Lecture 10:
Foundationalism, Coherentism and Reflective Equilibrium Theory

In previous lectures, we were primarily concerned with what the “moral game” was all about. Is it a matter of detecting non-natural properties as Moore thought, or intuiting moral truths that are as much a part of the fabric of the universe as geometrical truths as Ross thought? Is it a matter of expression and communication of emotion as Stevenson thought, or of issuing universal prescriptions as Hare thought? Or perhaps it is the game of trying to figure out what an ideal observer would approve of? We are now going to move to a different topic: the nature of reasoning in ethics. Of course, we have not totally neglected this field. Each of the metaethical theories we have looked at so far has something to say about moral reasoning. For example, Stevenson thinks that the main role for reason in moral thinking is to determine the facts of the case, and that once these are determined, reason runs out and all that is left is non-rational methods of persuading people to share one's attitudes. Hare, on the other hand, thinks that reasoning in ethics can take us much further. It involves “golden rule” arguments which, he thinks, have some hope of bringing about rational moral convergence. But what if we are not persuaded of the plausibility of any of these metaethics, or of any general view about the nature of the moral game? Where does that leave us? Fortunately, twentieth century moral philosophy provides an answer to that question, something called “reflective equilibrium” methods which were developed by John Rawls.

Before turning to reflective equilibrium theory, it will be useful to start with some general epistemology. This is because reflective equilibrium methods are generally seen as a variant of an approach to justification developed outside of ethics and which Rawls applied to ethics, coherentism. I will therefore begin with an introduction to two classic views on the structure of
justification, coherentism and foundationalism.

Foundationalism. Foundationalism is the view that knowledge has a two part structure. This structure is caught by the analogy of a building sitting on a foundation. The foundation supports the building and the foundation is, in turn, supported by nothing but itself. (Or more precisely, it is supported by the earth which is, in turn, supported by nothing -- but what the heck, it is only an analogy.) The foundationalist believes that knowledge has a similar structure. There are less basic and more basic beliefs. The less basic, the superstructure, are supported by the more basic, and the most basic are foundational, that is, they are in some way self-supporting or without need of support. In the building analogy, the support relation is physical support. In the case of knowledge, the support relation is justification. Thus, less basic beliefs are justified in terms of more basic beliefs, which are justified by most basic beliefs which are not in need of external justification, perhaps because they are self-justifying in some way.¹

According to foundationalism, for a belief to be justified, it must be either a foundational belief that is not in need of external justification or a superstructural belief, one that obtains justification in terms of other beliefs. Many foundationalists have supported this view about the structure of justification with the following argument. Suppose I have belief 1 which needs justification. It will be justified in terms of belief 2, which is, in turn, justified in terms of belief 3, and so on. There are three possibilities. First, this chain of justifying beliefs might go on forever. But if so, I never get to the point where I actually justify belief 1. Second, the chain of justifications might loop back into a circle: 1 is justified in terms of 2 which is justified in terms of 3, . . . , which is justified in terms of 1. But a circular justification is no justification at all. So if we are ever to really justify 1, a third possibility must be true: the chain of justifications must
stop at some belief that does not need external justification, at a foundation. So, foundationalists argue, if we are to have justified beliefs at all, foundationalism must be true. Of course, this does not mean any beliefs really are foundational. But, the foundationalist says, if there are none, then there is no knowledge and we are left with skepticism.

Which beliefs are foundational? Putting morality aside for the moment, foundationalists often take the direct testimony of our physical senses, of our eyes, ears, and so on, to yield foundational beliefs. I do not need to prove that John, who was deathly ill a few days before, is now up and running. I simply see it. I can then use this and other sensory beliefs to help justify (or refute) beliefs that cannot be known true or false by direct perception, for example, the belief that a certain medicine is effective in curing John’s illness. I have some (though not conclusive) reason to believe it is effective because I saw John sick, I saw John take the medicine, and I then saw John up and running. Again, I might take as foundational certain mathematical beliefs such as that the shortest distance between two points is a straight line. I can then use these foundational beliefs to prove or refute non-foundational mathematical beliefs, say, that when a line cuts two parallel lines, the alternate interior angles are equal. Thus, still putting morality aside, traditional foundationalists commonly said that foundational beliefs have at least two sources, our physical senses and some sort of mathematical intuition.

Coherentism. A good way to introduce coherentism is with a problem coherentists find with foundationalism. The problem is that, the coherentist says, there are no foundational beliefs of the sort the foundationalist wants. There are plenty that seem at first glance to work, but on further reflection, they do not. To see what they have in mind, here is a simplified example from the history of astronomy. Suppose that I am a sixteenth or seventeenth century astronomer trying
to construct a model for the sun, the earth, the planets, and so on. There were two models on the


table at that time, the geo-centric model developed by Ptolemy which has the earth at the center


and the sun and other planets going around it, and a helio-centric model championed by


Copernicus which puts the sun at the center and has the planets, including the earth, going

around it. To determine which is correct, I turn to sensory observations. I make a large number


of observations of the sun, the planets, and so on, or borrow ones from Tycho Brahe, and try to


fit them into the models.


Foundationalists might characterize this project as follows. Our observations of the


planets, the stars, and so on, count as our foundation. We do not need to prove that Saturn is in a


particular location in the sky at a particular time. We simply observe it, perhaps with the aid of


instruments like telescopes and, I suppose, sextons. We then use these foundational


observations to establish, or refute, particular models. Using them, I reject the earth-centered


model and accept the sun-centered model. For the sun-centered model is consistent with our


observations while the earth centered model is not.


This foundationalist characterization seems reasonable at first. But coherentists point out


a problem with it. Suppose that after carefully collecting many thousands of observations, we try


to figure out which astronomical model to go with. We find that many of our observations fit an


earth-centered model, but that some do not. The model says that certain celestial objects should


have been in certain places in the sky at certain times, and yet they were not. So we turn to a


sun-centered model. Once again, though we find that this model works pretty well, and that a


great many of our observations fit it, some do not. What should we say now? We could say that


neither model is right and that we should seek another approach. But suppose that no matter
what model we try, some recorded observations just do not fit. If this happens, foundationalists
might say that, since our observations yield foundational beliefs, the fact that none of our models
fits all of them means we have not yet found the right model. But there is another possibility.
We could reason that one of the models we are trying out is likely correct, or nearly correct, but
that some of our observations are faulty. Perhaps we made a mistake measuring, or some odd
viewing conditions made it hard to see accurately where the planets and stars were. If some of
our observations are faulty, of course it will be impossible to find the correct model if we require
that it fit all our observations. If this happens, and it often does in science, we have to decide not
just which astronomical model fits the observations, but also which observations are likely to be
faulty. But how do we do that?

Here is where the coherentist steps in. She says that the mistake that the foundationalist
made was in thinking that one set of beliefs, our observations, were foundational and that all
justification began with them and then moved in a linear fashion to superstructural beliefs. But,
she says, that oversimplifies things. Perhaps, she reasons, we should turn everything on its
head. So far we have been trying to determine which model is correct in terms of fitting our
observations. But instead, if we could figure out which is likely to be the correct model, we
could use it to toss out any observations that conflict with that model. To do this, we need a way
to test our models at least partly independently of fidelity to observation. A popular way to
proceed is in terms of simplicity. On this approach, we should go for simpler, rather than more
complex models. For a variety of reasons, the earth-centered model was more complex than the
sun-centered model. So on the basis of simplicity, we should go with the sun-centered model
and should reject as inaccurate those observations that conflict with it.
What would the coherentist have us learn from this over-simplified bit of scientific 
history? I begin with a lesson specifically about this case, and will then generalize. According 
to the foundationalist, in this case, there are foundational observations which are used to confirm 
or disconfirm astronomical models. Testing goes in one direction, from observations to models. 
But we now see that things are more complex. Just as we test models by their fidelity to 
observation, so do with test observations by their consistency with the best model determined, 
perhaps, by simplicity. When we try to choose between earth-centered and sun-centered models, 
we need to balance a number of factors including observations and simplicity. We seek a model 
that is best in light of all these factors. That a model is simple counts in its favor. That a model 
accounts well for the observational data counts in its favor. No model is likely, in interesting 
cases, to account perfectly for all the observations, and no model is simplest in all respects. 
Further, we can expect that some observations are accurate and some are inaccurate. 
Observations that fit an otherwise preferred model are more likely to be accurate. Those that do 
not are less likely to be accurate. In short, says the coherentist, what we face is a complex puzzle 
which cannot be solved by the relatively simple test that the foundationalist would have us use. 
Rather, we must seek the most coherent whole where we must work with diverse and often 
conflicting factors.

Coherentists believe this lesson is quite general. The coherentist view of justification 
says, roughly, that a belief is justified if and only if it fits into the best coherent whole. This goes 
both for both beliefs foundationalists take to be foundational, and for those foundationalists take 
to be superstructural. It goes for all beliefs. No beliefs are given an absolutely privileged, 
foundational status, not even seemingly straightforward beliefs based on direct sensory
observation. For coherentists, justification is a holistic notion. A belief is not justified as true, or rejected as false, because of itself or because of its relation to a small set of other beliefs. Rather, it is justified if and only if it is contained in the best coherent whole. Of course, as a practical matter, to determine whether or not the sun goes around the earth or the earth around the sun, we cannot wait for a complete theory of everything. So in practice, we try to test large chunks of beliefs rather than complete theories of everything. But the larger the chunk we are dealing with, the better. We do not feel secure in our astronomical model till we can weave it into an extensive theory of celestial mechanics which, in turn, we tie to general laws of physics. The larger the coherent structures that we embed our beliefs in, the more secure they are.

What exactly is involved with coherence? I do not have space to go into this in any detail, but a few words are needed. At a minimum, coherence is logical consistency. If I believe an astronomical model that says a certain planet should be in a certain place, and observation places the planet elsewhere, my beliefs are not consistent and the system is not coherent. Either my astronomical model, or my observation, must be rejected. But coherence requires more than consistency. The belief that the sun will rise tomorrow is consistent with the belief that the blue sofa will be comfortable for Aunt Eloise to sit on, but this does not help to justify my belief that the sun will rise or that Aunt Eloise will like the sofa. Coherence also involves various relations between beliefs. For example, the belief that the sun will rise tomorrow is tightly linked to various beliefs about the planets, about physics, and so on. A thousand beliefs might be consistent with one another, but be unrelated. Such a system is less coherent than one in which they are related as the laws of physics are to the rising of the sun tomorrow.

*Convergence Between Foundationalism and Coherentism.* Foundationalism and
coherentism are very different views about justification. However, as often happens in philosophy, as each view gets modified to deal with problems and make it more plausible, they tend to look more and more like one another. As I am mainly interested in coherentism here, I will focus on a way it might come to seem more like foundationalism and will put aside ways that foundationalism might come to look more like coherentism. I will add a qualification to the coherentist picture that many coherentists accept, one that makes it seem more like foundationalism. This qualification has been particularly important in the literature on moral coherentism. I said coherentism rejects the distinction between foundational and superstructural beliefs. All beliefs are, in principle, open to revision and rejection, and all can benefit from external justification. Justification is given by how well they fit into a coherent whole. However, most of us are uncomfortable with the total rejection of the foundationalist’s distinction. Surely there is something to it. Surely, for example, observations of the locations of the planets are more trustworthy, more foundational in some sense, then our astronomical theories. Surely our observations have some sort of epistemological privilege that our theories lack. And surely, if coherentism completely denies this common sense idea, there is something wrong with coherentism.

Fortunately, coherentists can agree with this. Foundationalists are getting at something real when they distinguish foundational from superstructural beliefs. Coherentists can agree that observational beliefs, and many others that foundationalists talk about, have some sort of privileged status. They merely deny that foundationalists understand this privileged status correctly. Foundationalists want to make such beliefs privileged in a strong way: they have no need of any external justification. Coherentists assign these beliefs a weaker privilege. They can
say that some beliefs foundationalists call “foundational” are what moral philosopher, John Rawls, called “provisionally fixed points.” These are beliefs that are not exactly foundational, but which nevertheless have a kind of privileged status. They are relatively immune to revision in the sense that to reject a provisionally fixed point, a great deal of evidence against it is needed. To reject a relatively unfixed point, on the other hand, relatively little evidence against it is needed.

Why are some beliefs relatively fixed points in this sense? Here is a possible answer. Some beliefs are of great importance in the sense that, if we were to reject them, there would have to be massive revisions in our other beliefs. Those are the relatively fixed points. Other beliefs are not important in the sense that rejecting them would not require great revisions elsewhere. These are the relatively unfixed points. Suppose I saw someone I took to be Jim in the distance. If later, Gordon tells me that he was somewhere else with Jim at that time, I will be quite willing to revise my opinion since I can easily account for my error without significant revisions in my beliefs. All I have to say is that at that distance, and with my vision, it is easy to make a mistake and to confuse Jim with a large rather oddly shaped rock. But suppose that the story were different. I met Jim for lunch, spent a pleasant hour with him, talking about our normal topics, saying our normal things, eating our usual foods. Gordon later tells me that he was somewhere else with Jim at exactly that time. Will I revise my belief that I was with Jim? Should I admit that Gordon was with him at that time and that I was not? Unless there are a lot of good reasons that go beyond Gordon’s word, I would be foolish to change my belief. After all, if I were to say that I was not with Jim, I would have to flesh the story out somehow. For example, I would have to say that I was hallucinating, or confusing real life with a realistic
dream, or that there were some super aliens around who love to impersonate humans and are very good at it. These are possible, but highly unlikely since I am only rarely taken to hallucinating, or confusing dreams with reality, and I have never had any reason to think that super aliens ever impersonate anyone in Bozeman, Montana -- their normal range is Washington D.C. Indeed, it is hard to think of even a reasonably plausible story here except for one: for some reason, Gordon is not telling the truth. Of course, this too might be hard to believe. But it is a lot more reasonable to believe that he lied than to believe I have been subject to such realistic hallucinations or cosmic scams.

The belief that I saw Jim, when the person I saw was in the distance, is not a relatively fixed point. It does not take much counter-evidence to overthrow it. The belief that I saw Jim when we had lunch together is a relatively fixed point. It would be unreasonable to give the belief up without a huge amount of evidence. To give up the former belief would require minimal changes in my belief system while to give up the latter would involve massive reformations. In the same way, no single scientific observation, say an observation of the location of a planet, yields a particularly fixed point. If a single, or small number, of observations conflict with a very good astronomical theory, I should often reject the observations and not the theory since there are all sorts of ways that observations can be screwed up. But the mass of my observations as a whole constitute quite a fixed point. For the only way that huge numbers of them can be faulty is if the world were radically different from what we think. Perhaps we would have to give up our usual theory of optics, or even more radically, postulate a Cartesian evil God who deceives us at every point. If a given astronomical theory conflicts with huge numbers of observations, we would generally be wise to reject the theory, not the
observations.

This tends to push coherentism in the direction of foundationalism to some extent. Relatively fixed points are the coherentist’s version of the foundationalist’s foundations. Thus, like foundationalism, coherentism can distinguish between two kinds of beliefs and give a kind of epistemological privilege to one of them, though not as strong a privilege as foundationalists assign to foundational beliefs.

In sum, there are two general models of the structure of justification. First, there is foundationalism according to which there are two different kinds of beliefs, foundational ones and superstructural ones. Foundational beliefs are not in need of external justification, while superstructural ones are. Superstructural beliefs are justified, ultimately, in terms of foundational beliefs. Justification therefore goes in one direction, from foundation to superstructure. Second, there is coherentism. It rejects the sharp dichotomy between foundational and superstructural beliefs. All beliefs are in more or less the same boat. All are in need of external justification. This justification comes from fitting them into the best coherent whole. This, in turn means that just as theory can be tested by observation, so can observation be tested by theory. Good clean observations can support or cast doubt on theory, and good clean theory can support or cast doubt on the accuracy of observation.

*Foundationalism and Coherentism as Methods.* Strictly speaking, foundationalism and coherentism are not “methods” of reasoning, moral or otherwise, in the way that the double blind experiment is a method for investigating the efficacy of drugs. Still, they are sometimes called “methods.” And it is true that although coherentists and foundationalists will often proceed in much the same ways when investigating issues, there are also some important
differences which many think are particularly important in moral philosophy. This warrants calling them different “methods” in a broad sense. It is worth bringing out some of the central differences between foundationalism and coherentism as methods.

The first is this. The foundationalist will say that when a genuinely foundational belief conflicts with a superstructural belief, the superstructural belief must go. So long as it is clear that the supposed foundational belief really is a foundational belief, it must always win when in conflict with a superstructural belief. The rejection of the superstructural belief is a “forced move.” But for a coherentist, when the any two beliefs conflict, a choice must be made. If one of the beliefs is a relatively fixed point, there may be an initial preference in favor of retaining that belief and rejecting the other, but it can be overridden in the interests of obtaining a more coherent whole. There may be an “encouraged” move, but no forced move.

A second major methodological difference between coherentists and foundationalists is this. For the foundationalist, any real knowledge must be rooted in the foundation. So, the development of genuine knowledge, whether or not moral knowledge, has a natural first step, to identify the foundation. Foundationalists will feel pressed to begin with what Descartes called “First Philosophy.” They will feel skeptical of any current beliefs until they have a First Philosophy. They will compare beliefs without foundations to building a skyscraper on the sand. Many philosophers have viewed it as being their particular job to work up these foundations, leaving superstructures to others. Coherentists, on the other hand, do not believe there is a foundation for all knowledge, so they will not believe there is a First Philosophy, something we must begin with. Rather, they will say that any epistemological project must begin pretty much from where we currently are, from our current set of beliefs. Our job is to begin from this
current set of beliefs, to seek out incoherences, and to try to deal with them, while at the same
time extending our beliefs into new areas.

This difference between foundationalism and coherentism is sometimes described in
terms of an analogy we owe to Neurath, a member of the “Vienna Circle,” an influential group of
logical positivists from the first half of this century. The idea is that we are to compare our
current set of beliefs to a boat that we are in. Our beliefs have various problems, for example,
they might be inconsistent in one way or another, or they might be incomplete. In the same way,
our boat can have various problems. For example, it might be too small for the number of people
on it, or it might have leaks. There are two approaches we might take to repairing the boat. We
might take it to a dry dock so that we can have good solid ground under our feet as we repair it.
Or we might try to make repairs while at sea, perhaps yanking out old planks and rapidly
slapping in new one’s before too much water rushes in, or sending out swimmers to work on the
outside of the hull who must deal with cold, mountainous seas, and sharks. Obviously, the
preferred course of action is to seek a dry dock. There is nothing like solid ground when one
starts dismantling one’s boat. But what if we cannot get to dry dock in time? Or what if there
simply is no dry dock to get to? Repairs must then be done on the fly.

The foundationalist project can be viewed as the search for a dry dock from which to
work on our beliefs. Just as we need a good solid foundation from which to repair our boat, so,
the foundationalist says, we need a good sold foundation from which to work on our beliefs.
Coherentists, on the other hand, deny that there is a dry dock. There is no use even looking. We
must work with what we have from where we are. If that seems unsatisfactory to those of
foundationalist inclination, it is too bad. For it is all we have. To update Neurath’s example, the
foundationalist’s search for secure foundations from which to proceed is like the Apollo 13’s crew deciding to wait till they got back to earth before making repairs to their spaceship. If they had waited, they never would have made it back and they never would have repaired their ship. They wisely chose to work from where they were, in space, with nothing but the resources available to them in space, and with no solid ground to stand on. Indeed, this analogy does not go far enough for the coherentists. In the Apollo 13 case, there was, in fact, an earth to return to. Apollo 13 might well have made it back to earth without repairs, even though the crew would have been dead. For the coherentist, matters are even more dire. For the better analogy would be where the crew decided to wait till they got back to earth when there were no earth at all to get back to.

There is a third difference between foundationalist and coherentist methods related to the last. In a sense, justification for the foundationalist can be *piecemeal* in a way that it cannot be for the coherentist. For the foundationalist, a belief is justified if and only if properly related to the foundation. Now return to our analogy with a building. In a large building, I can lay my foundation for the east wing, and build that wing, before beginning to lay the foundation for the west wing. So long as the east wing’s foundation is firm, what I build on it is firm, even though the west wing’s foundation has not been started. In the same way, with knowledge, a belief can be justified if properly linked to foundational beliefs even if the foundation is incomplete. For example, a handful of observations may be sufficient to completely refute an astronomical model. Justification for coherentists, on the other hand, is *holistic*. It turns on how the whole works. A handful of observations might count heavily against an astronomical model, but they cannot be conclusive till we have a very large chunk of knowledge, and in the ideal case, an
entire system of knowledge, filled in. For, depending on how the whole works out, we might
ultimately want to reject observations that once seemed accurate and which were once taken to
refute an astronomical model just as we might, in a criminal trial, come to reject a supposed
eyewitness account that we once accepted. How likely this is turns on many factors, especially
on how large a chunk of coherent belief we can embed those observations in. As our
astronomical observations fit into larger and larger structures, especially more and more detailed
physical theories, our confidence in them justly increases. But it can never be perfect till the
whole is filled in. In this sense, justification for coherentists cannot be piecemeal in a way that it
can be for coherentists.

Coherentist Methods and Relativism. It is likely that coherentist methods are
relativistic. “Relativism” is one of those tricky words that can have a number of meanings and
which should be avoided if possible. Unfortunately, it is often unavoidable. In ethics it typically
refers to the following view: what is right and wrong for a person (as judger and/or as agent) is
relative to the standards of his group (cultural relativity), or to his own individual standards
(individual relativism), or to some other standards which can differ from person to person. I call
this “ethical relativism.” It is a relativism about what standards correctly apply to people and
therefore about which acts are right and wrong for them. I am here interested in another kind of
relativism, a kind of epistemological relativism. I will speak of a method for justifying beliefs,
whether or not moral beliefs, as relativistic if two people can correctly employ that method and
come to different conclusions. When this happens, one belief will be justified for one person and
another belief will be justified for the other person.

Coherentist methods are probably relativistic in this sense. This is true for several
reasons. First, remember, for coherentism, there is no foundation from which we begin. Each thinker must begin from where he or she is. If we begin in different places, the effort to find coherence might lead us to different places. Return to our boat analogy. If you begin with a schooner and I begin with a frigate, and we both repair and update our ships perfectly, you will still have a schooner and I will still have a frigate, though better ones than we had before.

Second, even if we begin in exactly the same places, we might end up in different places. This is because coherentism, unlike foundationalism, does not present us with forced choices, even though there might be “encouraged” ones. This means that two people facing the same choices might make different decisions, and once this starts, differences can expand.

Admittedly, one could argue that, our boat analogy not withstanding, when all is said and done, a genuinely complete, coherent system of beliefs will have to be identical to any other genuinely complete, coherent system of beliefs. One could argue that sooner or later, there will always be something that pushes divergent views back together. This is a claim sometimes made on behalf of science: wherever we begin, and whatever choices we make along the way, sooner or later the facts will push us into agreement, if we live long enough and continue to sincerely investigate. Maybe this is so. But I know no proof that convergence will ultimately result. It seems to be more a statement of faith than anything else. And in any event, even if it is true in science, it is much less likely in the moral sphere. It seems likely that coherentism is a relativistic method.

Coherentist Methods as Conservative. Coherentist methods can be rather conservative. We begin from where we are, seek out incoherences, and modify our beliefs so as to render them coherent. There is an assumption that the smallest modifications possible are to be preferred.
After all, what could justify more massive changes when smaller ones can produce coherence? That would seem to require some external stance, some “God’s eye point of view,” from which we can evaluate the entire system of beliefs and reject them in a deep way. Again our boat analogy is useful. It is highly unlikely that we will be able to make radical changes in a boat that is at sea. A secure dry dock would seem necessary for that, and by hypothesis, there is no dry dock for knowledge.

Some have taken the conservativism of coherentism to be an objection to the view. However, we should not push this point too far. Coherentist methods do have the possibility of generating more radical changes than might appear likely at first glance. For one thing, it is possible that a person’s beliefs have significant conflicts and incoherences. It may not be possible to render them coherent without fairly massive change. Further, we must not forget the role that the formation of new beliefs might have. For coherentism, our job is not just to render our current beliefs coherent, but also to extend our beliefs into new areas. Thus scientists will continue to observe and to experiment, and moral thinkers will continue to acquire experience and knowledge of the world. Old ways of thinking might come to seem seriously, and not just modestly, inadequate as our collection of beliefs grows. Efforts to accommodate new beliefs by slight modification here and there might grow more and more uncomfortable. What Thomas Kuhn, an historian of science, called “anomalies” might build up which resist coherence with our present beliefs. This could, in principle, motivate what Kuhn called a “revolution.” Kuhn, of course, was talking about scientific revolutions, but the same is possible for moral revolutions.4

**Examples of Coherentist Reasoning.** It is worth illustrating coherentism with a distinctive form of argument that is popular in modern moral discussions by philosophers. This is the
consistency argument. Suppose John and Sue have different beliefs on a particular matter. John believes P and Sue believes Q. In the effort to persuade her to change her mind and to adopt P, John gets Sue to make a variety of moral judgments about various particular cases and about various general principles. He then points out to Sue that her belief that Q is inconsistent with some other belief she expressed, say R. Sue must now resolve the inconsistency in some way or another. Now John’s goal is to get Sue to give up Q and to embrace his own view, P. But in fact there is no forced move here, no one way for Sue to resolve the contradiction. It can be resolved by giving up Q, but it can also be resolved by giving up R. It might even be possible for her to argue that Q and R are not inconsistent after all, that the inconsistency was merely apparent. Which route Sue takes is a matter of choice, though a choice that must be made in light of the larger picture. John’s job is to continue to elicit judgments from Sue that show any moves she makes will produce new incoherences in her beliefs except for one move: giving up Q and adopting P.

Here is an example of a consistency argument: Robert Nozick’s argument in favor of vegetarianism. Though not a major focus of his book -- the topic only comes up in passing -- Nozick asks whether or not it is morally permissible to eat animals for food. He assumes that it is wrong to cause suffering in animals for its own sake, out of cruelty. But he considers the possibility that it is okay when our goal is not to cause suffering and death per se, but rather to eat the animal. He is not interested in the case in which eating animals is necessary for life or even health. He assumes that health can be maintained by most people in the U.S. today without eating meat. He is instead interested in the case in which we eat meat simply for the pleasures of the palate. He begins with an analogy. He asks us to imagine the following case.
Suppose then that I enjoy swinging a baseball bat. It happens that in front of the only place to swing it stands a cow. Swinging the bat unfortunately would involve smashing the cow's head. But I wouldn't get fun from doing that; the pleasure comes from exercising my muscles, swinging well, and so on. It's unfortunate that as a side effect (not a means) of my doing this, the animal's skull gets smashed. To be sure, I could forego swinging the bat, and instead bend down and touch my toes or do some other exercise. But this wouldn't be as enjoyable as swinging the bat; I won't get as much fun, pleasure, or delight out of it. So the question is: would it be all right for me to swing the bat in order to get the extra pleasure of swinging it as compared to the best available alternative activity that does not involving harming the animal? (p. 37)

The judgment he wishes to elicit is, of course, that it is wrong to swing the bat under these conditions. It is wrong to inflict that much pain, even on a cow, for such a trivial reason. And, he wants to say, it is inconsistent to judge that it is wrong to swing the bat and right to kill and eat animals. For both are done for nothing but rather trivial pleasures. How might we resolve this inconsistency? Nozick wishes us to resolve it by giving up the judgment that it is right to kill and eat animals for the pleasures of the palate. However, since this is a consistency argument, that is not a forced move. We have three possible moves. First, we can conclude we ought not to eat meat. Second, we can conclude that it is okay to bash the cow. Or third, we can try to argue that there is no inconsistency here after all, that it is merely apparent. To show this, we would have to show that the case of bashing the cow and the case of eating meat are, in fact, very different form one another in some morally significant way, making it possible to accept both judgments. For Nozick to complete his argument (which he does not do), he must show
that, once all is said and done, the most reasonable move in light of a coherent, complete moral theory is the first, to give up the belief that eating meat is legitimate.

*John Rawls and the Method of Reflective Equilibrium.* It is time to leave general views about the nature of justification and turn to moral philosophy. John Rawls, who is arguably the most influential American moral philosopher of the century, brought coherentist methods to center stage in moral theory in his monumental book, *A Theory of Justice*. He refers to his coherentist method as the method of “reflective equilibrium.” Given how rarely philosophers agree on anything, there is surprisingly wide agreement on the use of various versions of reflective equilibrium methods. Indeed, in one form or another, they can be thought of as the default approach to moral reasoning among contemporary Anglo-American philosophers, at least for those who have not committed to some other method because of their adoption of some particular view of the nature of the moral game. Most of the elements of reflective equilibrium theory are found in *Theory of Justice*, but some are not fully developed.

I begin with Rawls’ understanding, from *A Theory of Justice*, of what we are trying to do when we do moral philosophy. He tells us that it is an attempt to describe our moral capacities, though this is not really quite accurate: as we will see, going through the reflective equilibrium process not only describes, but modifies our moral capacities. His own concern is with our sense of justice, rather than with our moral capacities as a whole, but I will keep the broader focus. Though not necessarily inconsistent with objectivist views of morality, the method of reflective equilibrium, for Rawls, does not assume that there are objective moral truths or that there is one correct moral conception. It is, he says several times, doing a kind of psychology, simply investigating the shape of our moral sensibilities without even assuming that our moral
What is it to correctly describe our moral capacities? At a first approximation, it is to come up with principles that are in accord with the “everyday moral judgments” we make. But Rawls quickly replaces the notion of an everyday judgment with the notion of a considered judgment. These are judgments “in which our moral capacities are most likely to be displayed without distortion” (p. 47) and are “rendered under conditions favorable to the exercise of the sense of justice and therefore in circumstances where the more common excuses and explanations for making a mistake do not obtain” (p. 47-8). These conditions are, he says, simply the conditions necessary for good judgment on any topic. The judgments must be made without hesitation, when we are not upset or frightened, and when we do not stand to gain personally. The person making the judgments must have the ability, the opportunity, and the desire to reach a correct decision. Such a person is referred to by philosophers as a “competent judge.” An example of what Rawls takes to be a considered judgment, at least for most contemporary Americans, is that religious intolerance and racial discrimination are unjust. Indeed, he not only views these as considered judgments, but as being among our most firm considered judgments. They are, he thinks, “provisionally fixed points.” From these examples, it should be clear that both considered judgments and fixed points can vary from person to person and culture to culture. That is the nature of coherentism.

So moral philosophy is an attempt to describe our moral senses. This involves coming up with principles which match our considered judgments, and especially our fixed points. But this is just a first step in understanding the method. For even our most considered judgments may be “subject to certain irregularities and distortions despite the fact that they are rendered under
favorable circumstances” (p. 48). So they should be given a fair amount of priority in our choice of principles, but not an absolute priority. Suppose we come up with a set of principles that match and account for a great many of our considered judgments, but not all of them. If the principles constitute an intuitively appealing account of our sense of justice, we may prefer to revise our previous considered particular judgments rather than the principles, especially, but not only if, we can find an explanation for the deviations which undermine our confidence in our original judgments. For example, we may come to view some of our judgments as being the results of bias we might have because of our social position or our upbringing. Knowing this might serve to undermine them, and strengthen the case for giving them up when they conflict with intuitively appealing general principles. So when we find conflicts between considered particular judgments and intuitively plausible principles, we have a choice. We either keep our original considered judgments and reject (or modify) the intuitively plausible principles, or we keep the principles and reject the original considered judgment. We construct a moral philosophy by working back and forth, sometimes modifying principles to bring them into accord with our considered judgments, and sometimes modifying our considered judgments to bring them into accord with our principles. The end state of this process is a set of general moral principles that are in reflective equilibrium with the surviving considered judgments. Rawls speaks of these principles as “regulative” of our moral sense. They describe the shape of our moral psychology, or of a modified, improved version of it. Of course, any state of reflective equilibrium we actually reach in the real world is likely to be unstable as further reflection or new information forces us to go back and rework aspects of it.

Rawls tries to clarify his approach by an analogy – actually, he uses two but I put one
aside here. He compares the project of describing our sense of justice to the problem of
describing our sense of grammaticalness for sentences in our native language (p. 47). In
describing our sense of grammaticalness, we try to formulate principles that make the same
grammatical discriminations that native speakers make. In the same way, in describing our sense
of justice, we seek principles that make the same carefully considered moral discriminations that
we make. This analogy is helpful, but should not be pushed too far. For when we construct rules
of grammar, we do not work back and forth between considered grammatical judgments about
particular sentences and general principles the way Rawls thinks we do when we construct a
moral philosophy. There is little, in the grammar case, that seems analogous to tossing a
particular considered judgment to preserve an attractive grammatical rule. That is the whole
point of the dreaded category of the “irregular” in grammar.

A final point needs to be emphasized since it is central to the reflective equilibrium
method, that of mutual support. As a form of coherentism, reflective equilibrium methods do not
give ultimate priority to either our considered moral judgments about particular cases or
intuitively plausible principles that purport to describe our sense of justice. Each of these
components has some initial plausibility, but also obtains greater plausibility by finding a place
within a coherent whole. In other words, that a set of principles accurately describes our
considered particular judgments helps support those principles, but it also helps support the
considered particular judgments. The two components mutual support one another.

My explanation of Rawls' view so far has been rather abstract. So as not to lose touch
with what it is all about, we should pause to look at a simple example of a disequilibrium.
Suppose that George, after taking a variety of anthropology classes, becomes enamored of
traditional, technologically primitive ways of life, ways that are fast disappearing from our world. He comes to believe that those ways of life are valuable and should be protected as far as possible from the encroachment of modern civilization. George has also been an environmentalist for many years, and is especially interested in the protection of marine mammals. His beliefs in the value of environmental protection and in the value of protecting traditional societies have long sat comfortably together. He has viewed members of traditional societies as natural conservationists and as living in harmony with their environment. But a problem occurs. He reads in the newspaper of a group of Native Americans in Alaska wishing to reestablish traditional whale hunting. As an environmentalist, and also as a person with a great love of the giant marine mammals, George has long been opposed to hunting whales. And yet, the hunt had for centuries, if not millennia, been part of traditional culture. George now faces a conflict. He comes to see that his support for traditional cultures may not be fully consistent with his environmentalism. This conflict intensifies when he goes to the library to do a little research on the relations not just between contemporary traditional cultures and their environments, but between prehistoric cultures and their environments. He finds that such cultures were not quite as environmentally benign as he had long assumed. He reads with dismay chapters of *Quaternary Extinctions* in which experts argue that prehistoric peoples were destructive much as modern humans are. Wherever humans appeared, species disappeared. This pattern is found through much of the world, from North America to Polynesia. It is speculated, he reads, that such destruction was the result of causes much like the modern causes of the destruction of species. These include over hunting, habitat destruction, and the introduction of exotic species such as rats.
Faced with this dilemma, George is in a state of disequilibrium. The solution is not obvious. He must give up, or at least modify, his support of environmental causes, his support of traditional societies, or both. Many moves are open to him. For example, he could weaken his support of traditional cultures and say they are only to be supported in so far as they are not environmentally destructive beyond a point. Or he could, in the name of preserving traditional cultures, modify his support of environmental causes. He could say that whaling, which would never be permissible for a modern, technological society, is legitimate as part of a traditional culture. And so far as the issue of Native Alaskan whaling goes, he even has the possibility of trying to preserve both his commitment to the environment and to traditional cultures without modification. For he could argue that strictly speaking, once disrupted, genuinely traditional cultures cannot be reestablished. The traditional lives of Alaskan natives are largely gone, replaced by something different. And even if reestablishment were in principle possible, reintroducing the hunt does not do the job. Rather than reestablishing the traditional culture, it is nothing but a symbolic reminder of that now dead culture. And, he could reason, the great sea mammals need not be sacrificed as a mere symbol of a dead culture.

I will not try to solve George’s problem for him. There may be several solutions that are equally good from the point of reflective equilibrium. On the other hand, when everything is taken into account, there might be only one solution that works very well. That is not something that can be determined in advance. But at least at first glance, the possibility of multiple solutions, multiple equilibria, is strong.

*Narrow and Wide Reflective Equilibrium*. What I have been describing so far is called the method of “narrow” reflective equilibrium. It involves seeking a set of intuitively plausible
general principles that can serve as the regulative principles for a modified set of considered judgments about particular cases. Some philosophers are content with seeking a narrow reflective equilibrium, that is, with devising a set of general moral principles that cohere with and unify our more particular moral judgments and do it in a way we find intuitively plausible. But *A Theory of Justice* has an expanded method which Rawls further developed in later essays and which has grabbed the attention of many writers. This is the method of “wide” reflective equilibrium. Rawls does not work out in much detail how this works, so I will appeal, in my explanation, to ideas from Norman Daniels, one of the most influential early commentators on wide reflective equilibrium methods.¹⁰

The method of wide reflective equilibrium adds extra complexities to the narrow method. But the basic idea is simple and can be gotten from a straightforward example which will then need generalization. Suppose you go through the entire narrow reflective equilibrium process and that you end up with some intuitively plausible general principles that seem to serve well as regulative of your surviving considered particular judgments. But suppose also that you are a religious person and believe in a divine command approach to morality: morality is the command of God as found in certain authoritative documents, say, the Bible. You now notice, however, that your moral beliefs in narrow reflective equilibrium are out of sync with the moral beliefs that your religious beliefs commit you to. You are therefore in a narrow reflective equilibrium, but in a wider disequilibrium. Something has to give way. You have at least the following options, which are not exhaustive or mutually exclusive. You can modify your moral beliefs to be consistent with the requirements of your religious beliefs. You can modify your divine command theory, say, going for a natural law theory instead. You can reinterpret the relevant parts of the
This illustrates that a reflective equilibrium which only focuses on rendering particular and moral judgments and general principles coherent is, indeed, quite narrow. There are many other things that might need to be brought in, and in particular, anything that might have an impact on one's moral beliefs. One must bring in to the mix anything that might be part of an argument for or against a set of moral principles. Often these are beliefs about the nature of morality. In the previous example, it was the belief that morality is the command of God. But there are other possibilities. For example, we have already seen that some philosophers, such as R. M. Hare, think that moral language is to be understood as universally prescriptive, and that this can be the basis for a philosophical argument for utilitarianism. If one is inclined toward this theory of moral language, it must be brought into the mix. Others think that since morality is made for humans, a particular conception of human nature can be part of an argument for or against a set of moral principles. If this is true, then conceptions of human nature must be brought into the mix. Narrow reflective equilibrium can be sought on its own, but it is a curiously atheoretical approach, perhaps appropriate for one who does not have any wider commitments of, for example, a metaethical sort. But if one does have such commitments, or wishes to form them, they must be taken into account too. Further, there is no way to tell, in advance, what else might be relevant. Depending on what one’s underlying conception of morality is, any number of other matters could turn out to be relevant.

So, a wide reflective equilibrium involves an equilibrium involving three elements. First, there are our considered judgments about particular cases, with special attention being given to our provisionally fixed points. Second, there are intuitively plausible general moral principles.
And third, there are various background theories that might serve as the basis for abstract philosophical arguments for various kinds of moral principles. We need to find someway to modify each of these elements so as to find an intuitively acceptable whole.

Another way to put together the same ideas is this. For the wide reflective equilibrium theorist, moral philosophy involves at least two projects. First, there is the project of trying to render our particular judgments coherent by finding some intuitively plausible general regulative principles for them, even if this involves giving up some of our particular judgments. Second, it involves finding more abstract arguments, such as religious arguments or arguments from the nature of moral language, for various sets of principles. Ideally, these two projects should complement one another and leave us in the same place. If they do not, something has to give. Once we decide what gives, we have a wide reflective equilibrium. Now all this may seem bizarrely complex. But though it is complex, it is not bizarrely so. In fact, when one is dealing with most any theoretically interesting area, one finds at least this much complexity. Often, in science, researchers working in one area, A, come to various conclusions about some matter and that researchers in a second area, B, also come to some conclusions about that matter. And it could turn out that the research in A has implications for the research in B and that the conclusions in A force researchers in B to rethink some of their conclusions, and vice versa. Here is an example. Many disciplines are called upon to determine the nature of the colonization of the Americas by humans, and results from one discipline might be inconsistent with results from other disciplines, in which case we must rethink things and modify something somewhere. For example, geologists might tell us that there was only a single, very specific time slot when people could have crossed from Asia to the Americas. But it might be that genetics tells us that
there were likely two distinct migrations at different times. One of these has to be wrong. So either we must rework the genetics data, or figure out what went wrong with the geological data. This is analogous to what we did with the two philosophical projects.

As with narrow reflective equilibrium methods, wide methods make use of the notion of mutual support. The various components, particular considered judgments, intuitive principles, and background theories that can generate philosophical arguments for and against a set of principles all have some sort of independent plausibility, but they all also support one another. That a given plausible background theory generates a philosophical argument that supports principles that are in equilibrium with our considered particular judgments raises the plausibility of all the components.

Rawls does not himself embrace a divine command theory of morality, but he does have an underlying conception of morality, or at least of the corner of morality he is concerned with. It is important for those who wish to understand Rawls to notice a progressive narrowing of his focus over time. By the end of his career, he gave up the search for what he came to call a “comprehensive” moral or philosophical doctrine in favor of a narrowly “political” doctrine. But putting that aside, in *Theory of Justice*, he is attracted to a neo-Kantian picture of the nature of morality. Rawls is most attracted to Kant’s idea that the correct moral principles are those chosen by the members of a “Kingdom of Ends,” a group of rational beings who are both legislators and subjects. Precisely what this means for Kant is not important here, but what is important is that Rawls tries to make more precise this very obscure idea, and indeed, to “operationalize” it.

For those deprived souls not acquainted with Rawls, it is worth summarizing what he
does here and then see how it ties in with his reflective equilibrium methodology. In *Theory of Justice*, Rawls is out to discover the most basic principles of justice for a modern society such as our own. He settles on two principles, which I will just state and not try to explain. First, there is the greatest equal liberties principle which says we should seek the greatest liberty compatible with a like liberty for all. Second, there is a two-part principle covering economic justice. The first part is called the “difference principle” and says that we should arrange things so that any inequalities in wealth, income, and other “primary goods” benefit everyone, and especially the least well-off. Second, there is the “fair equality of opportunity principle” which says that the better positions, those that get the greater share of the primary goods, should be open to all.

What is interesting here is his argument for these two principles. He has us consider an imaginary “Original Position” and says that the correct principles of justice are ones that would be chosen by people in this Original Position in a sort of imaginary contract. This is a purely hypothetical situation that has never existed and cannot really exist but which we can enter in imagination, as a thought experiment, whenever we like. Imagine the Original Position peopled by beings much like us but with the following features – this is not a comprehensive list but will give you the basic idea. First, they have the best general knowledge we can give them about everything relevant to the decisions they will be called upon to make, for example, about economics and politics. Second, they are behind Rawls' famous “veil of ignorance,” that is, they lack all particular knowledge about themselves. They do not know who they are in the real world, what their station is, and so on. This means that when they reason with one another about what principles to adopt, they cannot tailor their arguments to benefit themselves and must take a more impartial point of view. Third, imagine that they are rational in the sense that they want
more rather than less of the primary goods, the basic means for a good life such as wealth and income, and they are reasonably good at getting them, even though they do not know just what they want them for since they are behind the veil of ignorance. Once Rawls fully defines the Original Position, he gives its residents the job of selecting the principles governing the basic structure of society. Note that they are not choosing principles for society behind the veil of ignorance in the Original Position, but for real society. These beings are choosing for themselves on the assumption that once they choose, they will leave the Original Position and enter a real world governed by those principles. Hopefully, this is enough to give the flavor of the theory.

Where does reflective equilibrium theory come in? Rawls says that he has certain “background theories” about the nature of persons, society, morality, and so on. For example, as mentioned, he assumes that morality should be something chosen by something like the members of a Kingdom of Ends. He then uses those background assumptions to construct a kind of apparatus or machinery, the Original Position argument, which generates moral principles. He must now bring all this – his background theories, the apparatus, and the resulting general principles – into contact with the considered judgments and principles he ended with when he achieved narrow reflective equilibrium. If the principles reached in his narrow reflective equilibrium match those which result from his Original Position argument, he is in wide reflective equilibrium. If not, something must give. Either he must modify some of the conclusion he came to when he reached narrow reflective equilibrium, or he most modify something in his more abstract argument involving background theories and the Original Position.
Reflective Equilibrium Theory, Convergence and Objectivity. Rawls refuses to take up the question of whether or not the method of wide reflective equilibrium, if successfully carried out, can lead to something we might call “objectivity” in ethics, though his analogy between doing moral philosophy and finding the principles of grammar certainly seem to point away from any kind of objectivity. If this is our line, we are likely to say that success is better thought of as producing convergence in our moral opinions, what is sometimes called “inter-subjectivity,” which is a very weak sort of objectivity, if it is objectivity in any meaningful sense. This has not stopped others from considering the possibility that we could get at some sort of objective moral truth using reflective equilibrium methods. Interestingly, the philosopher of science, Richard Boyd, thinks that reflective equilibrium methods are the proper methods of science, and not just of ethics, and that they can, although they need not, lead us to objective scientific truth. One might try to argue similarly for ethics.

There are two issues here. First, are Rawlsian wide reflective equilibrium methods likely to lead to significant convergence, and second, is this convergence a sign that we have, or might be, getting at some sort of objective moral truth? Norman Daniels thinks that the method of wide reflective equilibrium may aid in convergence by reducing intractable moral disagreements to possibly more tractable disagreements about background theories. For example, we might find the problem of abortion is in itself an intractable problem, but if we can come to some agreement about, say, the theory of the person and how that theory relates to morality, we might be able to use that agreement on background theories to motivate greater agreement about abortion. At least that is his idea. Personally, I am skeptical about this. It is not obvious to me that disagreements about various background theories, such as the nature of the person, the existence
of God, or the nature of moral language, are any more tractable than disagreements about concrete moral issues such as abortion. And the history of philosophy since Daniels made that suggestion certainly does not give us any reason to be sanguine about that possibility.

But what if the method of wide reflective equilibrium really did help us to converge in our moral judgments to a significant extent. Daniels does not think that such convergence is either necessary or sufficient for us to conclude that we have objective moral truth. But he thinks that such convergence is at least some evidence that we are reaching moral truth. Consider arguments about some realm that have the following general form:

1. Suppose that in an area of inquiry, the methods used are successful in the sense that they produce convergence and growth of knowledge.
2. Suppose further that the only plausible account of the success of these methods is that they lead us to better approximations to truths of the kind relevant to the inquiry.
3. Then we should adopt a realist account of the relevant objects of inquiry.

This is part of a classic defense of scientific realism. In science, we introduce various kinds of theoretical entities, that is, entities that we do not in any real sense directly detect, but which we posit because they help us explain and predict various things. Examples include atoms as well as the ever smaller subatomic particles such as electrons, neutrons and protons, not to mention gluons and quarks. We posit these things, and claim that they have certain characteristics (such as electrons have negative charge) because doing so lets us explain and predict. But do we really have reason to believe these things exist? Some philosophers, including those called "instrumentalists," so no. These are simply "useful fictions." No one really knows, or can know, what the world of the very small is like, but if these fictions help us control things, we might as
well go with them, so long as we do not lose track of their “useful fiction” status. But others think we have good reason to think these things are real, and at least part of the reason we should believe in them is that they are part of our most coherent picture of the universe, a picture we come to using reflective equilibrium methods. Now obviously, the argument for the reality of these theoretical entities will have to be more complex than this. But the idea Daniels suggests is that something similar might be said for moral realism. And Boyd, the philosopher of science mentioned above, thinks that just as reflective equilibrium methods can lead us to objective truth in science, so can they in ethics.

Admittedly, this is an interesting argument. But I am skeptical, and Daniels himself adds some “qualifying remarks” to his discussion of the argument. These qualifying remarks are not trivial and go some way to undermining the analogy between ethics and science even if they share the reflective equilibrium method. First, arguments for scientific realism focus not only on convergence, but also on other factors pointing to the cumulative nature of science, for example, ‘take off’ effects (the phenomenon of a particular science metaphorically taking off, the way, at some point genetics did) and the absence of “sects” in science. Second, we cannot assume that if the search for reflective equilibrium produces consensus in two areas of inquiry, then it does so for the same reason, that is, because it leads to truth. Third, the argument for scientific realism requires a causal account of knowledge. We lack a comparable account for moral knowledge. And fourth, there is nothing in ethics analogous to experimentation in science. We need an account of what we can learn from moral practice to fill this out. It seems to me that Daniels does a pretty good job of undermining his own hope the reflective equilibrium methods will produce objective results in ethics even if they can in science.
Problems with the Method of Reflective Equilibrium: Whether or not the reflective equilibrium method can plausibly be said to lead to objective knowledge in ethics comparable to what we have in science, it might nevertheless be the appropriate method for ethics. And as I said, most contemporary moral philosophers use it or some variation on it. So it is worth noting some of the basic problems with the view. I do not want to say that these problems cannot be solved, but I will not try to solve them here.

One major problem critics have raised is that of circularity. What do I mean by circularity, and why is circularity a problem? Circularity can be illustrated by a simple example. Suppose I am wondering whether or not to trust Sue’s word on some matter. She tells me “Jack will vouch for me.” I respond “Fine, but I do not know Jack. Who will vouch for him?” “No problem! I will vouch for Jack. I give you my word, he can be trusted.” Obviously, in order for Jack to be trusted, he needs support from something other than Sue, since Sue herself depends on Jack to support her word. For Sue to vouch for Jack and Jack to vouch for Sue is circular and offers no real support. In the same way, suppose I believe that P, and when asked why, I offer Q as evidence. When asked why I believe Q, I say because of R. And when asked to defend R, I respond “Because P.” This is circular.

In general, coherentist accounts of justification face the possible charge of circularity since various elements in our web of knowledge offer each mutual support, as explained above. Many people think that this notion of mutual support is quite suspect, and is, indeed, just a nice word for something philosophically nasty, circularity. In Rawls’ case the circularity might arise if we only focus on narrow reflective equilibrium, and also if we turn to wide reflective equilibrium. First, narrow reflective equilibrium. A given considered judgment has an initial
plausibility of its own. It is then used to test various general principles. However, any considered judgment only has a provisionally plausible plausibility. It can be rejected. It is only retained if it can be made to cohere with intuitively plausible principles. So our considered judgments about particular cases are used to support various principles, and intuitively plausible principles are used to support judgments about particular cases. If the same principles that are supported by particular considered judgments also support those particular considered judgments, there is a circle. Next, wide reflective equilibrium. Suppose we have considered judgments about particular cases and general principles in a coherent whole: narrow reflective equilibrium. Call the general principles in this mix $P$. We now throw into the mix various background theories which support philosophical arguments that lead to various principles. Suppose that our chosen background theories yield philosophical arguments that also lead us to principles $P$. Then we have what seems to be a circularity: our background theories support our work that we did in narrow reflective equilibrium and the work we did in narrow reflective equilibrium further supports our background theories.

A second problem people have raised for reflective equilibrium methods is the indeterminacy problem. Suppose that an incoherence is found in Environmentalist George’s beliefs in the example used above. He finds that his support of environmental causes is not fully consistent with his support of traditional societies. A choice must be made. But the test of coherence does not dictate which of the many possible ways of resolving the conflict is best. The only real restriction is that any choice must not only work for the case in hand, but also for other cases. For the goal is not coherence only between a few beliefs, but for one’s beliefs generally and it is possible that there are many choices compatible with this restriction. In other
words, two people can start with exactly the same beliefs with exactly the same incoherences, and end in very different places. The resulting equilibria will therefore be arbitrary in an important sense. Now Daniels will argue that once George brings in various background theories on which to base philosophical arguments for and against various principles, this indeterminacy will be reduced. But as I said, I see no reason to think this is true.

A third objection is that reflective equilibrium methods are really nothing but intuitionistic methods in disguise. Indeed, many reflective equilibrium theorists use the language of intuitionism, speaking of considered particular judgments as intuitions and of intuitive general principles. Since I argued that intuitionism is suspect, so might reflective equilibrium be. This one I think we can respond to. Daniels points out that whatever the language, this is not intuitionism, at least if we take intuitionism in its standard form. For intuitionism is generally foundationalist and picks out some set of moral beliefs that are, in some sense, deeply privileged, for example, self-evident. This is not part of reflective equilibrium methodology. No special significant priority is even granted to considered judgments, though admittedly, being a relatively fixed point is a modest kind of privilege. Further, reflective equilibrium theory does not require any deeply troubling metaphysical or epistemological views of the sort we find in Moore and Ross. It is consistent with such views, but is independent of them.

A fourth objection is more troublesome and grows out of the response just given to the previous objection. It is that reflective equilibrium methods are, in effect, intuitionism without the fairy tales of epistemic privilege and odd metaphysical entities. But without such a story, these intuitions become mere opinions which are often the result of self-deception, self-interest, historical and cultural accident, and so on. All reflective equilibrium methods do is systematize
and modestly rationalize our already existing biases, and they do it in a rather conservative way. This, critics say, does not sound like a process that can possibly lead to real justification of our beliefs. So, for example, it is likely that ancient societies that practiced slavery, widespread infanticide, and, in the case of Rome, gladiatorial games, would probably have little trouble spinning a coherent picture of their moral lives. But Daniels thinks this is not fair. For though there is something to the idea that narrow reflective equilibrium methods just systematize our biases, and do it in a rather conservative way, wide reflective equilibrium methods in fact allow more extensive revision of moral judgments. No type of considered moral judgment is immune to revision, though we will try to avoid giving up some. If, as a matter of fact, as part of our background theories, we come to view whole sets of our moral beliefs as having a corrupt origin and as being mere biases due, say, to our socio-economic position, then that gives us reason, in seeking wide reflective equilibrium, to reject them, or at least discount their authority.

Whether or not the reflective equilibrium theorist can solve these and other problems is unclear and a matter of some debate in the literature. But they are at least partly avoided if we buy Rawls’ concept of what moral philosophy is trying to do. It is not like physics. It is not an attempt to get at some sort of objective moral world. He does not deny such an objective moral world exists, but he is not committed to it either. Rather, as we saw, he characterizes moral philosophy as a kind of psychology, as an effort to characterize our moral sense. Even if the problems described above are devastating problems for the study of morality conceived of as an objective realm, it is not obvious that these are devastating problems for the study of morality as Rawls conceives it.
Endnotes

1. For the comparison to a building, see Reid (Active, p. 234). It is also found in many other places, for example, in Descartes who writes “my method imitates that of the architect. When an architect wants to build a house which is stable on ground where there is a sandy topsoil over underlying rock, or clay, or some other firm base, he beings by diffing out a set of trenches from which he removes the sane, and anything resting on or mixed in with the sane, so that he can lay his foundations on firm soil.” (Replies 7, AT 7:537).

2. Those acquainted with these discussions will notice that I am ignoring the massive debate about exactly what counts as foundational sensory beliefs, the search for the elusive “sense data.” The details of that debate are irrelevant here.


6. Most of the material in Theory of Justice on reflective equilibrium appears in Chapter 4 and Chapter 9.

7. Before continuing, I should point out that some authors question whether Rawls' reflective equilibrium method really is coherentist as opposed to a “modest” foundationalism because of its use of provisionally fixed points. But as I explained, a fully developed coherentism and a fully developed foundationalism tend to look more and more like each other. It signifies little whether we call Rawls' view a modified coherentism or a modified foundationalism.

8. Some of these ideas were better worked out by Rawls after he wrote A Theory of Justice. See in particular, “The Independence of Moral Theory, Proceedings and Addresses of the American Philosophical Association XLVII (1975), p. 5-22.

9. Paul Martin and Richard Klein, Quaternary Extinctions: A Prehistoric Revolution (Tucson:
Endnotes


11. For the contrast to physics, see his brief remark in _A Theory of Justice_, p. 49.