CHAPTER 7

Social Contract Theory

The social contract theory, also known as contractarianism, originated as a political theory and only later developed into a theory of morality. It tells us that laws are just if, and only if, they reflect the terms of a social contract that free, equal, and rational people would accept as the basis of a cooperative life together. Its view of morality stems directly from that political ideal: *actions are morally right just because they are permitted by rules that free, equal, and rational people would agree to live by, on the condition that others obey these rules as well.*

A. THE BACKGROUND OF THE SOCIAL CONTRACT THEORY

The political origins of the social contract theory can be traced back to the ancient Greeks. Early in the *Republic*, Plato's brothers tell Socrates that they find the social contract view both appealing and troubling. They challenge Socrates to tell them what is wrong with it. His answer takes up almost the whole of the book, a testament to the power of contractarianism.

Here is the story that Socrates heard. We are all by nature largely, or entirely, self-interested. What we want is power over others, physical security, plenty of money, and sensual pleasure. Our deepest goal is to lord it over everyone else. Who among us wouldn't want the power of the president or the wealth of Bill Gates—or, ideally, both?

This points to an obvious problem. Everyone wants to be at the top of the heap, and only a few can make it there. Further, no one wants to be a patsy, the person who gets stepped on as others climb the ladder of success. We each want to be number one. But we know that the chances of making it are slim, and we want to avoid being trampled as others claw their way to the top. So what do we do?

If we are rational, we will each agree to curb our self-interest and cooperate with one another. We'll do this *conditionally*—that is, on the condition that others do so as well. A complete free-for-all is going to make everyone miserable. If we all stop trying to get the better of each other, and instead agree to seek a little less for ourselves, then we'll all be better off.

That is what reason and morality require of us, according to the social contract theory. Starting with the assumptions that we each are largely motivated by self-interest, and that it is rational to be that way, contractarianism tells us that we each do best for ourselves by agreeing to limit the direct pursuit of self-interest and accept a bargain that gets us a pretty decent life. That everyone gets such a life means that we give up the chance of an absolutely fabulous life. But we also protect ourselves from a really terrible one, a life in which we are in the thick of a cutthroat competition, vulnerable to the attacks of everyone around us. That is a deal worth making. Here's why.

B. THE PRISONER'S DILEMMA

Consider life's basic scenario: There is intense competition for scarce resources. We each want as much of those resources as we can get. Being rational, we each try to get as much as we can, knowing that more for us means less for someone else. Things are going to get very bad, very quickly.

This is what happened when baseball players, Tour de France cyclists, and Olympic weight lifters began to take increasingly dangerous anabolic steroids, in a bid to gain a competitive edge and lucrative championships. This is what happens when a politician starts a smear campaign and his opponent feels the need to ramp up the abuse in order to stand a fighting chance in the race. This is what always happens in turf battles over the spoils of an illegal drug trade.

These cases all share the same essential features. In each, there is mounting competition over a scarce resource, and many are trying their best to increase their share of it. That seems to be rational, and yet, if everyone stopped being so selfish, each person would be better off.

These sorts of situations, in which everyone would be better off by scaling back their pursuit of self-interest, are known as **prisoner's dilemmas**. The name comes from a scenario, introduced by economists, in which two thieves (call them Al and Bob) are caught and sent to separate detention cells. Being rational, Al and Bob previously made a deal with each other: if they get caught, they'll each keep silent, to thwart the police and protect themselves. Now that they have been captured, the police tell each one the same thing: "If you keep your promise to your partner by keeping quiet, and he rats you out, then he's off the hook, and you're looking at a six-year sentence. If *you* break your word and snitch on him, while he remains silent, you're home free, while he spends the next six years in jail. If you both keep quiet, you'll each get two years. But if you both confess, you'll each get four."

The following diagram will help you keep track of the options. Each number represents years in jail. The first number in each pair is Al's prison sentence; the second is Bob's.

Suppose that both criminals know about the various outcomes, and that both have only one concern at this point: to minimize their jail time. If they are both rational, what are they going to do?

You might think that it's impossible to know the answer, since you don't know enough about Al or Bob, their bond with each other, their trustworthiness, and so on, to make an informed guess. But really, there is no doubt that each is going to confess. They are going to break their promise to each other, landing themselves a four-year sentence apiece. That's a far cry from getting off scot-free, and double the two years they'd get if they each kept quiet.

The important point is that remaining silent is the cooperative strategy. Silence here means keeping one's word, honoring the terms of the deal. Confession is a betrayal, breaking one's promise, abandoning a partner.

Al and Bob are going to betray each other. That's certain. They'll do this because they know the odds, because they are self-interested, and because they are rational.

Why will they confess? Because *no matter what his accomplice does, each criminal will be better off by confessing.*

Consider Al's choices. Suppose that

Bob remains silent. Then if Al confesses, Al is home free. If Al keeps his mouth

		Bob: Remains Silent (Cooperation)	Bob: Confesses (Betrayal)
Al	Remains Silent (Cooperation)	2, 2	6, 0
Al	Confesses (Betrayal)	0, 6	4, 4

shut, Al gets two years. So if Bob remains silent, Al should confess. That will minimize his jail time. That is what he most wants. So, if Al is rational, he will confess.

Now suppose that

Bob confesses. Then if Al confesses, Al gets four years in jail. Silence gets him six. So if Bob confesses, Al should confess, too.

Thus, either way, Al does best for himself by spilling the beans and breaking his promise to Bob. And of course Bob is reasoning in the same way. So they are both going to confess and end up with four years in jail.

The prisoner's dilemma isn't just some interesting thought experiment. It's real life. There are countless cases in which the rational pursuit of self-interest will lead people to refuse to cooperate with one another, even though this leaves everyone much worse off.

C. COOPERATION AND THE STATE OF NATURE

So why don't competitors cooperate? The answer is simple: because it is so risky. The criminals in the prisoner's dilemma could cooperate. But that would mean taking a chance at a six-year sentence and betting everything on your partner's good faith. Unilaterally keeping silent, refusing the use of steroids, forsaking negative campaigning or violence—these are strategies for suckers. Those who adopt them may be virtuous, but they are the ones who will be left behind, rotting in jail, economically struggling, off the Olympic podium, or the victim of an enemy's gunshot. If enough people are willing to do what it takes to ensure that they get ahead, then you've either got to join in the competition or be the sacrificial lamb.

Englishman Thomas Hobbes (1588–1679), the founder of modern contractarianism, was especially concerned with one sort of prisoner's dilemma. He invited the readers of his magnum opus, *Leviathan*, to imagine a situation in which there was no government, no central authority, no group with the exclusive power to enforce its will on others. He called this situation the **state of nature**. And he thought it was the worst place you could ever be.

In his words, the state of nature is a "war of all against all, in which the life of man is solitary, poor, nasty, brutish and short." People ruthlessly compete with one another for whatever goods are available. Cooperation is a sham, and trust is nonexistent. Hobbes himself lived through a state of nature—the English Civil War—and thus had first-hand knowledge of its miseries. If you've ever read *The Lord of the Flies*, you have an idea of what Hobbes is talking about. As I write this, I can turn on my television and see pictures of states of nature from around the world—in parts of Syria, Iraq, and Sudan. The scenes are terrible.

The Hobbesian state of nature is a prisoner's dilemma. By seeking to maximize self-interest, everyone is going to be worse off. In such dire circumstances, everyone is competing to gain as much as he can, at the expense of others. With so much at stake, an all-out competition is bound to be very bad for almost everyone. No one is so smart or strong or well-connected as to be free from danger.

There is an escape from the state of nature, and the exit strategy is the same for all prisoner's dilemmas. We need two things: beneficial rules that require cooperation and punish betrayal, and an enforcer who ensures that these rules are obeyed.

The rules are the terms of the social contract. They require us to give up the freedom to attack and to kill others, to cheat them and lie to them, to beat and threaten them and take from them whatever we can. In exchange for giving up these freedoms (and others), we gain the many advantages of cooperation. It is rational to give up some of your freedom, provided that you stand a good chance of getting something even better in return. The peace and stability of a well-ordered society is worth it. That is the promise of the social contract.

But you need more than good rules of cooperation to escape from a prisoner's dilemma. You also need a way to make sure the rules are kept.

The state of nature comes to an end when people agree with one another to give up their unlimited freedoms and to cooperate on terms that are beneficial to all. The problem with agreements, though, is that they can be broken. And without a strong incentive to keep their promises, people in prisoner's dilemmas are going to break them. Just think of Al and Bob in our original example.

What's needed is a powerful person (or group) whose threats give everyone excellent reason to keep their word. The central power doesn't have to be a government—it could be a mob boss, who threatens Al and Bob with death if they were to break their silence. It could be the International Olympic Committee, with the power to suspend or disqualify athletes who test positive for illegal substances. But in the most general case, in which we are faced with anarchy and are trying to escape from utter lawlessness, what we need is a government to enforce basic rules of cooperation. Without a central government, the situation will spiral downhill into a battleground of competing factions and individuals, warlords and gang bosses, each vying for as much power and wealth as possible. A war of all against all won't be far behind.

D. THE ADVANTAGES OF CONTRACTARIANISM

Contractarianism has many advantages. One of these is that contractarianism explains and justifies the content of the basic moral rules. On the contractarian account, the moral rules are ones that are meant to govern social cooperation. When trying to figure out which standards are genuinely moral ones, contractarians ask us to imagine a group of free, equal, and rational people who are seeking terms of cooperation that each could reasonably accept. The rules they select to govern their lives together are the moral rules. These will closely match the central moral rules we have long taken for granted.

John Rawls (1921–2002), the most famous twentieth-century social contract theorist, had a specific test for determining the rules that the ideal social contractors would support. In his *Theory of Justice* (1971), by most accounts the most important work of political philosophy written in the last century, Rawls has us envision contractors behind a **veil of ignorance**. This is an imaginary device that erases all knowledge of your distinctive traits. Those behind the veil know that they have certain basic human needs and wants, but they know nothing of their religious identity, their ethnicity, their social or economic status, their sex, or their moral character. The idea is to put everyone on an equal footing, so that the choices they make are completely fair.

When placed behind a veil of ignorance, or in some other condition of equality and freedom, what social rules will rational people select? These will almost certainly include prohibitions of killing, rape, battery, theft, and fraud, and rules that require keeping one's word, returning what one owes, and being respectful of others. Contractarianism thus easily accounts for why the central moral rules are what they are—rational, self-interested people, free of coercion, would agree to obey them, so long as others are willing to obey them, too.

The rules of cooperation must be designed to benefit everyone, not just a few. Otherwise, only a few would rationally endorse them, while the rest would rationally ignore them. This allows the contractarian to explain why slavery and racial and sexual discrimination are so deeply immoral. Biased policies undermine the primary point of morality—to create fair terms of cooperation that could earn the backing of everyone. Even if oppressed people identify with the interests of their oppressors, and staunchly defend the system of discrimination, that does not make it right. The correct moral rules are those that free people would endorse for their *mutual* benefit—not for the benefit of one group over another.

A second benefit of contractarianism is that it can explain the objectivity of morality. Moral rules, on this view, are objective. Anyone can be mistaken about what morality requires. Personal opinion isn't the final authority in ethics. Neither is the law or conventional wisdom—whole societies can be mistaken about what is right and wrong, because they may be mistaken about what free, equal, and rational people would include in their ideal social code.

Thus contractarians have an answer to a perennial challenge: if morality isn't a human creation, where did it come from? If contractarianism is correct, morality does not come from God. Nor does it come from human opinion. Rather, morality is the set of rules that would be agreed to by people who are very like us, only more rational and wholly free, and who are selecting terms of cooperation that will benefit each and every one of them.

Thus contractarians don't have to picture moral rules as eternally true. And they can deny that moral rules are just like the rules of logic or of natural science—other areas where we acknowledge the existence of objective truths. The moral rules are the outcomes of rational choice, tailored to the specifics of human nature and the typical situations that humans find themselves in. This removes the mystery of objective morality. Even if God doesn't exist, there can still be objective values, so long as there are mutually beneficial rules that people would agree to if they were positioned as equals, fully rational and free.

A third benefit of contractarianism is that it explains why it is sometimes acceptable to break the moral rules. Moral rules are designed for cooperative living. But when cooperation collapses, the entire point of morality disappears. When things become so bad that the state of nature approaches, or has been reached, then the ordinary moral rules lose their force.

One way to put this idea is to say that every moral rule has a built-in escape clause: do not kill, cheat, intimidate, and so on, *so long as others are obeying this rule as well.* When those around you are saying one thing and doing another, and cannot be counted on to limit the pursuit of their self-interest, then you are freed of your ordinary moral obligations to them.

The basis of morality is cooperation. And that requires trust. When that trust is gone, you are effectively in a state of nature. The moral rules don't apply there, because the basic requirement of moral life—that each person be willing to cooperate on fair terms that benefit everyone—is not met.

This explains why you aren't bound to keep promises made at gunpoint, or to be the only taxpayer in a land of tax cheats. It explains why you don't have to wait patiently in line when many others are cutting in, or to obey a curfew or a handgun law if everyone else is violating it. When you can't rely on others, there is no point in making the sacrifices that cooperative living requires. There is no moral duty to play the sucker.

E. THE ROLE OF CONSENT

Most of us believe that we have a moral duty to honor our commitments. And a contract is a commitment—it is a promise given in exchange for some expected benefit. A social contract differs from other contracts only in the extent of the duties it imposes and the benefits it creates. Since we are morally required to keep our promises, we have a duty to honor the terms of the social contract.

But have we actually promised to live up to any social contract? The Pilgrims did, when they paused before the shores of Massachusetts and together signed the Mayflower Compact in 1620. In ancient Athens, free men were brought to the public forum and directly asked to promise obedience to their city—or leave, without penalty. Naturalized citizens in the United States have long been required to pledge allegiance to the nation's laws. But relatively few adults nowadays have done any such thing. It seems, therefore, that we are not really parties to any such contract, and so are not bound to obey its terms.

Contractarianism would be in deep trouble if it claimed that our moral and legal duties applied only to those who agreed to accept them. *But it makes no such claim*. The social contract that fixes our basic moral duties is not one that any of us has *actually* consented to; rather, it is one that we each *would* agree to were we all free and rational and seeking terms of mutually beneficial cooperation. So the fact that we have never signed a social contract or verbally announced our allegiance to one does not undermine the contractarian project.

Contractarianism does not require you to do whatever the existing laws and social customs tell you to do. Those standards are partly a product of ignorance, past deception and fraud, and imperfect political compromise. We are morally required to live up to the standards that free, rational people would accept as the terms of their cooperative living. It's safe to say that no existing set of laws perfectly lines up with those terms.

Thus contractarianism isn't a simple recipe to do whatever your society says. Rather, it provides a way to evaluate society's actual rules, by seeing how close (or how far) they are to the ideal social code that would be adopted if we were freer, more equal, and more rational than we are. If contractarianism is correct, this ideal social code is the moral law.

F. DISAGREEMENT AMONG THE CONTRACTORS

If the social contract theory is correct, then the moral rules are those that free, equal, and rational people would agree to live by. But what happens if such people disagree with one another? For instance, what if these idealized contractors can't reach a deal about the conditions under which a nation should go to war, or about the kind of aid we owe to the very poor? What happens then?

Rawls solved this problem by making every contractor a clone of every other. Behind the veil of ignorance, all of your distinguishing features go away. No one is any different from anyone else. And so there is no reason to expect any disagreement.

But Hobbes and other contractarians won't stand for this. They can't see why I should follow the rules of someone who is so completely unlike me—a person who is not only absolutely rational but also stripped of all knowledge of his social status, his friendships and family situation, his desires, interests, and hopes. Hobbes and his followers insist that the moral rules are those that we, *situated as we are*, would rationally agree to, provided of course that others would agree to live by them as well.

It's not easy to know how to solve this disagreement between contractarians. On the one hand, Rawls's view is likely to be fairer, since any information that could prejudice our choices is kept from us as we select rules to live by. But Hobbes also has a point, in that we want to make it rational, if we can, for everyone to live by the moral rules. Why should I live according to the rules set by some person who isn't at all like the real me? That's a pretty good question.

I'm sure that you've already figured out that I am not able to answer every good ethical question. This is another one I am going to leave for your consideration. Instead, let's return to our original problem: what should we say when the people choosing the social rules disagree with one another?

Perhaps Rawls is right, and there won't be any disagreement. But what if he's wrong? If contractors disagree, then the actions or policies they disagree about are morally neutral. They are neither required nor forbidden. That's because the moral rules are ones that *all* contractors would agree to. If there are some matters that they can't agree on, then these are not covered by the moral rules.

This could be pretty bad. Or it might be just fine. It all depends on where the disagreement arises (if it ever does). If there are only small pockets of disagreement, regarding relatively trivial matters, then this is hardly a problem. But what if contractors can't agree about war

policy, about whether executions are just, about how to treat the poorest among us? Then this is really serious, since we do think that morality must weigh in on these issues.

So, how much disagreement will there be? There is no easy way to know. We can provide answers only after we know how to describe the contractors and their position of choice. Will they be clones of one another, situated behind the veil of ignorance? Or will they be aware of their different personalities and life situations? Will they be more or less equally situated, or are some going to have a lot more leverage than others? When we say that they are rational, do we have Kant's conception in mind? Or Hobbes's, according to which rationality amounts to reliably serving your self-interest? Or some other conception?

Answers to these questions will make a big difference in deciding on the specific moral rules that a social contract theory favors. These answers will also determine the amount of agreement we can expect from the contractors. There is no shortcut to discovering these answers. To get them, contractarians must defend their own specific version of the theory against competing versions. That is a major undertaking. Until it is done, we cannot know just what the moral rules are or how much contractual disagreement to expect.

G. CONCLUSION

Contractarianism starts with a very promising idea: morality is essentially a social matter, and it is made up of the rules that we would accept if we were free, equal, and fully rational. The heart of the theory is an ideal social code that serves as the true standard for what is right and wrong.

This theory has a lot going for it, as we've seen. It offers us a procedure for evaluating moral claims, and so offers the promise of being able to justify even our most basic moral views. It has an interesting explanation of the objectivity of morality. It can explain why we are sometimes allowed to break the moral rules. It does not require actual consent to the ideal social rules in order for them to genuinely apply to all people. In cases in which the contractors disagree with one another, the social contract theory ought to insist that actions are morally required only if all contractors agree. Whether this is a problem for the view is a matter left for your further reflection.

ESSENTIAL CONCEPTS

Prisoner's dilemma: a situation in which the pursuit of self-interest by all parties leads to a worse outcome than if each were to compromise.
State of nature: anarchy; a situation in which there was no government, no central authority, no group with the exclusive power to enforce its will on others.
Veil of ignorance: an imaginary device that erases all knowledge of your distinctive traits in preparation for selecting principles of justice or morality.

DISCUSSION QUESTIONS

1. What makes a situation a "prisoner's dilemma"? What is the rational thing to do in a prisoner's dilemma situation?
2. What is the state of nature, and why does Hobbes think that such a condition would be so bad? How does Hobbes think that people would be able to emerge from the state of nature?
3. How do contractarians justify moral rules against such things as slavery and torture? Do you find their justifications of such rules to be compelling?
4. Explain how a contractarian defends the objectivity of ethics. Do you find this defense plausible?
5. Suppose that the existing laws of a society require something that you regard as unjust. Does the social contract theory automatically support the morality of the existing law? Why or why not?
6. Would a group of free, equal, and rational people necessarily all agree on a set of rules to live by? If not, is this a problem for contractarianism?

5

Social Contract Theory and the Motive to Be Moral

> The question [why be moral] is on a par with the hazards of love; indeed, it is simply a special case. Those who love one another, or who acquire strong attachments to persons and to forms of life, at the same time become liable to ruin: their love makes them hostages to misfortune and the injustice of others. Friends and lovers take great chances to help each other; and members of families willingly do the same .... Once we love we are vulnerable.
>
> JOHN RAWLS, *A THEORY OF JUSTICE*

Carl owns a very profitable car dealership, and he attributes its success to long hours, talented workers, and, most important, using every trick in the book to manipulate buyers. The cars themselves are not particularly well constructed or fuel efficient, but he claims the exact opposite in his advertisements. Once customers are on his lot, his sales staff takes over, buttering up prospective buyers and seeking out their psychological vulnerabilities. Because they work on commission, it's in their best interest to charge the highest possible price for vehicles, so they budge little from the retail sticker price and secretly add on extra expenses for useless features. They especially inflate prices for women, racial minorities, and the elderly, who frequently end up spending a thousand dollars more on exactly the same vehicle that other customers buy. They coax low-income customers into purchasing luxury vehicles well beyond their price range; as long as loan companies are willing to foot the bill, it's no loss to Carl's dealership if the customers default on loan payments. And when cars come in for repair, the mechanics, who also work on commission, trick customers into paying for expensive repairs that they don't

need. At the end of the day, Carl and his workers go home to their families, giving little thought to the morality of their conduct during business hours.

Although Carl is a fictitious character, all these abuses are well documented among car dealerships. By breaking the rules of morality in seemingly undetectable ways, car dealers and mechanics routinely pad their pockets at the expense of unsuspecting customers. Attempts to cheat the system are clearly not confined to the business world. Over half of all college students cheat on exams, essays, or homework. One in five taxpayers thinks it is okay to cheat on taxes. With more serious offenses, 3 percent of adult Americans are currently behind bars, on probation, or on parole—and those are just the ones who have been caught.[1]

With human self-interest as strong as it is, what can motivate us to always follow the rules of morality? Asked more simply, "Why be moral?" Among the more common answers are these:

- Behaving morally is a matter of self-respect.
- People won't like us if we behave immorally.
- Society punishes immoral behavior.
- God tells us to be moral.
- Parents need to be moral role models for their children.

These are all good answers, and each may be a powerful motivation for the right person. With religious believers, for example, having faith in God and divine judgment might prompt them to act properly. With parents, the responsibility of raising another human being might force them to adopt a higher set of moral standards than they would otherwise. However, many of these answers won't apply to every person: nonbelievers, nonparents, people who don't respect themselves, people who think that they can escape punishment.

One of the more universal motivations to be moral is explained in a philosophical view known as **social contract theory**. The central idea is that people collectively agree to behave morally as a way to reduce social chaos and create peace. Through this agreement—or "contract"—I set aside my own individual hostilities toward others, and in exchange they set aside their hostilities toward me. Life is then better for all of us when we collectively follow basic moral rules.

There are two distinct components to the question "Why be moral?":

1. Why does society need moral rules?
2. Why should I be moral?

The first question asks for a justification for the institution of morality within our larger social framework. The second asks for reasons why I personally should be moral even when it does not appear to be in my interest. This chapter explores social contract theory's answers to both of these questions. We should note that

social contract theory is also an important political concept insofar as it explains where governments get their authority: Citizens agree to give governments power as a means of keeping society peaceful. However, our focus here is on social contract theory's answer to the uniquely ethical question "Why be moral?"

WHY DOES SOCIETY NEED MORAL RULES?

Why does society need moral rules? What does morality do for us that no other social arrangement does? Social contract theory's answer is forcefully presented in the book *Leviathan* (1651) by English philosopher Thomas Hobbes (1588–1679).

Hobbes and the State of Nature

Hobbes believed that human beings always act out of perceived self-interest; that is, we invariably seek gratification and avoid harm. His argument goes like this: Nature has made us basically equal in physical and mental abilities so that, even though one person may be somewhat stronger or smarter than another, each has the ability to harm and even kill the other, if not alone then in alliance with others. Furthermore, we all want to attain our goals such as having sufficient food, shelter, security, power, wealth, and other scarce resources. These two facts, equality of ability to harm and desire to satisfy our goals, lead to social instability:

> From this equality of ability arises equality of hope in the attaining of our ends. And therefore if any two people desire the same thing, which nevertheless they cannot both enjoy, they become enemies; and in the way to their end, which is principally their own preservation and sometimes their enjoyment only, endeavor to destroy or subdue one another. And from hence it comes to pass, that where an invader hath no more to fear, than another man's single power; if one plant, sow, build, or possess a convenient seat, others may probably be expected to come prepared with forces united, to dispossess, and deprive him, not only of the fruit of his labor, but also of his life or liberty. And the invader again is in the like danger of another.[2]

Given this state of insecurity, people have reason to fear one another. Hobbes calls this a **state of nature**, in which there are no common ways of life, no enforced laws or moral rules, and no justice or injustice, for these concepts do not apply. There are no reliable expectations about other people's behavior, except that they will follow their own inclinations and perceived interests, tending to be arbitrary, violent, and impulsive. The result is a war of all against all:

> Hereby it is manifest, that during the time men live without a common power to keep them all in awe, they are in that condition which is called war; and such a war, as is for *every man, against every man*. For war consists not in battle only or in the act of fighting; but in a tract of time,

> wherein the will to contend in battle is sufficiently known: and therefore the notion of *time*, is to be considered in the nature of war; as it is in the nature of weather. For as the nature of foul weather lies not in the shower or two of rain, but in an inclination thereto of many days together; so the nature of war consists not in actual fighting, but in the known disposition thereto, during all the time there is no disposition to the contrary.

Hobbes described the consequence of this warring state of nature here:

> In such a condition, there is no place for industry; because the fruit thereof is uncertain; and consequently no cultivating of the earth; no navigation, nor use of the comfortable buildings; no instruments of moving, and removing, such things as require much force; no knowledge of the face of the earth; no account of time; no arts; no literature; no society; and which is worst of all, continual fear, and danger of violent death; and the life of man solitary, poor, nasty, brutish and short.

But this state of nature, or more exactly, state of anarchy and chaos, is in no one's interest. We can all do better if we compromise, give up some of our natural liberty—to do as we please—so that we will all be more likely to get what we want: security, happiness, power, prosperity, and peace. So, selfish yet rational people that we are, according to Hobbes, we give up some of our liberty and agree to a *social contract*, or *covenant*. This agreement sets up both rules and a governing force: The rules create an atmosphere of peace, and the government ensures that we follow the rules out of fear of punishment. Only within this contract does morality arise and do justice and injustice come into being. Where there is no enforceable law, there is neither right nor wrong, justice nor injustice.

Thus, morality is a form of social control. We all opt for an enforceable set of rules such that if most of us obey them most of the time, then most of us will be better off most of the time. Perhaps a select few people may actually be better off in the state of nature, but the vast majority will be better off in a situation of security and mutual cooperation. Some people may cheat and thus go back on the social contract, but as long as the majority honors the contract most of the time, we will all flourish.

Hobbes does not claim that a pure state of nature ever existed or that humanity ever really formally entered into such a contract, although he notes that such a state actually exists among nations, so a "cold war" keeps us all in fear. Rather, Hobbes explains the function of morality. He answers the question "Why do we need morality?" Why? Because without it, existence would be an unbearable hell in which life is "solitary, poor, nasty, brutish and short."

Hobbesian Morality and *Lord of the Flies*

William Golding's classic novel *Lord of the Flies* (1954)[3] brilliantly portrays the Hobbesian account of morality. In this work, a group of boys, ages 6 to 12 years old, from an English private school, have been cast adrift on an uninhabited Pacific island and have created their own social system. For a while, the

constraints of civilized society keep things peaceful, but soon their system unravels into brutal chaos. The title *Lord of the Flies* comes from a translation of the Greek "Beelzebub," which is a name for the devil. Golding's point is that we need no external devil to bring about evil but that we have found the devil and he is *us*. Ever-present, ever-waiting for a moment to strike, the devil emerges from the depths of the subconscious whenever there is a conflict of interest or a moment of moral laziness. Let's consider some main themes of Golding's story, which illustrate how the dominance of the devil within us proceeds through fear, hysteria, violence, and ultimately leads to death.

In the novel, all the older boys recognize the necessity of procedural rules. During an assembly, only the boy who has the white conch shell, the symbol of authority, may speak. They choose the leader democratically and invest him with limited powers. Even the evil Roger, while taunting little Henry by throwing stones near him, manages to keep the stones from harming the child:

> Here, invisible yet strong, was the taboo of the old life. Round the squatting child was the protection of parents and school and policemen and the law. Roger's arm was conditioned by a civilization that knew nothing of him and was in ruins.

After some initial euphoria in being liberated from the adult world of constraints and entering an exciting world of fun in the sun, the children come up against the usual irritations of social existence: competition for power and status, neglect of social responsibility, failure of public policy, and escalating violence. Two boys, Ralph and Jack, vie for leadership, and a bitter rivalry emerges between them. As a compromise, a division of labor ensues in which Jack's choirboy hunters refuse to help the others in constructing shelters. Freeloading soon becomes common because most of the children leave their tasks to play on the beach. Neglect of duty results in their failure to be rescued by a passing airplane.

Civilization's power is weak and vulnerable to primitive, explosive passions. The sensitive Simon, the symbol of religious consciousness, is slaughtered by the group in a wild fury. Only Piggy and Ralph, mere observers of the homicide, feel sympathetic pangs of guilt at this atrocity.

Piggy (the incarnation of philosophy and culture) with his broken spectacles and asthma becomes ever more pathetic as the chaos increases. He reaches the depths of his ridiculous position after the rebels, led by Jack, steal his spectacles to harness the sun's rays for starting fires. Ralph, the emblem of not-too-bright but morally good civilized leadership, fails to persuade Jack to return the glasses, and Piggy then asserts his moral right to them:

> You're stronger than I am and you haven't got asthma. You can see.... But I don't ask for my glasses back, not as a favor. I don't ask you to be a sport ... not because you're strong, but because what's right's right. Give me my glasses.... You got to.

Piggy might as well have addressed the fire itself, for in this state of moral anarchy moral discourse is a foreign tongue that only incites the worst elements to greater immorality. Roger, perched on a cliff above, responds to moral

reasoning by dislodging a huge rock that hits Piggy and flings him to his death forty feet below.

A delegation starts out hunting pigs for meat. Then they find themselves enjoying the kill. To drown the initial shame over bloodthirstiness and take on a persona more compatible with their deed, the children paint themselves with colored mud. Being liberated from their social selves, they kill without remorse whoever gets in their way. The deaths of Simon and Piggy (the symbols of the religious and the philosophical, the two great fences blocking the descent to hell) and the final hunt with the "spear sharpened at both ends" signal for Ralph the depths of evil in the human heart.

Ironically, it is the British navy that finally comes to the rescue and saves Ralph (civilization) just when all seems lost. But, the symbol of the navy is a two-faced warning. On the one hand, it symbolizes that a military defense is unfortunately sometimes needed to save civilization from the barbarians (Hitler's Nazis or Jack and Roger's allies), but on the other hand it symbolizes the quest for blood and vengeance hidden in contemporary civilization. The children's world is really only a stage lower than the adult world from whence they come, and that shallow adult civilization could very well regress to tooth and claw if it were scratched too sharply. The children were saved by the adults, but who will save the adults who put so much emphasis on military enterprises and weapons systems in the name of so-called defense?

The fundamental ambiguity of human existence is visible in every section of the book, poignantly mirroring the human condition. Even Piggy's spectacles, the sole example of modern technology on the island, become a curse for the island as Jack uses them to ignite a forest fire that will smoke out their prey, Ralph, and burn down the entire forest and destroy the island's animal life. It is a symbol both of our penchant for misusing technology to vitiate the environment and our ability to create weapons that will lead to global suicide.

Social Order and the Benefits of Morality

We learn from *Lord of the Flies* that rules formed over the ages and internalized within us hold us back and hopefully defeat the devil in society, wherever that devil might reside. Again, from Hobbes's perspective, morality consists of a set of rules such that, if nearly everyone follows them, then nearly everyone will flourish. These rules restrict our freedom but promote greater freedom and well-being. More specifically, the five social benefits of establishing and following moral rules accomplish the following:

1. Keep society from falling apart.
2. Reduce human suffering.
3. Promote human flourishing.
4. Resolve conflicts of interest in just and orderly ways.
5. Assign praise and blame, reward and punishment, and guilt.

All these benefits have in common the fact that morality is a social activity: It has to do with society, not the individual in isolation. If only one person exists on an island, no morality exists; indeed, some behavior would be better for that person than others—such as eating coconuts rather than sand—but there would not be morality in the full meaning of that term. However, as soon as a second person appears on that island, morality also appears. *Morality* is thus a set of rules that enable us to reach our collective goals. Imagine what society would be like if we did whatever we pleased without obeying moral rules. I might promise to help you with your homework tomorrow if you wash my car today. You believe me. So you wash my car, but you are angered when I laugh at you tomorrow while driving off to the beach instead of helping you with your homework. Or you loan me money, but I run off with it. Or I lie to you or harm you when it is in my interest or even kill you when I feel the urge.

Under such circumstances, society would completely break down. Parents would abandon children, and spouses would betray each other whenever it was convenient. No one would have an incentive to help anyone else because cooperative agreements would not be recognized. Great suffering would go largely unhindered, and people would not be very happy. We would not flourish or reach our highest potential.

I visited the country of Kazakhstan shortly after the collapse of the Soviet Union, when it was undergoing a difficult shift from communism to democracy. During this transition with the state's power considerably withdrawn, crime was increasing and distrust was prevalent. At night, trying to navigate my way up the staircases in the apartment building where I was staying, I was in complete darkness. I asked why there were no light bulbs in the stair-wells, only to be told that the residents stole them, believing that, if they did not take them, their neighbors would. Absent a dominant authority, the social contract had eroded, and everyone had to struggle in the darkness—both literally and metaphorically.

We need moral rules to guide our actions in ways that light up our paths and prevent and reduce suffering, enhance human well-being (and animal well-being, for that matter), resolve our conflicts of interest according to recognizably fair rules, and assign responsibility for actions so that we can praise, blame, reward, and punish people according to how their actions reflect moral principles. In a world becoming ever more interdependent, with the threats of terrorism and genocide, we need a sense of global cooperation and a strong notion of moral responsibility. If the global community is to survive and flourish, we need morality as much now as we ever have in the past.

WHY SHOULD I BE MORAL?

Let's agree with Hobbes's social contract theory that moral rules are needed for social order: Morality serves as an important antidote to the state of nature, and unless there is general adherence to the moral point of view, society will break down. There remains, though, a nagging question: "Why should I join in?" If I'm sly enough,

I can break moral rules when they benefit me but never get caught and thus avoid being punished. What motivation is there for me to accept the moral viewpoint at all? This question was raised over two millennia ago by Plato in his dialogue, *The Republic*, where he tells the story of Gyges.

The Story of Gyges

In Plato's story, Gyges is a shepherd who stumbles upon a ring that at his command makes him invisible and, while in that state, he can indulge in his greed to the fullest without fear of getting caught. He can thus escape the restraints of society, its laws, and punishments. So, he kills the king, seduces his wife, and becomes king himself. The pertinent question raised by the story is this: Wouldn't we all do likewise if we too had this ring?

To sharpen this question, let's recast the Gyges story in contemporary terms. Suppose there were two brothers, Jim and Jack. Jim was a splendid fellow, kind and compassionate, almost saintly, always sacrificing for the poor, helping others. In fact, he was too good to be true. As a young man, he was framed by Jack for a serious crime, was imprisoned, and was constantly harassed and tortured by the guards and prisoners. When released, he could not secure employment and was forced to beg for his food. Now he lives as a streetperson in a large city, in poor health, without a family, and without shelter. People avoid him whenever they can because he looks dangerous. Yet, in truth, his heart is as pure as the driven snow.

Jack, the older brother who framed Jim, is as evil as Jim is good. He also is as "successful" as Jim is "unsuccessful." He is the embodiment of respectability and civic virtue. He is a rising and wealthy corporate executive who is praised by all for his astuteness and appearance of integrity (the latter of which he lacks completely). He is married to the most beautiful woman in the community, and his children all go to the best private schools. Jack's wife is completely taken in by his performance, and his children, who hardly know him, love him unconditionally. He is an elder in his church, on the board of directors of various charity groups, and he was voted the Ideal Citizen of his city. Teachers use him as an example of how one can be both morally virtuous and a successful entrepreneur. He is honored and admired by all. Yet he has attained all his success and wealth by ruthlessly destroying people who trusted him. He is in reality an evil man.

So, the question posed by the story of Gyges is this: If you had to make a choice between living either of these lives, which life would you choose? That of the unjust brother Jack who is incredibly successful or that of the just brother Jim who is incredibly unsuccessful?

Let's consider two reasons for opting to live the life of Jim, the good man who through no fault of his own is a social outcast. Plato argued that we should choose the life of the "unsuccessful" just person because it's to our advantage to be moral. He draws attention to the idea of the harmony of the soul and argues that immorality corrupts the inner person, whereas virtue purifies the inner person, so one is happy or unhappy in exact proportion to one's moral integrity. Asking to choose between being morally good and immoral is like asking to choose between being healthy and sick. Even if the immoral person has material

benefits, he cannot enjoy them in his awful state, whereas the good person may find joy in the simple pleasures despite poverty and ill fortune.

Is Plato correct? Is the harm that Jim suffers compensated by the inner goodness of his heart? Is the good that Jack experiences outweighed by the evil of his heart? Perhaps we don't know enough about the hearts of people to be certain who is better off, Jim or Jack. But perhaps we can imagine people like Jack who seem to flourish despite their wickedness. They may not fool us completely, but they seem satisfied with the lives they are living, moderately happy in their business and personal triumphs. And perhaps we know of some people like Jim who are really very sad despite their goodness. They wish they had meaningful work, a loving family, friends, and shelter; but they don't, and their virtue is insufficient to produce happiness. Some good people are unhappy, and some bad people seem to be happy. Hence, the Socratic answer on the health–sickness analogy may not be correct.

Plato's second answer is a religious response: God will reward or punish people on the basis of their virtue or vice. The promise is of eternal bliss for the virtuous and hard times for the vicious. God sees all and rewards with absolute justice according to individual moral merit. Accordingly, despite what may be their differing fates here on earth, Jim is infinitely better off than Jack. If religious ethics of this sort is true, it is in our self-interest to be moral. The good is really good for us. The religious person has good reason to choose the life of the destitute saint.

We'll take up the relationship of religion to morality in a later chapter, but we can say this much about the problem: Unfortunately, we do not know for certain whether there is a God or life after death. Many sincere people doubt or disbelieve religious doctrines, and it is not easy to prove them wrong. Even the devout have doubts and probably cannot be sure of the truth of the doctrine of life after death and the existence of God. In any case, millions of people are not religious, and the question of the relationship between self-interest and morality is a pressing one. Can a moral philosopher give a nonreligious answer as to why they should choose to be moral all of the time?

MORALITY, SELF-INTEREST, AND GAME THEORY

Attempting to prove that we should always be moral is an uphill battle because, as we've seen, countless situations may arise in which it's in our best interest to break the rules of morality as long as we don't get caught. Social contract theorists have recently attempted to resolve the conflict between morality and self-interest by drawing from a field of study called **game theory**. The idea behind game theory is to present situations in which players make decisions that will bring each of them the greatest benefit; these games then provide easy models for understanding more complex situations of social interaction in the real world. A simple game like Monopoly, for example, models the real dog-eat-dog world of business in which you need to kill the competition before the competition kills you. At the same time, Monopoly shows the devastating results

on society when a single person succeeds in owning everything. The most common game theory scenario in philosophy is the Prisoner's Dilemma, and this is frequently used to illuminate the tension between morality and self-interest.

Game 1: The Prisoner's Dilemma

The Prisoner's Dilemma scenario is this. The secret police in another country have arrested two of our spies, Sam and Sue. Prior to being caught, Sam and Sue have agreed to keep silent during interrogation if they are ever arrested. Now that they are in the hands of the enemy, they both know that if they adhere to their agreement to keep silent the police will be able to hold them for four months; but if they violate their agreement and both confess that they are spies, they will each get six years in prison. However, if one adheres and the other violates, the one who adheres will get nine years, and the one who confesses will be let go immediately. We might represent their plight with the following matrix. The figures on the left represent the amount of time Sam will spend in prison under the various alternatives, and the figures on the right represent the amount of time that Sue will spend in prison under those alternatives.

		Sue	
		Adheres	*Violates*
Sam	*Adheres*	4 months, 4 months	0 time, 9 years
	Violates	9 years, 0 time	6 years, 6 years

Initially, Sam reasons in this manner: Either Sue will adhere to the agreement or she will violate it. If Sue adheres, then Sam should violate because it's better for him to spend zero time in prison than four months. On the other hand, if Sue violates, then Sam should violate because it's better for him to spend six years in prison than nine years. Therefore, no matter what Sue does, it's in Sam's best interest to violate their agreement. However, Sue reasons exactly the same way about Sam and will conclude that it is in her best interest to violate the agreement. Here's the catch: If both reason in this way, they will obtain the second-worst position—six years each, which we know to be pretty awful. If they could only stick to their original agreement and stay silent, they could each do better—getting only four months. But how can they confidently do that without magically reading each other's minds to see the other's true intentions? They can't and thus each will be forced to look out for his or her own best interest and violate their original agreement.

In a nutshell, here's the lesson that the Prisoner's Dilemma teaches us about violating the rules of morality. It's better for me to secretly violate society's rules, regardless of what other people do. It would be nice if the Prisoner's Dilemma told us that adhering to morality was the best thing for me, but unfortunately it shows the opposite. What do we do now? Remember that the point of games like the Prisoner's Dilemma is to provide an easy model for understanding complex social situations, such as how I might benefit by adhering to the rules of

morality. The Prisoner's Dilemma, though, might not be a very good model for this. In particular, it inaccurately depicts moral choices as a one-shot event: Sam and Sue are in a single situation in which they must make a single choice about whether to adhere to or violate their initial agreement to stay silent. But morality is not a single-issue decision. On a daily basis, we decide whether or not to violate society's moral rules when we might benefit from deception. Should I cheat on my taxes? Should I rack up charges on a bogus credit card? Should I defraud a trusting buyer on eBay? Morality is more like a game in which each player takes several turns, so we need to consider a different game model.

Game 2: Cooperate or Cheat

Consider this alternative game theory scenario called Cooperate or Cheat.[4] In it there are two players and a banker who pays out money or fines to the players. Each player has two cards, labeled "Cooperate" and "Cheat." Each move consists of both players simultaneously laying down one of their cards. Suppose you and I are playing against one another. There are four possible outcomes:

Outcome 1. We both play Cooperate. The banker pays each of us $300. We are rewarded nicely.

Outcome 2. We both play Cheat. The banker fines each of us $10. We are punished for mutual defection.

Outcome 3. You play Cooperate and I play Cheat. The banker pays me $500 (Temptation money) and you are fined $100 (a Sucker fine).

Outcome 4. I play Cooperate and you play Cheat. The banker fines me $100 and pays you $500. This is the reverse of Outcome 3.

The game continues until the banker calls it quits. Theoretically, I could win a lot of money by always cheating. After twenty moves, I could hold the sum of $10,000—that is, if you are sucker enough to continue to play Cooperate, in which case you will be short $2,000. If you are rational, you won't do that. If we both continually cheat, we'll each end up minus $200 after twenty rounds.

Suppose we act on the principle "Always cooperate if the other fellow does and cheat only if he cheats first." If we both adhere to this principle, we'll each end up with $6,000 after our twenty rounds—not a bad reward! And, we have the prospects of winning more if we continue to act rationally.

We may conclude that rational self-interest over the long run would demand that you and I cooperate. While I might gain greater rewards by cheating, it comes at a high risk of winning much less. As contemporary social contractarian David Gauthier puts it, "Morality is a system of principles such that it is advantageous for everyone if everyone accepts and acts on it, yet acting on the system of principles requires that some persons perform disadvantageous acts."[5] The game of Cooperate or Cheat illustrates that morality is the price that we each have to pay to keep the minimal good that we have in a civilized society. We have to bear some disadvantage in loss of freedom (analogous to paying membership dues in an important organization) so that we can have both protection from the onslaughts of chaos and promotion of the good life. Because an

orderly society is no small benefit, even a selfish person who is rational should allow his or her freedom to be limited.

The answer, then, to the question "Why should I be moral" is that I allow some disadvantage for myself so that I may reap an overall, long-run advantage.

THE MOTIVE TO ALWAYS BE MORAL

The game of Cooperate or Cheat informs us that even the amoralist must generally adhere to the moral rule because it will give him or her some long-term advantage. There remains, however, a serious problem: The clever person will still break a moral rule whenever he or she can do so without getting detected and unduly undermining the whole system. This clever amoralist takes into account his overall impact on the social system and cheats whenever a careful cost–benefit analysis warrants it. Reaping the rewards of his clever deceit, he may even encourage moral education so that more people will be more dedicated to the moral rule, which in turn will allow him to cheat with greater confidence.

The Paradox of Morality and Advantage

Gauthier describes this problem of the clever amoralist through what he calls the **paradox of morality and advantage**. He writes,

> If it is morally right to do an act, then it must be reasonable to do it. If it is reasonable to do the act, then it must be in my interest to do it. But sometimes the requirements of morality are incompatible with the requirements of self-interest. Hence, we have a seeming contradiction: It both must be reasonable and need not be reasonable to meet our moral duties.[6]

Laid out more formally, the argument is this:

(1) If an act is morally right, then it must be reasonable to do it.
(2) If it is reasonable to do the act, then it must be in my interest to do it.
(3) But sometimes the requirements of morality are incompatible with the requirements of self-interest.
(4) Hence, a morally right act must be reasonable and need not be reasonable, which is a contradiction.

The problematic premise seems to be the second one claiming that our reasons for acting have to appeal to self-interest. For simplicity, let's call this the *principle of rational self-interest*.

Might we not doubt this principle of rational self-interest? Could we not have good reasons for doing something that goes against our interest? Suppose Lisa sees a small boy about to get run over by a car and, intending to save the child, hurls herself at the youngster, fully aware of the danger to herself. Lisa's interest is in no way tied up with the life of that child, but she still tries to save his life at great risk to her own. Isn't this a case of having a reason to go against one's self-interest?

I think that it is such a reason. The principle of rational self-interest seems unduly based on the position that people always act to satisfy their perceived best

interest—a view called psychological egoism, which we will critically examine in a later chapter. Sometimes, we have reasons to do things that go against our perceived self-interest. We find this, for example, when a poor person gives away money to help another poor person; so too with the student who refrains from cheating when she knows that she could easily escape detection. Being faithful, honest, generous, and kind often requires us to act against our own interest.

But you may object to this reasoning by saying, "It is perhaps against our immediate or *short-term* interest to be faithful, honest, generous, or kind; but in the long run, it really is likely to be in our best interest because the moral and altruistic life promises benefits and satisfactions that are not available to the immoral and stingy."

There seems to be merit in this response. The basis of it seems to be a plausible view of moral psychology that stipulates that character formation is not like a bathroom faucet that you can turn on and off at will. To have the benefits of the moral life—friendship, mutual love, inner peace, moral pride or satisfaction, and freedom from moral guilt—one has to have a certain kind of reliable character. All in all, these benefits are very much worth having. Indeed, life without them may not be worth living. Thus, we can assert that for every person (insofar as he or she is rational) the deeply moral life is the best sort of life that he or she can live. Hence, it follows that it is reasonable to develop such a deeply moral character—or to continue to develop it because our upbringing partly forms it for most of us.

Those raised in a normal social context will feel deep distress at the thought of harming others or doing what is immoral and feel deep satisfaction in being moral. For such people, the combination of internal and external sanctions may well bring prudence and morality close together. This situation may not apply, however, to people not brought up in a moral context. Should this dismay us? No. As Gregory Kavka says, we should not perceive "an immoralist's gloating that it does not pay him to be moral … as a victory over us. It is more like the pathetic boast of a deaf person that he saves money because it does not pay him to buy opera records."[7] The immoralist is a Scrooge who takes pride in not having to buy Christmas presents because he has no friends.

The Modified Principle of Rational Self-Interest

We want to say, then, that the choice of the moral point of view is not an arbitrary choice but a rational one. Some kinds of lives are better than others: A human life without the benefits of morality is not an ideal or fulfilled life; it lacks too much that makes for human flourishing. The occasional acts through which we sacrifice our self-interest within the general flow of a satisfied life are unavoidable risks that reasonable people will take. Although you can lose by betting on morality, you are almost certain to lose if you bet against it.

Therefore, the principle of rational self-interest must be restated in a modified form:

> *Modified principle of rational self-interest.* If it is reasonable to choose a life plan L, which includes the possibility of doing act A, then it must be in

> my interest (or at least not against it) to choose L, even though A itself may not be in my self-interest.

Now there is no longer anything paradoxical in doing something not in one's interest because, although the individual moral act may occasionally conflict with one's self-interest, the entire life plan in which the act is embedded and from which it flows is not against the individual's self-interest. For instance, although you might be able to cheat a company or a country out of some money that would leave you materially better off, it would be contrary to the *form of life* to which you have committed yourself and that has generally been rewarding.

Furthermore, character is important and habits force us into predictable behavior. Once we obtain the kind of character necessary for the moral life—once we become *virtuous*—we will not be able to turn morality on and off like a faucet. When we yield to temptation, we will experience alienation in going against this well-formed character. The guilt will torment us, greatly diminishing any ill-gotten gains.

The modified principle of rational self-interest answers several moral questions raised throughout this chapter: Should I act immorally if I wear the ring of Gyges? Should I break the social contract if I can get away with it? The answer in both cases is no. First, it is sometimes reasonable to act morally even when those actions do not immediately involve our self-interest. Second, and more important, a life without spontaneous and deliberate moral kindness may not be worth living. This helps explain why Carl and his employees at the car dealership should behave morally, even if it means risking fewer sales with less profit. If they adopted a moral form of life that's not overburdened with a desire for private financial gain, they may feel more rewarded in their business lives by not cheating their customers.

Of course, there's no guarantee that morality will produce success and happiness. Jim—the moral yet unsuccessful brother discussed earlier in this chapter—is not happy. In a sense, morality is a rational gamble. It doesn't guarantee success or happiness. Life is tragic. The good fail and the bad—the Jacks of life—seem to prosper. Yet the moral person is prepared for this eventuality. John Rawls sums up the vulnerability of the moral life this way:

> A just person is not prepared to do certain things, and so in the face of evil circumstances he may decide to chance death rather than to act unjustly. Yet although it is true enough that for the sake of justice a man may lose his life where another would live to a later day, the just man does what all things considered he most wants; in this sense he is not defeated by ill fortune, the possibility of which he foresaw. The question is on a par with the hazards of love; indeed, it is simply a special case. Those who love one another, or who acquire strong attachments to persons and to forms of life, at the same time become liable to ruin: their love makes them hostages to misfortune and the injustice of others. Friends and lovers take great chances to help each other; and members of families willingly do the same.... Once we love we are vulnerable.[8]

We can, however, take steps to lessen the vulnerability by working together for a more moral society, by bringing up our children to have keener moral

sensitivities and good habits so that there are fewer Jacks around. We can establish a more just society so that people are less tempted to cheat and more inclined to cooperate, once they see that we are all working together for a happier world, a mutual back-scratching world, if you like. In general, the more just the political order, the more likely it will be that the good will prosper, and the more likely that self-interest and morality will converge.

CONCLUSION

In this chapter, we've examined social contract theory's explanation of moral motivation as expressed in two questions: "Why does society need moral rules?" and "Why should I be moral?" Hobbes argues that because humans always act out of perceived self-interest people are naturally driven into conflict with everyone—the state of nature. The solution is for us to create a social contract: By giving up some of our liberty and adopting moral rules, we gain peace. Thus, the answer to the first question ("Why does society need moral rules?") is that morality is a much needed mechanism of social control.

Social contract theory's answer to the second question ("Why should I be moral?") is more complicated as the game Cooperate or Cheat shows. Ultimately, I should be moral because, by occasionally allowing some disadvantage for myself, I may obtain an overall, long-term advantage. Even when it seems as though I can break moral rules without getting caught, I still need to consistently follow them because, although an individual moral act may sometimes be at odds with my self-interest, the complete moral form of life in which the act is rooted is not against my self-interest.

NOTES

1. Donald McCabe, "Cheating in Academic Institutions: A Decade of Research," *Ethics & Behavior* 11, no. 3 (2001): 219–232; "Cheating on Spouse or Taxes Morally Acceptable for Many" in LiveScience (www.livescience.com), March 28, 2006; James Vicini, "US Has the Most Prisoners in the World," Reuters, December 9, 2006.
2. Thomas Hobbes, *Leviathan* (1642). A recent edition is by Edwin Curly (Indianapolis: Hackett, 1994). All quoted material in this section is from this work.
3. William Golding, *Lord of the Flies* (New York: Putnam, 1954). All quoted material in this section is from this work.
4. Robert Axelrod, *The Evolution of Cooperation* (New York: Basic Books, 1984); Robert Axelrod and William Hamilton, "The Evolution of Cooperation," *Science* 211 (1981): 1390–1396. The game of Cooperate or Cheat is sometimes called the Iterated Prisoner's Dilemma.
5. David Gauthier, "Morality and Advantage," *Philosophical Review* 76 (1967): 460–475.

JÜRGEN HABERMAS

Discourse Ethics

Jürgen Habermas, a German philosopher born in 1929, is a member of the Frankfurt School and a proponent of critical theory. Started by a group of German sociologists, political scientists, and philosophers between the world wars, critical theory is an interdisciplinary analysis of social, economic, and cultural phenomena. The founders of critical theory borrowed heavily from Marx and Freud, and rejected all forms of irrationality.

Habermas develops a theory of rationality that is particularly sensitive to social and ethical issues. He criticizes theories that reject moral truths; he sees such theories as not sufficiently protecting us from social domination and repression. By modifying some ideas from Kant and G.H. Mead, Habermas develops an approach based on the conditions of rational discussion and intersubjective agreement.

As you read the selection, ask yourself how well Habermas characterizes the conditions for rational discussion. How plausible it is to justify moral norms on the basis of argument and agreement alone?

Universalization

In what follows, I presuppose that a theory of [moral] argumentation must take the form of an "informal logic," because it is impossible to *force* agreement on theoretical and moral-practical issues either by means of deduction or on the basis of empirical evidence.

In theoretical discourse the gap between particular observations and general hypotheses is bridged by some canon or other of induction. An analogous bridging principle is needed for practical discourse. Accordingly, all studies of the logic of moral argumentation end up having to introduce a moral principle as a rule of argumentation that has a function equivalent to the principle of induction in the discourse of the empirical sciences.

Interestingly enough, in trying to identify such a moral principle, philosophers of diverse backgrounds always come up with principles whose basic idea is the same. *All* variants of cognitivist ethics[1] take their bearings from the basic intuition contained in Kant's categorical imperative. What I am concerned with here is not the diversity of Kantian formulations but their under-lying idea, which is designed to take into account the impersonal or general character of valid universal commands. The moral principle is so conceived as to exclude as invalid any norm that could not meet with the qualified assent of all who are or might he affected by it. This bridging principle, which makes consensus possible, ensures that only those norms are accepted as valid that express a *general will*. As Kant noted time and again, moral norms must be suitable for expression as "universal laws." He focuses on "that inner contradiction which promptly arises for an agent's maxim when his behavior can lead to its desired goal only upon the condition that it is not universally followed."

The principle of universalization is by no means exhausted by the requirement that moral norms must take the *form* of unconditionally universal "ought" statements. The *grammatical form* of normative statements alone, which does not permit such sentences to refer to or be addressed to particular groups or individuals, is not a sufficient condition for valid moral commands, for we could give such universal form to commands that are plainly immoral. What is more, in some respects the requirement of formal universality may well be too restrictive; it may make sense to submit nonmoral norms of action (whose range of jurisdiction is socially and spatiotemporally limited) to a practical discourse (restricted in this case to those affected and hence relative), and to test them for generalizability.

Other philosophers subscribe to a less formalistic view of the consistency required by the principle of universality. Their aim is to avoid the contradictions that occur when equal cases are treated unequally and unequal ones equally. R.M. Hare has given this requirement the form of a semantic postulate. As we do when we attribute descriptive predicates ("is red"), so we should attribute normative predicates ("is of value," "is good," "is right") in *conformity with a rule*, using the same linguistic expression in all cases that are the same in the respects relevant to the particular case. Applied to moral norms, Hare's consistency postulate comes to this: every individual, before making a particular norm the basis for his moral judgment, should test whether he can advocate or "will" the adoption of this norm by every other individual in a comparable situation. This or another similar postulate is suitable to serve as a moral principle only if it is conceived as a warrant of impartiality in the process of judging. But one can hardly derive the meaning of impartiality from the notion of consistent language use.

Kurt Baier and Bernard Gert come closer to this meaning of the principle of universalization when they argue that valid moral norms must be generally teachable and publicly defendable. The same is true of Marcus Singer when he proposes the requirement that norms are valid only if they ensure equality of treatment. The intuition expressed in the idea of the generalizability of maxims intends something more than this, namely, that valid norms must *deserve* recognition by *all* concerned. True impartiality pertains only to that standpoint from which one can generalize precisely those norms that can count on universal assent because they perceptibly embody an interest common to all affected. It is these norms that deserve intersubjective recognition. Thus the impartiality of judgment is expressed in a principle that constrains *all* affected to adopt the perspectives of *all others* in the balancing of interests. The principle of universalization is intended to compel the *universal exchange of roles* that G.H.Mead called "ideal role taking" or "universal discourse." Thus every valid norm has to fulfill the following condition:

> (U) *All* affected can accept the consequences and the side effects its *general* observance can be anticipated to have for the satisfaction of *everyone's* interests (and these consequences are preferred to those of known alternative possibilities for regulation).

Discourse ethics

We should not mistake this principle of universalization (U) for the following principle, which already contains the distinctive idea of an ethics of discourse.

> (D) Only those norms can claim to be valid that meet (or could meet) with the approval of all affected in their capacity *as participants in a practical discourse.*

This principle of discourse ethics (D), to which I will return after offering my justification for (U), already *presupposes* that we *can* justify our choice of a norm. At this point in my argument, that presupposition is what is at issue. I have introduced (U) as a rule of argumentation that makes agreement in practical discourses possible whenever matters of concern to all are open to regulation in the equal interest of everyone. Once this bridging principle has been justified, we will be able to make the transition to discourse ethics. I have formulated (U) in a way that precludes a monological application of the principle. First, (U) regulates only argumentation among a plurality of participants; second, it suggests the perspective of real-life argumentation, in which all affected are admitted

as participants. In this respect my universalization principle differs from the one John Rawls proposes.

Rawls wants to ensure impartial consideration of all affected interests by putting the moral judge into a fictitious "original position," where differences of power are eliminated, equal freedoms for all are guaranteed, and the individual is left in a condition of ignorance with regard to the position he might occupy in a future social order. Like Kant, Rawls operationalizes the standpoint of impartiality in such a way that every individual can undertake to justify basic norms on his own. The same holds for the moral philosopher himself. It is only logical, therefore, that Rawls views the substantive parts of his study, not as the *contribution* of a participant in argumentation to a process of discursive will formation regarding the basic institutions of late capitalist society, but as the outcome of a "theory of justice," which he as an expert is qualified to construct.

If we keep in mind the action-coordinating function that normative validity claims play in the communicative practice of everyday life, we see why the problems to be resolved in moral argumentation cannot be handled monologically but require a cooperative effort. By entering into a process of moral argumentation, the participants continue their communicative action in a reflexive attitude with the aim of restoring a consensus that has been disrupted. Moral argumentation thus serves to settle conflicts of action by consensual means. Conflicts in the domain of norm-guided interactions can be traced directly to some disruption of a normative consensus. Repairing a disrupted consensus can mean one of two things: restoring intersubjective recognition of a validity claim after it has become controversial or assuring intersubjective recognition for a new validity claim that is a substitute for the old one. Agreement of this kind expresses a *common will*. If moral argumentation is to produce this kind of agreement, however, it is not enough for the individual to reflect on whether he can assent to a norm. It is not even enough for each individual to reflect in this way and then to register his vote. What is needed is a "real" process of argumentation in which the individuals concerned cooperate. Only an intersubjective process of reaching understanding can produce an agreement that is reflexive in nature; only it can give the participants the knowledge that they have collectively become convinced of something.

From this viewpoint, the categorical imperative needs to be reformulated as follows: "Rather than ascribing as valid to all others any maxim that I can will to be a universal law, I must submit my maxim to all others for purposes of discursively testing its claim to universality. The emphasis shifts from what each can will without contradiction to be a general law, to what all can will in agreement to be a universal norm." This version of the universality principle does in fact entail the idea of a

cooperative process of argumentation. For one thing, nothing better prevents others from perspectivally distorting one's own interests than actual participation. It is in this pragmatic sense that the individual is the last court of appeal for judging what is in his best interest. On the other hand, the descriptive terms in which each individual perceives his interests must be open to criticism by others. Needs and wants are interpreted in the light of cultural values. Since cultural values are always components of intersubjectively shared traditions, the revision of the values used to interpret needs and wants cannot be a matter for individuals to handle monologically.

The rules behind an ideal-guided moral discourse

We must return to the justification of the principle of universalization. We are now in a position to specify the role that the transcendental-pragmatic argument can play in this process. Its function is to help to show that the principle of universalization, which acts as a rule of argumentation, is implied by the presuppositions of argumentation in general. This requirement is met if the following can be shown:

> Every person who accepts the universal and necessary communicative presuppositions of argumentative speech and who knows what it means to justify a norm of action implicitly presupposes as valid the principle of universalization, whether in the form I gave it above or in an equivalent form.

It makes sense to distinguish three levels of presuppositions of argumentation along the lines suggested by Aristotle: those *at* the logical level of products, those at the dialectical level of procedures and those at the rhetorical level of processes. First, reasoning or argumentation is designed to *produce* intrinsically cogent arguments with which we can redeem or repudiate claims to validity. This is the level at which I would situate the rules of a minimal logic currently being discussed by Popperians, for example, and the consistency requirements proposed by Hare and others. For simplicity I will follow the catalog of presuppositions of argumentation drawn up by R.Alexy. For the logical-semantic level, the following rules can serve as *examples:*

(1.1) No speaker may contradict himself.
(1.2) Every speaker who applies predicate *F* to object *A* must be prepared to apply *F* to all other objects resembling *A* in all relevant aspects.
(1.3) Different speakers may not use the same expression with different meanings.

The presuppositions of argumentation at this level are logical and semantic rules that have no ethical content. They are not a suitable point of departure for a transcendental-pragmatic argument.

In *procedural* terms, arguments are processes of reaching understanding that are ordered in such a way that proponents and opponents, having assumed a hypothetical attitude and being relieved of the pressures of action and experience, can test validity claims that have become problematic. At this level are located the pragmatic presuppositions of a special form of interaction, namely everything necessary for a search for truth organized in the form of a competition. Examples include recognition of the accountability and truthfulness of all participants in the search. At this level I also situate general rules of Jurisdiction and relevance that regulate themes for discussion, Contributions to the argument, etc. Again I cite a few examples from Alexy's catalog of rules:

(2.1) Every speaker may assert only what he really believes.
(2.2) A person who disputes a proposition or norm not under discussion must provide a reason for wanting to do so.

Some of these rules obviously have an ethical import. At this level what comes to the fore are presuppositions common both to discourses and to action oriented to reaching understanding as such, e.g., presuppositions about relations of mutual recognition.

But to fall back here directly on the basis of argumentation in action theory would be to put the cart before the horse. Yet the presuppositions of an unrestrained competition for better arguments are relevant to our purpose in that they are irreconcilable with traditional ethical philosophies that have to protect a dogmatic core of fundamental convictions from all criticism.

Finally, in *process* terms, argumentative speech is a process of communication that, in light of its goal of reaching a rationally motivated agreement, must satisfy improbable conditions. In argumentative speech we see the structures of a speech situation immune to repression and inequality in a particular way: it presents itself as a form of communication that adequately approximates ideal conditions. This is why I tried at one time to describe the presuppositions of argumentation as the defining characteristics of an ideal speech situation. I cannot here undertake the elaboration, revision, and clarification that my earlier analysis requires, and accordingly, the present essay is rightly characterized as a sketch or a proposal. The intention of my earlier analysis still seems correct to me, namely the reconstruction of the general symmetry conditions that every competent speaker who believes he is engaging in an argumentation must presuppose as adequately fulfilled. The pre-supposition of something like an "unrestricted communication

community," an idea that Apel developed following Peirce and Mead, can be demonstrated through systematic analysis of performative contradictions. Participants in argumentation cannot avoid the presupposition that, owing to certain characteristics that require formal description, the structure of their communication rules out all external or internal coercion other than the force of the better argument and thereby also neutralizes all motives other than that of the co-operative search for truth.

Following my analysis, R.Alexy has suggested the following rules of discourse for this level:

(3.1) Every subject with the competence to speak and act is allowed to take part in a discourse.

(3.2)

a. Everyone is allowed to question any assertion whatever.
b. Everyone is allowed to introduce any assertion whatever into the discourse.
c. Everyone is allowed to express his attitudes, desires, and needs.

(3.3) No speaker may be prevented, by internal or external coercion, from exercising his rights as laid down in (3.1) and (3.2).

A few explanations are in order here. Rule (3.1) defines the set of potential participants. It includes all subjects without exception who have the capacity to take part in argumentation. Rule (3.2) guarantees all participants equal opportunity to contribute to the argumentation and to put forth their own arguments. Rule (3.3) sets down conditions under which the rights to universal access and to equal participation can be enjoyed equally by all, that is, without the possibility of repression, be it ever so subtle or covert.

If these considerations are to amount to more than a definition favoring an ideal form of communication and thus prejudging everything else, we must show that these rules of discourse are not mere *conventions;* rather, they are inescapable presuppositions. The presuppositions themselves are identified by convincing a person who contests the hypothetical reconstructions offered that he is caught up in performative contradictions.

Study questions

1 How does Habermas define a "bridging principle"? Why are such principles needed? What is Habermas's bridging principle for ethics?

2 What is cognitivist ethics? How does Habermas argue that Kantian ethics is the ground for all cognitivist ethics?
3 Explain Habermas's principles (U) and (D). How does (U) bring in consequences?
4 For Habermas, what is the significance of "recognition" for ethics? How does recognition impact the formulation of (U)?
5 Why does Habermas avoid what he calls "monological application" of (U)? Whom would he accuse of monological application?
6 In what ways are the approaches of Habermas and Rawls similar? How does Habermas criticize Rawls's approach?
7 What is Habermas's "unrestricted communication community"? How do participants in argumentation discover it?
8 How does Habermas guarantee that the rules for discourse are not merely arbitrary conventions?

For further study

This selection has excerpts, sometimes simplified in wording, from Jürgen Habermas's "Discourse ethics: Notes on a program of philosophical justification," in *Moral Consciousness and Communicative Action* (Cambridge, MA: MIT Press, 1990), translated by C.Lenhardt and S. Weber Nicholsen. For other ethical writings, see his *Legitimation Crisis* (Boston: Beacon Press, 1975), translated by T.McCarthy—especially Part III; *Justification and Application: Remarks on Discourse Ethics* (Cambridge, MA: MIT Press, 1994), translated by C.Cronin—especially Chapters 1 and 2, which respond to criticisms; and *Between Facts and Norms,* (Cambridge, MA: MIT Press, 1996), translated by W. Rehg—especially Chapter 3. For analysis and criticism of Habermas's discourse ethics, see James Swindal's *Reflection Revisited: Jürgen Habermas's Emancipative Theory of Truth* (New York: Fordham University Press, 1999), Chapters 4 and 5; William Rehg's *Insight and Solidarity* (Berkeley: University of California Press, 1994); Seyla Benhabib's *Critique Norm and Utopia* (New York: Columbia University Press, 1986); Benhabib's edited collection of articles on discourse ethics in *The Communicative Ethics Controversy* (Cambridge, MA: MIT Press, 1990); Nancy Fraser's "What's critical about critical theory?: The case of Habermas and gender," in *Unruly Practices* (Minneapolis: University of Minnesota Press, 1989); and Joseph Heath's "The Problem of Foundationalism in Habermas's Discourse Ethics," *Philosophy and Social Criticism* 21 (1995): 77– 100.

Related readings in this volume include Hare, Kant, Nagel, O'Neill, and Sartre (who also develop principles of universalization); Frankena and Rawls (who provide somewhat related views about the method to be