothes wet, being late for an appointment) the total costs of attempting to alleviate world poverty are uncertain, and may involve creating more poverty, political corruption, and other negative outcomes.[65] Rightness as Fairness thus holds that in order to determine whether and to what extent we have duties to alleviate world poverty, we must engage in fair negotiation taking into account the real costs and benefits of alleviating poverty in order to arrive at an answer as to whether the costs of giving up most of our wealth to alleviate world poverty are 'comparable' to the costs of not taking action. In other words, according to Rightness as Fairness, whether a cost is 'comparable' is not something that can be settled 'from on high' by a philosopher. It is something that must be fairly negotiated by real people, in the real world, given their actual lives and the costs as they encounter them.

This, I submit, is a convincing analysis. Rightness as Fairness does not give us a 'pat' answer as to whether and to what extent we should seek to alleviate world poverty. Rather, it presents a method for determining what our duties are, and at what cost. The method, specifically, is to set up international negotiating institutions and procedures that approximate a fair method for (A) determining the best methods for helping people if they indeed want to be helped (in line with the Principle of Positive Fairness), and for (B)

distributing costs. And although the world is currently doing this in a very imperfect way—through international institutions and organizations negotiating 'fair trade' agreements, and so on—the fact we are attempting to address issues of poverty and economic inequality in such a manner sits well with Rightness as Fairness. Moral rightness is about negotiating our duties in the real world, with real people, through as fair of a process as possible.

3.4 Distribution of Scarce Medical Resources

Rightness as Fairness has similarly convincing implications for cases in biomedical ethics—for instance, the issue of how to distribute scarce medical resources such as hospital beds or transplantable organs.

Biomedical ethicists have formulated and defended many different answers to how scarce resources should be distributed, including:[66]

1. Scarce resources should be given to those 'first in line.'
2. Scarce resources should be utilized in whichever manner maximizes average Quality-of-Life-Years (QALYS), or patients most likely to use the resources best.
3. Scarce resources should be diverted to those most in need (most serious cases).
4. Scarce resources should be diverted to those most deserving of them (those who have made good life choices and/or have the greatest social value, viz. 'VIPs', parents, etc.).

All of these answers, however, seem to have problems. For instance, if a scarce resource (e.g. a transplantable organ) is given to those first in line, people in greater need may die (a person first in line may be able to live three years without the organ, whereas a person later in line may only be able to live until next week). Similarly, if scarce resources are

utilized to maximize QALYs, such a policy would seem to wrongly discriminate against the elderly and unhealthy, since younger, healthier patients can be expected to benefit more from scarce resources. Giving to those most in need, however—the elderly and most unhealthy—would seem to waste important resources (diverting organs to people who are likely to die relatively soon anyway, possibly leaving patients who are more likely to live without the organs necessary to survive). Finally, of course, diverting scarce resources to the 'more deserving'—to parents over non-parents, wealthy or powerful 'VIPs' over the poor—seems wrongly discriminatory.

Rightness as Fairness provides a telling and nuanced answer. According to Rightness as Fairness, there (once again) is no simple answer: instead, morality requires us to settle the issue through a fair deliberative process that treats all stakeholders equitably. Notice that this is how we already aim to resolve such dilemmas in practice. In addition to having ethics committees with representatives of stakeholders that deliberate on how scarce resources are to be utilized, we also have a broadly fair democratic process for arriving at legislation to govern the use of scarce resources. Although ethics committees and the democratic process are far from perfectly fair in practice, Rightness as Fairness directs us to aim to make them fairer, and to then abide by the results of their deliberations. Finally, insofar as all of the aforementioned answers to the scarce resources issue single out particular stakeholders to the detriment of others ('first-in-line' is to the maximum advantage of those first in line, 'maximize QALYs' is to the maximum advantage of those who stand to make the best use of scarce resources, etc.), Rightness as Fairness plausibly entails a fair compromise between all of these options. Specifically, suggests:

1. Setting aside some resources (viz. X number of vital organs) for those 'first in line'

2. Setting aside some resources to maximize QALYs.
3. Setting aside some resources for those most in need.
4. Setting aside some resources for those that are the 'most deserving.'
5. Etc.,
6. Where the amount of resources distributed to each class of persons is negotiated broadly in proportion to the number of individuals in each stakeholder group and the relative strength of their interests (because people tend to have much stronger interests in avoiding death than other things—including 'being first in line' or 'deserving' organs—such negotiations should presumably prioritize saving lives to some extent over these other considerations).

A fair compromise between all of these answers is intuitively fair and right. We commonly recognize, for instance, that some people—the President of the United States, parents of young children—have a unique sort of claim to scarce medical resources in light of their responsibilities and accomplishments. However, we also commonly recognize that medical resources should not merely be directed to the most deserving, and indeed, that even people who have made poor life choices—smokers, drug-users, alcohol abusers—have lives worth saving, and should therefore have some claim to scarce resources (though Rightness as Fairness allows that a fair process of public deliberation may see fit to divert smaller amounts of scarce resources to such people, on account of their poor life decisions), etc.

As such, Rightness as Fairness provides an illuminating answer to applied ethical issues regarding scarce resources. First, it entails—in line with commonsense—that there is no simple, one-size-fits-all answer to the question of what ought to be done in cases of

scarce resources. Second, it requires us to arrive at an answer to specific issues (how to distribute organs, hospital beds, etc.) through fair deliberative processes (e.g. ethics committees comprised by stakeholder representatives). Third, it entails that such a fair process should result in a substantially fair conclusion—a fair compromise between existing answers, given that each such answer (first-in-line, maximize QALYs, etc.) currently favors some stakeholders over others. I believe that all of these implications are clearly plausible, and indeed, commonsense.

3.5 The Ethical Treatment of Animals

Finally, let us consider the ethical treatment of non-human animals. Although Rightness as Fairness once again holds that there are no simple answers, it does lend support to two answers—responsible, compassionate animal husbandry and conservation efforts—over others, including vegetarianism, veganism, and current factory farming practices. Allow me to explain.

On the one hand, animals in nature face all kinds of coercive horrors, such as starvation and disease. According to the Principle of Negative Fairness, these natural horrors are a moral issue: we should care about the coercive horrors that animals experience in nature. Simply leaving animals alone in nature—however much animal advocates may like to romanticize it—is, on Rightness as Fairness, not fair to animals. Just as there is nothing fair about leaving fellow human beings to suffer or die from starvation or disease, so too is there nothing fair about leaving animals to suffer and die from such things in nature. On the other hand, prevailing 'factory farming' methods—methods which treat animals cruelly, allowing them to live only short, miserable lives—is also unfair to animals, as animal rights advocates point out. Such methods simply ignore their lives and

wellbeing, and use them merely for our own purposes (for cheap consumption). Finally, although it might be nice if human beings had the time, energy, and resources to save every diseased or starving animal from the horrors of nature, this would be unfair to us: it would require us to spend our lives—day and night—being dedicated to 'saving animals from nature', regardless of the costs we might have to thereby incur.

How, then, does Rightness as Fairness entail that we morally ought to treat animals? The answer is that we have a duty to deliberate in a manner that is fair to us and to animals about how to advance their welfare and our own. On the one hand, human beings tend to enjoy consuming animal products; such consumption is deeply embedded in many cultures and traditions around the world (I say this, as an aside, as someone who lived a vegetarian lifestyle for the better part of a decade). On the other hand, animals have an interest in living comfortable lives, protected from the many horrors of nature (starvation or disease). We can advance both sets of interests—treating human beings and animals fairly—by engaging in both (A) compassionate animal husbandry, giving farm animals comfortable and reasonably long lives on pastures, while ultimately consuming them for the sake of profits that not only benefit human beings but also animals (insofar as profits from animal agriculture may in turn be used to give more animals decent, comfortable lives in a humane animal agriculture industry), and (B) negotiating conservation efforts to protect wild animals and their habitats (for although we may not reasonably be able to help wild animals avoid disease, we can help them enjoy more comfortable and plentiful lives by protecting them from human encroachment and interference, and ought to do so insofar as it is fair to us). These I believe, are sound conclusions. Although Rightness as Fairness does not entail veganism or vegetarianism—as these practices, while not killing animals for

human purposes, simply leave farm animals to suffer from natural sources of coercion and deprive humans of traditional and longstanding means of sustenance—it requires a compassionate approach to the treatment of animals that is, as far as possible, fair to both them and us.

4 Conclusion

This chapter argued that Four Principles of Fairness, and a general analysis of Rightness as Fairness comprised by their conjunction, emerge from the Moral Original Position, the method Chapter 5 argued specifies the Categorical-Instrumental Imperative's requirements. I have argued that morality is a matter of acting fairly in four ways: coercion-avoidance and minimization (the Principle of Negative Fairness), assisting others who would benefit from and desire our assistance (the Principle of Positive Fairness), applying these two principles by way of process of fair negotiation (the Principle of Fair Negotiation), and developing dispositions to conform to these first three principles (the Principle of the Virtues of Fairness). Finally, I showed that although Rightness as Fairness does not entail *a priori* answers to many applied ethical questions, it provides compelling moral guidance, as it requires solving moral problems through a process of 'principled fair negotiation' that merges reflection on principles with actual negotiation with others. I argued that this is a compelling picture—providing uniquely attractive answers to moral questions ranging from suicide to Trolley Cases, torture, and the ethical treatment of animals.does provide significant, nuanced guidance—which, I have argued, is precisely what we should expect of a sound moral theory. In real life, morality is almost never as simple as merely applying some abstract moral principle(s), such as the Principle of Utility, the Categorical Imperative, Rossian *prima facie* moral rules, or even virtues of character, to

a complex ethical issue. Rather, in real life, morality is a matter of (1) having correct principles in mind (the Principles of Negative and Positive Fairness), but also (2) negotiating with other people, and other sentient beings, to arrive at fair compromises in cases of conflict, given full (or at least emerging) knowledge of psychological, social, and other empirical facts. Indeed, Rightness as Fairness explains and justifies how we actually go about settling moral problems in the real world. We do not solve moral problems 'solipsistically', or merely thinking about moral problems as isolated thinkers. Instead, we set up ethics boards with the aim of giving medical stakeholders a fair say over what ought to be done; we set up governments with the aim of giving citizens a fair say over what ought to be done in their nation; and we set up non-governmental, international organizations with the aim of giving humanity a fair say over moral issues (torture, war, and so on) that potentially affect all of us. Rightness as Fairness therefore provides a compelling new framework for resolving applied moral issues.

Notes

1. Rawls (1999a): 4-5, 7-8, 215-6.
2. Anderson (2014).
3. Farquhar (1984).
4. Bettman, Luce, & Payne (1998).
5. Slovik (1995)
6. Williams (1981): 1–19.
7. Stocker (1976).
8. Annas (1984).
9. Wolf (1982).
10. Young (1998).
11. Wolf (2012): 71, italics added.
12. Baron (2008).
13. Cottingham (1983).
14. Nathanson (2015).
15. Singer (1972).
16. Kuhse & Singer (1985).
17. Murphy (2000).
18. Herman (1993): chs. 1&2.
19. Baron (1995, 2008).
20. Annas (2008).
21. Schwitzgebel (2015).
22. Taurek (1977).

23. Sanders (1988).
24. Cohen (2014).
25. Foot (1967).
26. Thomson & Parent (1986)
27. Thomson (2008b).
28. Fiala (2005).
29. Hill (2007).
30. Allhoff (2006, 2012).
31. Steinhoff (2013).
32. Arrigo (2004).
33. Sussman (2005).
34. Arvan (2013b)
35. Van Gelder *et al.* (2013).
36. Hirschi (2004).
37. Gottfredson & Hirschi (1990).
38. Wilson & Herrnstein (1985): 44-5.
39. Ersner-Hershfield, Wimmer, & Knutson (2009).
40. Ersner-Hershfield, Garton, *et al*. (2009).
41. Hershfield *et al.* (2011).
42. Nathanson (2015): §3.b.i.
43. Kant [1785]: 4:394, 4:399.
44. Skelton (2012): §5.
45. Hursthouse (2013): §3.1.

46. Dancy (2013): §8.

47. Ross [1930]: 19, 30, 31, 33.

48. Das (2003).

49. Svensson (2010)

50. Ross [1930]: 42.

51. Hursthouse (1999): ch. 1.

52. Dancy (1983).

53. Kant [1785]: 4:422-3, 4:429-31.

54. Potter (1993).

55. Cushman, Young, & Hauser (2006).

56. Ross [1930]: 25, 27, 30.

57. Ross [1930]: 21.

58. Zagzebski (2004).

59. Beauchamp & Childress (2009): 13.

60. Singer (1972): 874.

61. Singer (1972): 874-5.

62. Arthur [1996].

63. Kekes (2002).

64. Schmidtz (2000): 685-6.

65. Schmidtz (2000): 686-8.

66. Vaughn (2009): 620-6.

References

Allhoff, F. (2012). *Terrorism, Ticking Time-Bombs, and Torture: A Philosophical Analysis* (Chicago: University of Chicago Press).

____ (2006). 'A Defense of Torture: Separation of Cases, Ticking Time-Bombs, and Moral Justification', *International Journal of Applied Philosophy* 19(2): 243-64.

Anderson, S. (2014). 'Coercion', in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2014 Edition), http://plato.stanford.edu/archives/spr2014/entries/coercion/.

Annas, J. (2008). 'Virtue Ethics and the Charge of Egoism', in P. Bloomfield (ed.), *Morality and Self-Interest* (Oxford: Oxford University Press): 205–23.

____ (1984). 'Personal Love and Kantian Ethics in Effi Briest', *Philosophy and Literature* 8: 15–31.

Arrigo, J.M. (2004). 'A Utilitarian Argument against Torture Interrogation of Terrorists', *Science and Engineering Ethics* 10(3): 543-72.

Arthur, J. [1996]. 'Famine Relief and the Ideal Moral Code', reprinted in S.M. Cahn & P. Markie (eds.), *Ethics: History, Theory, and Contemporary Issues*, 5th ed. (Oxford: Oxford University Press, 2012): 881-92.

Arvan, M. (2013a). 'A New Theory of Free Will', *Philosophical Forum* 44(1): 1-48.

Baron, M. (2008). 'Virtue Ethics, Kantian Ethics, and the "One Thought too Many" Objection', in M. Betzler (ed.) *Kant's Ethics of Virtue* (Berlin: Walter de Gruyter): 245-77.

_____ (1995). *Kantian Ethics Almost Without Apology* (Ithaca: Cornell University Press).

Beauchamp, T.L., Childress, J.F. (2009). *Principles of Biomedical Ethics*, 6th ed. (Oxford: Oxford University Press).

Bettman, J.R., Luce, M.F, Payne, J.W. (1998). 'Constructive consumer choice processes', *Journal of Consumer Research* 25: 187–217.

Cohen, Y. (2014). 'Don't Count on Taurek: Vindicating the Case for the Numbers Counting', *Res Publica* 20(3): 245-61.

Cottingham, J. (1983). 'Ethics and Impartiality', *Philosophical Studies* 43(1): 83–99.

Cushman, F., Young, L., Hauser, M. (2006). 'The Role of Conscious Reasoning and Intuition in Moral Judgment Testing Three Principles of Harm', *Psychological Science* 17(12): 1082-9.

Dancy, J. (2013). 'Moral Particularism', in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), http://plato.stanford.edu/archives/fall2013/entries/moral-particularism/.

____ (1983). 'Ethical Particularism and Morally Relevant Properties', *Mind* 92(368): 530-47.

Das, R. (2003). 'Virtue Ethics and Right Action', *Australasian Journal of Philosophy*, 81(3): 324-39.

Ersner-Hershfield, H., Garton, M.T., Ballard, K., Samanez-Larkin, G.R., Knutson, B. (2009). 'Don't Stop Thinking About Tomorrow: Individual Differences in Future Self-continuity Account for Saving', *Judgment and Decision Making* 4: 280-6.

____, Wimmer, G. E., Knutson, B. (2009). 'Saving for the Future Self: Neural Measures of Future Self-Continuity Predict Temporal Discounting', *Social Cognitive and Affective Neuroscience*, 4(1): 85-92.

Farquhar, P.H. (1984). 'Utility-assessment Methods', *Management Science* 30(11): 1283–1300.

Fiala, A. (2005). 'A Critique of Exceptions: Torture, Terrorism, and the Lesser Evil Argument', *International Journal of Applied Philosophy* 20(1): 127-42.

Foot, P. (1967). 'The Problem of Abortion and the Doctrine of Double Effect', *Oxford Review* 5: 5-15.

Gottfredson, M.R., Hirschi, T. (1990). *A General Theory of Crime* (Stanford, CA: Stanford University Press).

Herman, B. (1993). *The Practice of Moral Judgment* (Cambridge, MA: Harvard University Press).

Hershfield, H.E., Goldstein, D.G., Sharpe, W.F., Fox, J., Yeykelis, L., Carstensen, LL, *et al*. (2011). 'Increasing Saving Behavior Through Age-Progressed Renderings of the Future Self', *Journal of Marketing Research: November* 2011, Vol. 48, No. SPL: S23-S37.

Hill, D.J. (2007). 'Ticking Bombs, Torture, and the Analogy with Self-Defense', *American Philosophical Quarterly* 44(4): 395-404.

Hirschi, T. (2004). 'Self-control and Crime', In R.F. Baumeister & K.D. Vohs (eds.), *Handbook of Self-Regulation: Research, Theory, and Applications* (New York, NY: Guilford Press): 537–52.

Hursthouse, R. (2013). 'Virtue Ethics', in E.N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy* (Fall 2013 Edition), http://plato.stanford.edu/archives/fall2013/entries/ethics-virtue/.

____ (1999). *On Virtue Ethics* (Oxford: Oxford University Press).

Kant, I. [1785]. *Groundwork of the Metaphysics of Morals*, in M.J. Gregor (ed.), *The Cambridge Edition of the Works of Immanuel Kant: Practical Philosophy* (Cambridge: Cambridge University Press, 1996): 38-108.

Kekes, J. (2002). 'On the Supposed Obligation to Relieve Famine', *Philosophy* 77(4): 503-17.

Kuhse, H., Singer, P. (1985). *Should the Baby Live?: The Problem of Handicapped Infants* (Oxford: Oxford University Press).

Murphy, L.B. (2000). *Moral Demands in Nonideal Theory* (Oxford: Oxford University Press).

Nathanson, S. (2015). 'Act and Rule Utilitarianism', *Internet Encyclopedia of Philosophy*, available at: http://www.iep.utm.edu/util-a-r/, accessed 28 July 2015.

Potter, N. (1993). 'What Is Wrong with Kant's Four Examples', *Journal of Philosophical Research* 18: 213-29.

Rawls, J. (1999a). *A Theory of Justice: Revised Edition* (Cambridge, MA: The Belknap Press of Harvard University Press).

Ross, W.D. [1930]. *The Right and the Good* (Oxford: Oxford University Press, 2002).

Sanders, J.T. (1988). 'Why the Numbers Should Sometimes Count', *Philosophy and Public Affairs* 17(1): 3-14.

Schmidtz, D. (2000). 'Islands in a Sea of Obligation: Limits of the Duty to Rescue', *Law and Philosophy* 19(6): 683-705.

Schwitzgebel, E. (2015). 'Cheeseburger Ethics', *Aeon,* http://aeon.co/magazine/philosophy/how-often-do-ethics-professors-call-their-mothers/, accessed 15 July 2015.

Singer, P. (1972). 'Famine, Affluence, and Morality', *Philosophy and Public Affairs* 1(3): 229-43.

Skelton, A. (2012). 'William David Ross', in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2012 Edition), http://plato.stanford.edu/archives/sum2012/entries/william-david-ross/:

Slote, M. (2009). *Moral Sentimentalism* (Oxford: Oxford University Press).

Slovik, P. (1995). 'The Construction of Preference', *American Psychologist* 50(5): 364-71.

Steinhoff, U. (2013). *On the Ethics of Torture* (Albany: State University of New York Press).

Stocker, M. (1976). 'The Schizophrenia of Modern Ethical Theories', *The Journal of Philosophy* 73(14): 453-66.

Sussman, D. (2005). 'What's Wrong with Torture?', *Philosophy & Public Affairs* 33(1): 1–33.

Taurek, J.M. (1977). 'Should the Numbers Count?', *Philosophy and Public Affairs* 6(4): 293-316.

Thomson, J.J. (2008b). 'Turning the Trolley', *Philosophy & Public Affairs* 36(4): 359-74.

Thomson, J.J., Parent, W. (1986). *Killing, Letting Die, and the Trolley Problem* (Cambridge: Harvard University Press).

Van Gelder, J.L., Hershfield, H.E., Nordgren, L.F. (2013). 'Vividness of the Future Self Predicts Delinquency', *Psychological Science* 24(6): 974-80.

Vaughn, L. (2009). *Bioethics: Principles, Issues, and Cases* (Oxford: Oxford University Press).

Williams, B. (1981). 'Persons, Character, and Morality', in B. Williams, *Moral Luck* (Cambridge: Cambridge University Press): 1–19.

Wilson, J.Q., Herrnstein, R.J. (1985). *Crime & Human Nature: The Definitive Study of the Causes of Crime* (New York, NY: Free Press).

Wolf, S. (2012). 'One Thought Too Many': Love, Morality, and the Ordering of Luck, Value, and Commitment', in U. Heuer & G. Lang (eds.), *Themes From the Ethics of Bernard Williams* (Oxford: Oxford University Press): ch. 3.

Wolf, S. (1982). 'Moral Saints', *Journal of Philosophy* 79(8): 419-39.

Young, W.E. (1998). 'Resentment and Impartiality', *Southern Journal of Philosophy* 36(1): 103-30.

Zagzebski, L. (2004). *Divine Motivation Theory* (Cambridge: Cambridge University Press).