Morality as a social contract

If we expect people to be moral, we need to have some idea of why they would. So we need to know why people do things. One obvious answer is ‘self-interest’ – people will do what they think will benefit them in some ways. So one answer to the question ‘why should I be moral?’ is that, in some way, being moral is in my self-interest. In other words, I will benefit from being a moral person and acting morally. We can develop this answer to argue that morality is, in fact, an agreement about how to behave, an agreement that people have reached because they realise that certain ways of behaving, which we call moral, are in their self-interest.

This view understands morality as a means to an end; the end, for each person, is doing what is best for themselves. However, if you ask people ‘why did you do that?’, they don’t always cite some benefit to themselves. They may give some other reason, e.g. ‘I thought it would help him’. This raises the questions – what reasons do we act on and what reasons should we act on?

Philosophers have given different theories of reasons, some arguing that rationality is just a matter of pursuing your self-interest intelligently, others arguing that we have reasons to act in ways that aren’t about self-interest.

SELF-INTEREST AND RATIONAL EGOISM

The answer, that one should be moral because it is in one’s self-interest, has the advantage of simplicity. First, it is obvious that individuals are motivated by their self-interest. Second, this doesn’t need justifying: it would be strange to ask the question ‘why should I do what is in my self-interest?’. The basic desires to stay alive and stay free of pain and the more complex desire for happiness are part of human nature.

Morality can require that, at times, we give up something we want for ourselves for the sake of someone else. But we cannot assume that people are interested – or as interested – in other people’s well-being as they are in their own. Perhaps we would help each other when it is no cost to ourselves; but if there is a competition between getting what one wants and helping others, we cannot assume that people will sacrifice their self-interest and be altruistic.

What is ‘reasonable’ here? There is a particular view of what it is rational to do, usually adopted in economics and politics when trying to predict what people will do. If people are motivated by self-interest, then it is rational for them to do what benefits themselves. A rational person is someone who selects the means to their end of self-interest. They will consider what they need to do to get what will benefit themselves, and then, if they can, do it. A simple example: I’m thirsty, I want a drink (my end). I know where to find water, so I go there (my means) and get the water.

This ability to work out the right means to that end defines what rationality is, on this view. People are ‘rational egoists’.
MORALITY AS AN AGREEMENT

Being rational, we can see that it is very much in our self-interest that other people do not harm us – physically, or emotionally, or financially – when pursuing their self-interest. It would be good if there were constraints on what other people did, constraints that they followed. On the other hand, it looks like it is not such a good thing to be constrained oneself. If by stealing people’s wallets I could get rich without much work, this looks like a good means to benefiting myself. But if I am constrained not to steal, then this easy path to wealth is no longer an option. So perhaps I ended up poorer, which is not in my self-interest.

We can see how this problem could be solved: we need to agree not to do things that would harm other people in exchange for them not doing things that harm us. While that means that my self-interest will suffer a little, it would be much worse for me if other people harmed me.

Another example: suppose I want a house to live in, but I can’t build one on my own. I need the cooperation of other people to help. It would be a very good thing if I could trust other people to cooperate if they say they will. Perhaps someone agrees to help me build my house if I first help him build his. Despite the extra work for me, I decide it is worth it – after all, having a house is better than having no house, even if I have to build two houses – his and mine – to get my house. But can I trust him to help me build my house when we have finished his? From his point of view, it is a lot of work to help me build mine, and he gains nothing because he already has his house! I might decide that, because he is self-interested, I shouldn’t trust him. So I don’t help him build his house. So we both end up without houses.

Again, we can see how to overcome this: we need to be able to trust other people to keep their word, so that we cooperate together on projects that will benefit us both.

Agreeing to live by the rules of morality – and then actually doing so! – solves the problems that acting on rational self-interest raises. Knowing that someone else will act morally means that we do not need to fear them harming us or breaking their word. The benefits this brings are very large, but the costs of signing up to the agreement ourselves are, by comparison, relatively small. There are certain ways in which we can no longer pursue our self-interest, but we are protected from harm and able to achieve more benefits for ourselves through cooperation than we could achieve alone.

Of course, it is only worth signing up to if it does in fact produce cooperation. But people will only cooperate (willingly) if they feel that the agreement is fair. Being self-interested, no one wants to sign up to an agreement which benefits other people more than it benefits them – because then they will be at a relative disadvantage. So the agreement must be fair – not because otherwise it would be immoral, but because otherwise it wouldn’t work.

THE PRISONER’S DILEMMA

If everyone acts on their self-interest, it may seem that we will each do the best for ourselves. But we can show that this is false, and that one great advantage of morality, for everybody, is that it creates trust and cooperation. The Prisoner’s Dilemma is a
A fictional scenario invented in the 1950s. Suppose two men, Adam and Barry, are arrested for a crime, and held in separate rooms for interrogation. The prosecutor only has enough evidence to charge them with illegally possessing a gun, but if he can get one of them to confess, he can charge them with armed robbery. Adam is told this: ‘If you confess, your sentence will be reduced because you have helped the prosecution. In fact, if you confess and Barry doesn’t, we’ll let you go free, but Barry will serve 10 years in prison for armed robbery. If Barry confesses and you don’t, you will serve 10 years in prison, and he’ll go free. If you both confess, you’ll get a lighter sentence of 7 years each. If neither of you confess, then you’ll be charged with illegally possessing a gun, and you’ll both get 2 years in prison.’ Barry is given the same deal.

So there are four options, which we can display in a table:

<table>
<thead>
<tr>
<th></th>
<th>Barry confesses</th>
<th>Barry doesn’t confess</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adam confesses</td>
<td>Both get 7 years</td>
<td>Adam is free, Barry gets 10 years</td>
</tr>
<tr>
<td>Adam doesn’t confess</td>
<td>Adam gets 10 years, Barry is free</td>
<td>Both get 2 years</td>
</tr>
</tbody>
</table>

What should Adam do? If Barry confesses, then it is better that Adam confesses as well (he’ll get 7 years rather than 10). If Barry doesn’t confess, then once again it is better than Adam confesses (he’ll go free rather than get 2 years). So, from a purely self-interested point of view, Adam should confess.

Barry is thinking the same thing. So Barry also confesses. Because they both confess, they get 7 years. But it would have been better for them both not to confess, and only get 2 years each. If both of them act rationally from a self-interested point of view, the result is worse than if they had trusted each other to act in a way that benefited them both!

This clearly shows that acting only on one’s self-interest can lead to worse consequences for oneself than cooperating with others, even when there is a cost to cooperation. For Adam and Barry, not confessing meant giving up the possibility of freedom. But if they could trust each other to do this, they would both be better off (getting 2 years) than they are in a situation in which they cannot trust each other and they both confess (getting 7 years).

**THE ‘FREE RIDER’ PROBLEM**

The Prisoner’s Dilemma shows that cooperating can be better than everyone acting out of self-interest. Suppose that, realising the benefits of morality, we agree with each other to act morally. Now we know that, by and large, we can trust each other.

Suppose now that Barry doesn’t confess, trusting that Adam won’t confess either. Adam suspects that Barry trusts him not to confess – so now the options Adam has to choose between are either not confessing, and getting 2 years, or confessing and walking free. This means that if Adam has good reason to think that Barry trusts him, he is better off acting self-interestedly.

Everyone acting morally is better than everyone acting self-interestedly. But if everyone else is acting morally, it is even better to act self-interestedly – at least if you can get away with it. This is called the ‘free rider’ problem – someone who does this gets the benefits
It turns out that whether I have reason to conform to the expectations of morality depends on whether I can get away with acting immorally when other people are acting morally. If I can, then I have more reason to act immorally than to act morally.

Morality as an agreement is in everyone’s self-interest, collectively, to set up. But once it is set up, it is in each person’s self-interest, individually, to get away with breaking the agreement, if they can avoid punishment. Is there some way we could describe morality as an agreement that will help solve the free rider problem? This is one of the questions we discuss in the next section.

**EXACTLY WHAT KIND OF AGREEMENT COULD MORALITY BE?**

It may seem beside the point to talk about morality as an agreement – it clearly isn’t. We have never agreed to be moral, and nobody ever asked us.

**A tacit agreement**

Some philosophers respond that morality can be understood as a tacit – an unspoken – agreement. This view is defended by people who want to explain why we have the particular moral practices we do. So, for example, why has there been a ‘double standard’ about sex? In many societies, it is seen as morally bad for a woman to sleep with many men, but more acceptable for a man to sleep with many men. This seems very unfair. We can explain it if we think of our moral code as a tacit agreement between people who had or have power, in this case, men.

If morality is this kind of agreement based on power, then it is not an agreement for the mutual advantage of everyone. So you only have reason to be moral if you are one of the powerful people morality benefits. So if we take a realist approach to morality being an agreement, it may turn out that we don’t have reason to be moral.

**A hypothetical agreement**

We want to show why it is rational to be moral, where being moral is mutually advantageous. To show that it is rational, we don’t need to show that morality is an actual agreement, only that if we could make such an agreement, we should. Morality is a ‘hypothetical’ agreement, an agreement we would or should make because it is rational to do so. The point of looking at it this way is not to explain our moral code as it is, but to justify morality.

**Gauthier**

To show that it is rational to conform to the expectations of morality, David Gauthier argues that the situation without morality is like the Prisoner’s Dilemma – even if we try to cooperate with each other, because we are self-interested and can’t trust them, we will both end up in a situation which isn’t as good as it could be. This will motivate us to agree to morality.

To argue that it is rational for me to be moral, I need to imagine me in that situation; and the same for you. But we know that people are very different in power and ability – perhaps it would be more rational for powerful people to agree among themselves to enforce a morality that isn’t equally in everyone’s interests. However, Gauthier argues,
this situation will be unstable – the people who are not treated equally could threaten to upset the agreement. A stable agreement must be one in which no one feels coerced or cheated.

For the argument to work, individuals must be the best judges of what is in their self-interest and how to achieve this. Suppose that I thought I always made bad decisions – I might, then, feel it was rational for me to ask other people to decide what I should do. I could choose to submit to morality for this reason. But this isn’t how Gauthier understands the agreement – each person agrees because they think it will be the best for them.

**What do we agree to?**

We saw above that thinking of the agreement to be moral in terms of the Prisoner’s Dilemma leads to the free rider objection. If I am motivated by self-interest to agree to morality, then presumably I am still self-interested after making the agreement. But my self-interest will then lead me to be immoral when this is in my self-interest and I can get away with it. Realizing this, will we really trust other people? And if we don’t, then we haven’t got an agreement to be moral at all.

Gauthier argues that we can solve this problem if what we agree to is not simply to act morally, but to change what motivates us. We agree to adopt a new disposition, the disposition to be moral. In other words, we agree to become people who will not act on self-interest when this conflicts with acting morally. (Perhaps it makes little sense to talk about choosing a motivation. But Gauthier is not trying to tell a realistic story, but justify morality. What he argues is that it is rational to have this disposition to be moral.)

We can still object, however, that if our motivation to be moral rests ultimately on self-interest, it will not be strong enough to get us to act morally when this conflicts with self-interest. Human psychology doesn’t respond as well to these arguments about justification as it does to immediate self-interest. If this is right, then Gauthier’s story about why we ought to be moral cannot tell us why we are moral.

**TWO OBJECTIONS**

**Who’s in?**

A further objection to the view that morality is an agreement for mutual advantage is that it is only rational to make this agreement with people from whom you can benefit. This is in danger of leaving out some people, e.g. people with disabilities. If the cost of treating them morally is greater than the benefits that result from their cooperation, then it is not in our self-interest to include them in the agreement. But then we have no reason to treat people with disabilities well.

One possible response is that including them in the agreement is a kind of insurance – I could become disabled, and then the cost of being left outside the agreement would be very high. So I have reason to make sure the disabled are covered by the agreement to act morally, since I might be disabled one day.

This reply won’t work for other cases, though. Does morality cover how we treat animals or the environment? We cannot make an agreement with animals or the environment. We can, of course, make an agreement with other people to treat animals and the environment in certain ways. But why would we? What do I gain from how people treat...
animals? What do I gain from people not exploiting or polluting the environment (as long as it isn’t near me)?

**Morality before the agreement**

If morality is the agreement we make, then there are no moral rules for making the agreement. Gauthier agrees with this – the only reason the agreement needs to treat people equally is that it is more likely to break down if it doesn’t. Is this right? Life goes on despite a great deal of unfairness. If I think that I will do better ignoring morality when it suits me, it is not obvious that I am wrong.

Rather than arguing that the agreement is based just on self-interest, we could say that the agreement is itself a *moral* agreement. It expresses what morality is about, and signing up to the agreement is an expression of our commitment to morality, not an expression of self-interest.

This view is defended by Thomas Scanlon. He uses the idea of agreement to explain what morality is:

> An act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behaviour that no one could reasonably reject as a basis for informed, unforced, general agreement’. (What We Owe to Each Other, p. 153)

If I behave immorally, I am acting in a way that I cannot reasonably expect other people to accept. The motivation for being moral, then, is not self-interest, but wanting to be able to *justify* our behaviour to each other.