Phil 115, June 13, 2007
The argument from the original position:
set-up and intuitive presentation and the two principles over average utility

What is the role *of the original position in Rawls's theory?*

- On the *alchemy* interpretation: an argument that can be used to convince everyone, on the basis of uncontroversial platitudes about justice, that Rawls's conception is correct.
- On the *system* interpretation: Rawls proposes to ground an alternative to utilitarianism on a different conception of persons and society. The OP is a way of articulating what follows from viewing society as a fair system of cooperation among free and equal persons. The original position is a "device of representation."

How does Rawls characterize the initial situation, the "original position"?
The list of alternatives:

A. Two principles of justice
    1. Greatest equal liberty
    2. (a) Fair equality of opportunity
       (b) Difference principle
B. Mixed conceptions: Greatest equal liberty, but instead of fair equality and DP:
    1. Average utility
    2. Average utility, subject to:
        (a) a social minimum
        (b) a constraint on the degree of inequality
    3. Average utility, subject to (a) or (b) and fair equality of opportunity
C. Classical Teleological conceptions
    1. Classical (=total) utility
    2. Average utility
    3. Perfectionism: a teleological theory, where the good is not the satisfaction of desire, but instead something more objective: virtue, human excellence, etc.
D. Intuitionist conceptions: various ways of balancing a plurality of first principles
E. Egoistic conceptions
    1. First-person dictatorship: Everyone is to serve my interests
    2. Free-rider: Everyone but me is to act justly
    3. General: Everyone may do whatever is in his interests

The circumstances of justice:
The parties *do* know is that their society is in the "circumstances of justice." The circumstances make a system of cooperation regulated by a conception of justice both *possible* and *necessary*.

The "objective" circumstances:

(i) There is rough equality of persons (in Hobbes's sense), so that none can dominate the rest. Hence each must rely, to some extent, on the willing cooperation of others.

(ii) There is moderate scarcity (in Hume's sense): Natural and other resources are neither too plentiful, nor too scarce. If resources were too plentiful, as in the Garden of Eden, then there would be no need for cooperation, or cooperation would be so effortlessly productive that there would be no conflict over how its benefits were distributed. If

resources were too scarce, as in the case of a sinking ship, then there would be no benefit from cooperation, or no real possibility of stable cooperation at all.

The "subjective" circumstances:

(i) There is an *identity* of interests. People's needs and interests are sufficiently complementary to make mutually beneficial cooperation possible.

(ii) But there is also a *conflict* of interests. People have different conceptions of the good.

Formal constraints on principles:

Rawls suggests that some of the entries on the list of conceptions can be dismissed out of hand, because they violate certain formal constraints on principles of justice:

1. The first is generality. Principles should not include proper names or rigged definite descriptions. Why? The principles are supposed to hold unconditionally in perpetuity. They must be understandable to any generation, and so cannot require knowledge of particular persons or groups.
2. The second constraint is universality. Principles should apply to everyone. Among other things, this implies (i) that the principles must be simple enough for everyone to follow, and (ii) that everyone could follow the principles without undermining the point of the principle.
3. The third constraint is publicity. The parties are to assume that it will be mutual knowledge that the principles are universally accepted. Note that publicity is often *rejected* by utilitarians.
4. The fourth constraint is ordering. The principles must order conflicting claims. They should be complete and transitive.
5. The fifth and last constraint is, appropriately enough, finality. The principles are supposed to be "the final court of appeal in practical reasoning" (116).

These formal conditions rule out the egoistic principles.

- Free-rider and dictatorship violate generality.
- General egoism violates ordering.

The veil of ignorance:

- First, the parties do not know their own (i) social position, (ii) talents, (iii) conception of the good, or (iv) special features of their psychology, such as their aversion to risk.
- Second, the parties do not know specific information about their society. They do not know its level of development. But they do know that their society is in the circumstances of justice. Furthermore, they know the general facts about human societies, such as that a given conception of justice is stable.
- The veil of ignorance represents persons as equal. They are all symmetrically situated.
- The veil of ignorance also makes possible a unanimous choice.

The interests of the parties:

- Parties choose on the basis of primary goods: all-purpose means for pursuing one's conception of the good.
- "Of course, it may turn out, once the veil of ignorance is removed, that some of them for religious or other reasons may not, in fact, want more of these goods. But from the

standpoint of the original position, it is rational for the parties to suppose that they do want a larger share, since in any case, they are not compelled to accept more if they do not wish to" (123).

- The parties are "mutually disinterested." They do not object to others' having more *per se*. Nor do they desire that others have less *per se*. They care *per se* only about what they themselves have.

The strains of commitment:

- Each chooser also knows that he has, and that all the others have, a capacity for a sense of justice. The parties, therefore, "can rely on each other to understand and act in accordance with whatever principles are finally agreed to" (125).
- A consequence is that "they will not enter into agreements they know they cannot keep, or can do so only with great difficulty" (126). The parties take into account what Rawls calls the "strains of commitment." They *cannot* choose a set of principles if they believe that there is even a *chance* that once the veil is lifted, they will not be able to abide by those principles.
- For example, as parties in the original position, we cannot choose principles that would allow a ban on minority religions, because we know that we might not be able to live up to such principles. To choose such principles would be like negotiating in bad faith.

The intuitive argument for the two principles:

- Consider a single individual. There is no way for him to gain special advantages for himself.
- Nor does he have any reason to accept special disadvantages.
- This leaves him with an equal share.
- The unanimous choice seems to be a principle dividing social primary goods equally. This gives us equal liberty and fair equality of opportunity.
- However, some inequalities in certain primary social goods—income, wealth, authority, and responsibility—may increase everyone's absolute share of those goods, as measured from the initial benchmark of equality.
- Since he is rational, he has reason to want this. Since he is not moved by envy, he has no reason not to want this.
- "Because the parties start from an equal division of all social primary goods, those who benefit least have, so to speak, a veto" (131).
- This gives us the difference principle. Inequalities in these primary goods are acceptable, so long as they give the person with the least of these goods as much of these goods as possible.

Rawls's argument that the parties would reject classical utilitarianism:
Rawls argues that the parties would reject classical (=sum total) utilitarianism straightaway. This is because it has unacceptable consequences when applied to population policy.

Increasing the population has two effects on the sum of happiness.

- On the one hand, it adds new people, who may lead happy lives. This tends to increase total happiness.

- On the other hand, it reduces the resources available to the people already around, and hence lowers their happiness. This tends to decrease total happiness.

So long as the first effect—the boost that comes from adding new people—outweighs the second effect—the loss to the old people—classical utilitarianism demands that we add new people.

Yet it seems intuitively wrong, and, in any event, it would be rejected in the original position. "Since the parties aim to advance their own interests, they have no desire in any event to maximize the sum total of satisfaction" (141).

Notice here how the OP embodies a different conception of society and persons—as separate—and how this leads to a straightforward rejection of at least classical utilitarianism.

Average utilitarianism does not have this consequence. Average utilitarianism directs us to produce the greatest average of happiness. When adding people would lower the average level of happiness, average utilitarianism directs us *not* to add people.

The argument that the parties would choose average utilitarianism:
In fact, Rawls offers a *very* plausible argument for average utility.

- Consider any individual in the original position.
- She does not know who in society she will turn out to be.
- She knows that there is some (natural) number of people, n, but of course she does not know what number n is.
- If she becomes person 1, whoever that is, then her utility will be U1, whatever that is; if she becomes person 2, whoever that is, then her utility will be U2, whatever that is; and so on.
- So the expected value she confronts is:
  P1*U1 + P2*U2 +…+Pn*Un, where P1 is the probability of being person 1, and so on.
- Suppose that our individual chooser assumes that she has an equal chance of becoming anyone. Then the chance of becoming any particular person is 1 divided by n, and the expected value she confronts is:
  1/n*U1 + 1/n*U2 + … 1/n*Un.
- Now this is equivalent to:
  (U1 + U2 + … + Un)/n.
- Finally, suppose that what it is rational for the individual chooser to do is to maximize expected value.
- Then she will want to choose whatever principle makes (U1 + U2 + … + Un)/n as large as possible.
- This is the principle of average utility.

Therefore, someone in the original position who has reason
(i) to believe that she has an equal chance of becoming anyone and
(ii) to maximize her expected utility
will choose average utilitarianism, not Rawls's two principles.

Rawls's objections to the argument for average utilitarianism:
There are "no objective grounds in the initial situation for assuming that one has an equal chance of turning out to be anybody. That is, this assumption is not founded upon known features of one's society" (146).

Therefore, the only basis for the assumption of equiprobability is the principle of insufficient reason: in the absence of any reason to believe otherwise, we ought to assume, at least initially, that every outcome is equally likely.

Rawls argues that it is irrational apply the principle of insufficient reason in this case. In learning situations, the principle of insufficient reason makes sense, because it is an unbiased starting point. But the OP is *not* a learning situation. It is once and for all.

Moreover, in the average utilitarian argument, "the individual is thought to choose as if he has no aims at all which he counts as his own. He takes a chance on being any one of a number of persons complete with each individual's systems of ends, abilities, and social position. We may doubt whether this expectation is a meaningful one. Since there is no one scheme of aims by which its estimates have been arrived at, it lacks the necessary unity" (150). I can compare how I would like, given my final ends, being in your position to how I like, given my final ends, being in my position. But how am I to compare how I would like, if I had *your* final ends, being in your position, to how I like, given *my* final ends, being in my position? We seem to need a *fixed set* of final ends to evaluate different lives.

Notice that the average utilitarian might get around this problem by assuming hedonism: that *pleasure* is the only real end. But this in effect denies that people's final ends really are their *final ends*.

Rawls's maximin argument that the parties would choose his two principles:
The "maximin solution" is the option with the *best worst* outcome.

	Circumstances		
Decisions	C1	C2	C3
D1	-7	8	12
D2	-8	7	14
D3	5	6	8

Which is the maximin choice? Does the maximin choice change if the outcomes in C2 and C3 change? Does it change if the relative probabilities of C1, C2, and C3 change?

Rawls's argument for the two principles is then as follows:

(A) Although the maximin rule is not generally an appropriate rule for choice under uncertainty, it is appropriate when three conditions obtain.
(B) These three conditions obtain in the original position.
(C) The two principles provide a better worst outcome than the alternative principles. (I.e., the two principles are the maximin choice.)
(D) Therefore, the parties would choose the two principles in the original position.

As (A) concedes, maximin is not generally an appropriate rule:

	Heads	Tails
Bet 1	1 cent	$1,000,000
Bet 2	2 cents	$1

(A) claims only that maximin is appropriate when the following three conditions are met:

(i) One has no knowledge of the probabilities of various outcomes. So it makes sense to disregard probabilities, as maximin does.

(ii) One cares little for gains above the best worst outcome, which is guaranteed by making the maximin choice. I.e., maximin ensures a satisfactory minimum.

(iii) The other options have outcomes that one cannot accept.

Premise (B) asserts that in the original position, these three conditions are met "to a very high degree," with premise (C) claiming that the two principles represent the maximin choice.

(i) The veil of ignorance entails that parties have no basis for calculating probabilities.

(ii) We can see that the two principles provide a satisfactory minimum by considering what a society regulated by the two principles would be like.

(iii) Utilitarianism, for example, might justify slavery and serfdom and almost certainly justifies religious persecution. These are intolerable outcomes.

Consideration of the "strains of commitment" reinforces the parties' focus on the maximin choice.

Utilitarian responses to each of the alleged conditions that are supposed to make the choice of the two principles rational:

(i) The parties lack knowledge that they have an equal chance of being anyone only because Rawls *stipulates* that they lack that knowledge. Why not give them that knowledge? So long as they do not know who they will be, their choices will still be impartial and fair.

(ii) The claim that the two principles provide an adequate minimum rests on two strong empirical assumptions: (a) that human interests have a kind of threshold or cutoff point, such that reaching the threshold means everything and surpassing it means nothing, and (b) that the stock of resources in society is sufficient to raise everyone up to the threshold.

   a. First, these empirical assumptions seem highly implausible.
   b. Second, knowledge of them is excluded by the veil of ignorance.
   c. Finally, if there is such a cutoff point and if there are sufficient resources to raise everyone up to it, then average utility is maximized by doing so. Average utilitarianism will *also* guarantee this satisfactory minimum.

(iii) First, utilitarianism probably would not have intolerable outcomes. Utilitarianism would condone slavery only in bad conditions that are unlikely to occur. Second, the two principles would also justify slavery the same conditions. Recall that when things are really bad, the two principles give way to the general conception, which permits sacrifices of liberty for the sake of material benefits.