Class 7 - Reflective Equilibrium in Ethics

I. Ethics and Meta-Ethics

Rawls's main concern in *A Theory of Justice* is to establish a theoretical framework in which to conduct what we might call first-order ethical work.
To understand his goal, and more importantly his methods, it will be useful to distinguish first-order ethics (or ethics) from second-order ethics (or meta-ethics).
Ethics is the quest for universal, or universalizable, prescriptions for action, ones which are public and practicable and which override other normative claims (like those of manners, aesthetics, grammar, or social convention).
Meta-ethics is the study of the possibility of ethics, including the study of the meanings of ethical terms.
One of Rawls's central achievements was to almost single-handedly return the attention of vast numbers of philosophers from meta-ethics back to ethics.
The key element in Rawls's revolution is his adoption of Goodman's method of reflective equilibrium.
In order to explain how Rawls returned ethics to respectability by introducing the method of reflective equilibrium, I am going to provide a very rough caricature of about twenty-five hundred years of moral philosophy.

II. The Entire History of Moral Philosophy, from the Beginning of Time Until 1883

Until the early twentieth-century, most work in ethics was first-order.
Pre-modern Western philosophers like Aristotle, Cicero, and Augustine, as well as eastern philosophers, when writing about morality, tended to argue that people should adopt particular behaviors, practices, and virtues.
Aristotle, for example, defends moderateness in action, as a route toward one's own happiness, which he takes as an unassailable goal.
One might call Aristotle a foundationalist in ethics.
For Aristotle, all moral reasoning flows from, is derived from, first ethical principles: the quest for *eudaemonia*.
Aristotle's moral theory is not universal, in the sense that it tells us how to develop virtues for an individual, and justifies those virtues on the basis of that individual's interests.
Indeed, Aristotle believed that *eudaemonia* was not even possible for large swaths of the population, including women and people he considered natural slaves.
Still, he provided a set of public and practical prescriptions for developing one's character.
In the middle ages, Aristotle's virtues were re-interpreted by the Church as universal virtues; Aristotle's vices were re-imagined as cardinal sins; his moral theory was universalized and made catholic.

In the modern period and into the nineteenth century, two dominant ethical approaches were developed.
Kant's deontological (or duty-based or backward-looking) ethics used our duties as first principles.
Utilitarianism (or consequentialist, or forward-looking) ethics used hedonism as a first principle.
These theories are completely opposed in the content of their prescriptions.
They differ radically in their epistemic grounds.
But they are similar in that they are both, again, first-order ethical theories.

Kant's work in ethics starts with a very general claim that morality is possible.

This claim is taken as an obvious and unassailable truth.
Kant then works backwards: if moral action is possible, then morality must be autonomous, rather than heterogeneous; it must depend completely on oneself, and not be dependent on actions of others.
If the morality of my action depended on the actions of others, then it would not be possible to ensure that I behaved morally.
But, quests for happiness are inevitably dependent on the way the world around us responds to our actions.
We can not control the rest of the world; we can only control our own behaviors.
So, morality can not depend, in any way, on happiness.
Instead, moral theory must be a system of universal principles governing one's intentions, or duties, independent of the results of acting on those intentions.
Kant calls these universal principles the categorical imperative, a name which signifies both the certainty of their dictates and their utility as a first-order moral theory.
That is, the categorical imperative tells you how to behave and how not to behave.

Like Kant, Mill, the most prominent utilitarian philosopher, develops and defends the ground for his moral theory.
Like Kant, Mill relies on a basic, obvious assumption about morality.
Unlike Kant, Mill's particular assumption is that every one desires happiness.
Given this universal desire, Mill proceeds to argue that the only defensible moral theory ties the permissibility of action to the increase or decrease of pleasure and pain, both in oneself and in others.
An advantage of Mill's theory is that he identifies the good with something found in the world.
We can objectively measure the amount of pleasure and pain in our selves and, given the veracity of people's reports of their own mental states, others.

Unlike Kant, Mill does not argue *a priori* toward his first principle.
But, both philosophers derive their moral dictates from their first principles.
In this way, we can classify, very broadly and roughly, the traditional first-order moral theorists as foundationalists, in the limited sense that they defend a principled standard for morality, developing a theory which regiments or embodies that standard, and then they derive all particular moral dictates from those general principles.
The religious moralist, too, follows this general procedure, arguing from first principles (the perfect goodness of God) to particular moral claims (say, the Ten Commandments).

It is not the case that all philosophers before the late nineteenth century were only concerned with first-order moral theorizing.
Plato, in the *Euthyphro* and in the *Republic*, wondered about the possibility of morality.
For example, Thrasymachus argues, in the *Republic*, that the good is just what is in the interests of the powerful.
Hume argued for a fact/value distinction, that we can not derive an ought from an is, that there are no moral theories that derive strictly from the way the world is.
Hume claimed that we must start with a moral (or evaluative) premise in order to reach a moral (or evaluative) conclusion.
While this work (of both Plato and Hume) was profound, and perpetually influential, it did not derail later philosophers, specifically Kant and Mill, from proceeding with first-order ethics, from presenting and defending ethical theories.

III. The Birth of Meta-Ethics: A Tragedy

In the late nineteenth century, standard approaches to morality came under severe attack.
Most influentially, Nietzsche criticized the foundations of the standard ethical theories.
He denied Kant's basic claim that a universal moral theory was possible.
He derided Mill's concerns with the greatest good.
Instead, Nietzsche returned to a classical interest in one's own self-interest, and self-determination.
Whether one calls Nietzsche a nihilist, for denying universal principles, or one reads Nietzsche as presenting a more positive thesis, his negative attack on the possibility of a universal moral theory is clear.
In a sense, Nietzsche was the first of the twentieth-century second-order moral theorists.
He does not present a moral theory.
He does not tell us how to behave or how to determine how to behave.
He argues (if what Nietzsche does can be called arguing) that moral theory is impossible.
Like Plato's Thrasymachus, he is concerned with the possibility of moral theory, not with the dictates of moral theory.

In the early twentieth century, moral theorizing continued to turn from first-order to second-order, from determining basic principles of morality and their applications to the questions of what moral terms mean and whether first-order moral reasoning is possible.
With the linguistic turn in philosophy generally, moral philosophers started asking about the meaning of 'good', and related terms.
G.E. Moore's open question test (in *Principia Ethica*, 1903) denied the possibility of any naturalist ethical theory, like Mill's utilitarianism.
Mill's theory is naturalistic because it identifies the good with some property that we find in the world (viz. human pleasure).
In contrast, Moore thought that the good was separate from anything like happiness, conformity to social norms, or satisfaction of desires.
The good is a non-natural property, recognizable intuitively by our capacity for moral judgment.
Moore's anti-naturalism, like Hume's fact/value distinction, entails that moral theories can not have any non-moral grounding; we must start with ethical assumptions if we are to derive ethical conclusions.

Moore and the intuitionists who followed him claimed that our moral intuitions were just as secure as our empirical science.
Ethical assumptions, though, seem clearly liable to questions about legitimacy.
In a world of divergent cultures and divergent moral intuitions, Moore's claim that we have an inborn moral sense has come to seem implausible, presumptive, chauvinistic, and dogmatic.
Thus, Moore's work, though less virulent, and against his intentions, ended up denigrating moral theory, as traditionally conceived, just like Nietzsche's work.

Other philosophers in the early twentieth century responded to Moore by investigating the meaning of 'good', and related moral terms.
Non-cognitivists claimed that moral claims were not factual at all.
Some non-cognitivists, called emotivists, claimed that a statement like 'murder is wrong' actually is a (subjective) expression of distaste ("Murder? - Yech!") rather than an objective claim about the wrongness of murder.
Other non-cognitivists claimed that moral claims are just all false.
These error theorists (e.g. John Mackie) claimed that since 'wrong' in 'murder is wrong' is not a natural

property, and since it could not be a non-natural property (since we have no ability to know of non-natural properties) it must not be any property at all.

Thus, work in the late nineteenth- and early twentieth-centuries in ethics seemed to lead inexorably to the conclusion that first-order moral theorizing was fundamentally flawed.
Foundational principles, like the categorical imperative and the greatest happiness principle, seemed to be unknowable and indefensible.
Without secure foundational principles, there seemed to be no way to justify particular beliefs.

To make the claim that first-order ethics is lost without foundational moral theories more concrete, consider a controversial ethical claim, say that universal health care is a right.
How could we possibly defend such a claim?
We might appeal to notions of duty to others.
But, unless we can establish with certainty some fundamental theory of our rights and duties to others, such claims will be unconvincing.
Similarly, we might appeal to the benefits of universal health care.
But again, unless we are convinced of the importance of helping others, those benefits will not be morally motivating.
If we can not establish the meaning of 'good', and a groundwork for moral theorizing generally, all first-order moral theory seems pointless.

IV. Rawls and the Return of Ethics

Most moral philosophy for at least fifty years, from *Principia Ethica* until Rawls's *A Theory of Justice* (1971), focused directly on the problem of the nature of the good, on second-order (or meta-) ethics.
Rawls rescued traditional moral theorizing from the morass of abstract philosophizing.
He did so by developing a decision procedure for moral reasoning.
Relying heavily on Goodman's work, Rawls claimed that the foundational ethical principles that seemed impossible in light of the meta-ethical debates, were unnecessary as starting points.
Indeed, such a foundationalist position is implausible.

> It is obviously impossible to develop a substantive theory of justice founded solely on truths of logic and definition. The analysis of moral concepts and the a priori, however traditionally understood, is too slender a basis (51).

And:

> A conception of justice cannot be deduced from self-evident premises or conditions on principles; instead, it is a matter of the mutual support of many considerations, of everything fitting together into one coherent view (21).

Instead of starting with Kant's autonomy or Mill's hedonism, Rawls claims that we can do ethics by starting with some intuitive ethical judgments and working toward substantial ethical theories.
In particular, Rawls attempts to use the method of reflective equilibrium to settle questions about justice (as fairness) by starting with our intuitive judgments about what social arrangements are just.
Rawls wants to know if our society is just, in the sense of fairness, in basic rights and duties (responsibilities); specific principles of justice (how we make decisions, tax, etc.); and distribution of

social benefits.
Our goal, undertaken from what Rawls calls the original position, is to formulate the principles which guide the basic structure of society.
How should we distribute governing responsibilities?
What kinds of freedoms should we allow?
How do we determine tax rates, and other specific laws?
The principles of justice are not, in general, particular laws, but ways in which those laws are made.

In order to determine if our society is just, Rawls conducts a thought experiment.
How would we organize society, if we did not know who we were?
Rawls's hypothetical question resembles the question which motivates the social contract, which asked what kinds of rules we would set up if we were starting society.
The social contract theorists started by considering individuals in the state of nature.
Rawls starts his thought experiment by asking us to consider the original position, behind the veil of ignorance.
In the original position, we are rational, meaning that we pursue our ends as efficiently as we can.
We are mutually disinterested, meaning that we are primarily interested in our own ends, whatever they are.
We have a sense of justice, one that we assume develops naturally.
We are unaware of our race, sex, talents, conception of the good, religion, and social status; we do not know who we are.
By ignoring particular factors of our selves, we ensure that we do not choose principles of justice which favor only some members of society, and disadvantage others.
Also, we assume that in the original position all members of our society are roughly equal: we have the same rights and rational abilities.
Though we may have different values, no system of values is to be ranked higher than another.
So, we start with a fair situation.

Social contract theorists used the fiction of the state of nature to justify the rule of law or a particular government.
It's not always clear whether the social contract theorists take their stories to be fictions.
Indeed, were their stories about society developing civilized structures out of a state of nature to be true, their justifications of the rule of law would have most force.
So the social contract theorists ask us to pretend that their fictions are true.
For Rawls, in contrast, the original position is recognized explicitly as an instrumental hypothetical by all parties.

> In justice as fairness the original position of equality corresponds to the state of nature in the traditional theory of the social contract. This original position is not, of course, thought of as an actual historical state of affairs, much less as a primitive condition of culture. It is understood as a purely hypothetical situation characterized so as to lead to a certain conception of justice (12).

Once we are behind the veil of ignorance, we reason, from our original intuitions about justice to a full theory of justice.
Our original intuitions are limited and incomplete, in the sense that they do not determine difficult cases.
But, the final theory we construct will determine those cases.
For example, we can start our quest for a theory of justice with convictions that racial discrimination is unjust and that everyone should have access to clean water and a basic education, but not knowing which

tax rates are acceptable, or what kinds of distributions of wealth are unfair.
We develop a theory that generates the intuitive judgments that we accept.
Then, we proceed to apply it to the harder cases.

Rawls argues that we will arrive at two principles of justice.

1. A just society will insist on equality in basic rights and duties. Our basic rights and duties include the right to vote and hold office; freedom of speech, assembly, and thought; freedom from oppression and assault; the right to hold property and freedom from arbitrary search and seizure.
2. A just society will allow inequalities of social goods only if they are available to and benefit every one.

There are natural goods, like health and intelligence, which the state can not really regulate, and social goods, like money and opportunity, over which the state has some control.
Rawls argues that in the original position, we would accept some inequalities in social goods in addition to natural inequalities, but with limits.
In particular, he claims that we would choose inequalities according to a maximin principle: when choosing among various distributions of social goods, we should choose the one which maximizes the worst possibility (i.e. puts the least-well-off in the best position).
Thus, Rawls believes that we can allow inequalities in social goods, as long as every one benefits.

Working backwards, from intuitive judgments to general theories, we establish a theory of justice that yields our intuitive claims and which allows us to evaluate new or controversial arrangements.
We integrate new information, new data, by balancing our intuitive claims with our general theories of justice.
Working toward what Rawls called reflective equilibrium, we establish a theory which should be acceptable to all rational persons, and which does not require a fundamental commitment to any particular moral theory.

> In searching for the most favored description of this situation we work from both ends. We begin by describing it so that it represents generally shared and preferably weak conditions. We then see if these conditions are strong enough to yield a significant set of principles... We can either modify the account of the initial situation or we can revise our existing judgments, for even the judgments we take provisionally as fixed points are liable to revision. By going back and forth, sometimes altering the conditions of the contractual circumstances, at others withdrawing our judgments and conforming them to principle, I assume that eventually we shall find a description of the initial situation that both expresses reasonable conditions and yields principles which match our considered judgments duly pruned and adjusted. This state of affairs I refer to as reflective equilibrium (20).

V. The Linguistics Analogy

For support for his method, Rawls appeals to an analogy with linguistics.
We will look more carefully at the linguistics case in our next class.
For now, I will just mention that Chomsky posits a distinction between our competence with a language and performance with that language.

People often make grammatical errors, or have idiosyncratic ways of using a standard natural language, like English.
If our linguistic theories were required to account for all actual linguistic usage, if they were theories of actual performance, they would end up extremely messy and *ad hoc*.
Instead, we take linguistics to be a theory of linguistic competence.
Thus, we can abstract away from actual performance, and construct a neater theory of the language.
Our linguistic theory is still a theory of the language we use, but it is a theory of an idealized version of the language.
Similarly, says Rawls, our theory of justice is a theory of an idealized form of justice, one which may not match any one's intuitions precisely.
But, that idealization does not denigrate the theory as a theory of justice.

VI. Rationality and Reflective Equilibrium

Rawls, relying on Goodman's procedure, thus pulled a rabbit out of his hat.
He returned first-order moral theorizing to respectability by adapting the work we have already read about reasoning generally to moral reasoning.

One natural worry about Rawls's method is his reliance on what reasonable, or rational persons accept or learn in what he calls "a sense of justice [developed] under normal social circumstances" (46).
The assumption of rationality is closely related to another assumption, that reasonable persons will come to the same principles of justice.

> I shall not even ask whether the principles that characterize one person's considered judgments are the same as those that characterize another's. I shall take for granted that these principles are either approximately the same for persons whose judgments are in reflective equilibrium, or if not, that their judgments divide along a few main lines... (50).

The assumption of rationality seems defensible.
It is methodologically similar to the scientific principle of discounting outliers.
We want the principles of justice to be universally acceptable, but we will settle for acceptability by only rational persons.
In effect, the criterion of rationality is constitutive of personhood, so this criterion is nearly definitional.
Similarly, Rawls demands, "requisite intellectual capacity" (46).
On the other hand, we have to be sure that we do not eliminate dissenters from the principles of justice.
We have to be careful not to silence particular kinds of dissent by calling them irrational.

> The concept of rationality must be interpreted as far as possible in the narrow sense, standard in economic theory, of taking the most effective means to given ends... One must try to avoid introducing into it any controversial ethical elements (14)

The commitment to rationality plays a specific role when we have to decide among various proposed principles of justice.
We have to determine whether one conception of justice is more reasonable, more acceptable, than another.
Rawls points out that our need to evaluate competing principles involves theories of rational choice.

So, the two prices we pay for returning to first-order moral theory are, first, that we must rely on our intuitions about justice (as starting points for moral theorizing) and, second, that we must rely on a theory of rational choice, a scientific theory of how to decide among given options.
Both prices are defensible, commonsensically.
But, they are both liable to empirical criticism.
What would happen to Rawls's theory if it were shown that our moral intuitions were systematically corrupt?
What would happen to Rawls's theory if it were shown that we tend to make irrational decisions?

VII. Ethics and Science

Here is one more question to ponder.
Rawls relies on Goodman's procedure, which was taken as appropriate in philosophy of science, and in epistemology.
But, there seems to be an important difference between moral philosophy and science.

> Moral philosophy is Socratic: we may want to change our present considered judgments once their regulative principles are brought to light. And we may want to do this even those these principles are a perfect fit. A knowledge of these principles may suggest further reflections that lead us to revise our judgments... There is a contrast with physics. To take an extreme case, if we have an accurate account of the motions of the heavenly bodies that we do not find appealing, we cannot alter these motions to conform to a more attractive theory. It is simply good fortune that the principles of celestial mechanics have their intellectual beauty (49).

What is the significance of this distinction?