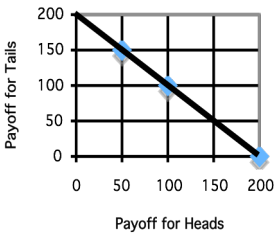
RAWLS AND RISK AVERSION

DREW SCHROEDER – PREPARED FOR MR-22, DEC. 2007

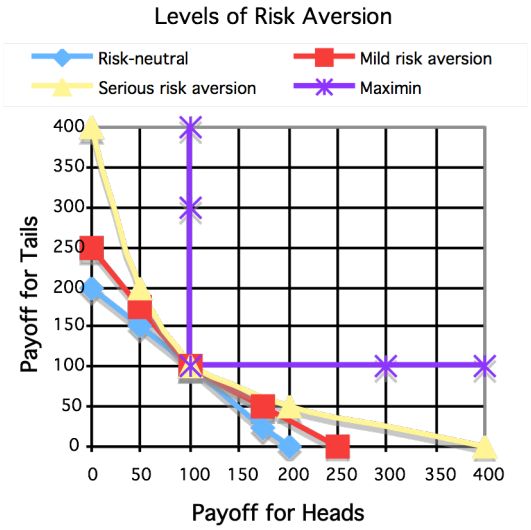
I. MAXIMIN AS MAXIMAL RISK AVERSION

Many of you have spoken to me about Rawls's argument from the original position, claiming Rawls is crazy to think that in such a position we'd all be in favor of the Difference Principle. This seems like a pretty good objection. For example, suppose there are three relevant groups in the population (A, B, and C), and consider the two distributions to the right, given in terms of dollars above the poverty line to each member of the group. From behind the veil of ignorance, which distribution would you favor? Factoring out the possibility of envy (which Rawls says is to be disregarded in the original position (124)) and remembering that all parties behind the veil of ignorance are supposed to be self-interested, I'd wager that 100% of you would choose #1. If that's right, though, then isn't Rawls wrong? Isn't it *false* that agents behind the veil of ignorance would choose to maximize the position of the worst off? The economist John Harsanyi actually used a device very much like Rawls's veil of ignorance before Rawls did, and he used it to justify utilitarianism. That is, Harsanyi argued that the parties behind a veil of ignorance would agree to the distribution that maximized overall utility.

	A	B	C
#1	$10	$1000	$9000
#2	$11	$11	$11



Harsanyi, along with many others writing about Rawls, complains that Rawls assumes the parties to the original position would be risk-averse – in fact, that they'd be *maximally* risk-averse. We can see why they say this by considering a simpler situation. Suppose that I'm about to flip a coin and I offer you a choice of gambles. If we simplify matters greatly and assume that money equates to utility, a typical expected utility calculation says you ought to be indifferent to being guaranteed $100, getting $200 if heads but nothing if tails, and getting $150 if tails and $50 if heads. The expected value of all those options is $100. In fact, you ought to be indifferent to all points on the line x+y=200. Call that the *indifference curve*. (See the graph to the left.) Most people, though, are not indifferent between those gambles. They display a level of *risk aversion*, especially in cases where the stakes are high. Most people prefer the guaranteed $100 to the 50% shot at $200. They'd give up the sure thing only to get, say, a 50% shot at $250. A more seriously risk-averse person might give up the $100 only for a 50% shot at $400. From those "data points" (and a few more), we can add some risk-averse indifference curves to the graph. Now, think about what the indifference curve would look like for someone out to make the worst outcome as good as possible – in other words, someone out to *maxi*mize the *min*imum value. Call this decision procedure "maximin". That person wouldn't accept any gamble unless it guaranteed her at least $100. She wouldn't, for example, take a gamble that paid $1000 for heads but only $99 for tails, since the worst possible outcome there ($99) is worse than $100. Her indifference curve around our original $100 guarantee would therefore make a right angle. Looking at the graph, we can see why adopting maximin amounts to having an infinite level of risk aversion.[1] Maximin, however, is just Rawls's Difference Principal.



Let's bring this all back to distributive justice.

[1] Could we have an acute indifference curve? We could. That would mean refusing to give up the guaranteed $100 for a gamble which gave $100 for tails and $200 if heads. You'd give up the $100 only for a gamble that paid more than $100 for *both* heads and tails – for example, one that paid $105 for heads and $110 for tails. And of course you'd prefer a guaranteed $105 to that gamble. In other words, it'd amount to being *so* risk-averse that you would reject a gamble that you knew would pay you more, just because you didn't like the idea of a gamble. There's a sense, then, in which one could be more risk-averse than Maximin, but it's a pretty crazy option, so I'll ignore it.

Harsanyi is positing that in the original position people would have no level of risk aversion. Rawls seems to be assuming an infinite level of risk aversion. What should we say? People do in fact get more risk-averse the larger the stakes. The original position is in many senses the biggest-stakes gamble possible: your whole life's prospects are being determined! So we'd expect people to be risk averse to some extent. We could claim that risk aversion is irrational and that the parties to the original position would therefore be risk-neutral, but that doesn't seem especially plausible. There doesn't seem to be anything wrong with being risk-averse. That said, infinite risk aversion seems crazy. There are surely *some* risks that anyone would take – say, a guaranteed dollar vs. a 99% chance at a million. So, neither risk-neutrality (Harsanyi) nor maximal risk-aversion (Rawls) seems plausible. In order to arrive at a principle of justice, we apparently need to figure out the appropriate level of risk aversion in the original position.

Rawls sometimes suggests that this is the way to proceed. He says, for example, that "from the standpoint of the original position, the parties will surely be very considerably risk-averse" (1974 143). I don't, however, think it's Rawls's primary strategy for two reasons. For one, even if we agree that some level of risk aversion is rational, it's absurd to think that *maximal* or *infinite* risk aversion is that level. Surely Rawls wasn't crazy, so we can conclude that this isn't how he meant to pursue the argument. Second, there doesn't seem to be a single rational level of risk aversion. If multiple levels of risk aversion can be rational, then how can we privilege one in the original position? In fact, Rawls says that one of the things blocked by the veil of ignorance is knowledge of one's level of risk aversion (149). It would be quite odd to prevent people from knowing their level of risk aversion and then to ask them to make a decision that requires specifying a level of risk aversion, when more than one level could be rational. For these reasons, it seems to make sense to ask whether we can make out an argument for the Difference Principle without getting tied up in questions of risk aversion.

II. MAXIMIN AS A RESPONSE TO IGNORANCE

It turns out we can make some headway, at least. The key is to note that Rawls's veil is one of ignorance, not mere risk or uncertainty. Recall the two distributions #1 and #2, above. In order to conclude that #1 is preferable to #2, we seem to require the assumption that the agent is equally likely to end up in group B or C, as opposed to group A. (Recall that Rawls stipulates that the parties to the original position are self-interested.) Or, rather, we need to assume that it's not overwhelmingly likely that the agent will be in group A. If she were virtually guaranteed to be in group A, then #2 would be the smarter choice – at least she'd have $11, instead of $10. But, Rawls says (§28), he's never said anything about how the positions in society are to be distributed. On what grounds can we take it for granted that we might end up in B or C? The assumption of equiprobability – which seems to drive Harsanyi's argument – is unjustified. In fact, *any* assumption about probabilities is unjustified. We can't even begin to calculate expected utility, since we can't begin to assign probabilities to the various possible outcomes. If we can't even begin to calculate expected utility, then the whole notion of risk aversion doesn't get a grip.

Let's make this slightly more concrete. Suppose I propose to you two distributions of grades for the class: on the first, we'll have 18 As and 2 Ds. On the second, we'll have 20 D+s. From a self-interested point of view, which should you choose? The first might seem to be the obvious choice, but what if I reminded you that I haven't specified how the grades will be distributed. Suppose your grade was to be given to you by your worst enemy, or suppose I told you you'd get last pick of grades. Then, the second distribution would be the smart choice. In order to begin any kind of calculation about what's in your best interest, you need to have some idea of how likely it is that grades will be assigned randomly, instead of by your worst enemy. But you have no idea which of those is more likely, so you seem to be paralyzed. It's not clear what the right thing to do is, but it's clear that it's *not* obviously to take the first distribution, merely because the overall average is higher. The situation behind the veil of ignorance, Rawls says, is like this. If it were a veil of *risk*, and you knew that positions would be randomly assigned, then perhaps some moderately risk-averse expected utility calculation would be appropriate. Or *perhaps* Harsanyi's risk-neutral utilitarianism would be chosen. It's a veil of ignorance, however, and so you're supposed to have no idea how likely it is that you'll end up in any given position. The assumption that you're equally likely to end up in each position is unjustified.

Now, this argument may do a good job of showing why it would be inappropriate to engage in a calculation of expected utility (even a somewhat risk-averse one), but it doesn't even begin to explain why maximin, or the Difference Principle, is the appropriate choice from behind the veil – even if maximizing average utility is *wrong*, that doesn't mean maximin is *right*. This points to a general problem with decision and game theory, branches of mathematics/economics/philosophy that purport to describe how ideally rational agents would act in certain well-defined circumstances. There are good theories of how to behave when interacting with other agents, who are either cooperating with you or working against you. There are good theories of how to behave when facing a situation where one has some even very imperfect idea of the likelihood of various outcomes. There isn't, however, an uncontroversial theory of how one ought to behave when one has no idea of how likely various outcomes are.

It's hard to even know how to proceed. The typical approach begins by positing a series of criteria a decision procedure ought to satisfy. For example, one act *weakly dominates* another when for all possible outcomes, the first is either just as good or better than the second, and for at least one outcome it's better. Suppose you're offered two bets: (1) if the coin is heads you win $50, tails you win $25, or (2) if the coin is heads you win $75, if tails you win $25. Even if you have no idea how likely heads is to come up – you have no idea if it's a fair coin – you'd surely choose (2). You'd do better if heads came up, and just as good if tails came up. Any decision procedure that had you pick (1), or that left you indifferent between (1) and (2) would be a bad decision procedure. Therefore, one of the constraints on a decision procedure is that it should never choose acts that are weakly dominated by other acts.[2] By coming up with a series of innocuous-sounding criteria, we can actually achieve some surprisingly concrete results. Luce and Raiffa give about a dozen plausible criteria. Unfortunately, they're mutually inconsistent: no decision procedure can satisfy all of them. Rejecting one or another of the criteria does, however, yield a determinate decision procedure. Among them are both maximin and the principle of insufficient reason. We know about the former. The latter proposes that, when one is completely ignorant about which of several states of affairs obtains or will obtain, one should arbitrarily assign equal probabilities to each state. Finally, then, we come to much more sensible arguments for both Harsanyi's and Rawls's positions. Rawls wasn't implausibly assuming people would be maximally risk averse in the original position. Harsanyi wasn't misinterpreting Rawls's veil as one of risk instead of ignorance, and he wasn't implausibly assuming that people behind the veil would be completely risk-neutral. Instead, Rawls suggests that the proper decision procedure under ignorance is maximin. Harsanyi thinks it's the principle of insufficient reason. (For more about this debate, see the end of this document.)

III. MAXIMIN AS UNCERTAINTY-AVERSION

Here's a related, though slightly different, way of thinking about the situation, due to Susan Hurley. Hurley suggests that risk aversion and uncertainty (or ignorance) aversion are separate things. Being averse to a known gamble is different than being averse to entering a situation where one has no information about what the odds are. Experimentally, people do display uncertainty aversion independent of risk aversion. Consider a result known as Ellsberg's Paradox (described in Daniel Ellsberg's Harvard Ph.D. dissertation in 1959). Suppose there is a bag containing 90 balls. You know that 30 are red and the other 60 are a mix of black and yellow, of unknown proportion. One ball is to be drawn from the bag at random. You're offered a choice of two bets: (a) you'll win $100 if the ball is red and nothing otherwise, or (b) you'll win $100 if it's black and nothing otherwise. Most people take (a). Now, the bag is refilled as before, and you're offered two different bets: (c) $100 if the ball is red or yellow, or (d) $100 if the ball is black or yellow. Most people – including the same people that chose (a) – choose (d). This should be puzzling. In preferring (a) to (b), people seem to be assigning greater probability to drawing a red than drawing a black ball. That is, they seem to be behaving as if they believe there are more red than black balls. In preferring (d) to (c), however, they seem to be supposing that there are more black balls than red

[2] Note that this already rules out the simple maximin approach: both options have minimum value $25, so both are equal according to simple maximin. That suggests we need a revised principle, *lexical maximin*, which begins by comparing the worst outcomes, then compares the second-worst outcome, and so forth. Lexical maximin appropriately prefers (2) to (1). Rawls' Difference Principle should be understood as an example of lexical maximin, not simple maximin.

balls. The point isn't about whether or not there are more red or black balls – we've stipulated that that's unknown. Rather, the thought is that *if* you prefer (a) to (b), *then* you ought to also prefer (c) to (d).

This causes a problem for decision theory, which generally requires that an agent's preferences be consistent. What should we say? Does Ellsberg's Paradox simply show that people are irrational? It might, but Hurley and others propose a different explanation: people are averse to uncertainty. That is, they're averse to gambling on situations for which they don't know the likelihood of winning. This is different than risk aversion. A *risk-averse* person doesn't like to gamble in situations where she *knows* the odds, if there's a chance of loss. An *uncertainty-averse* person doesn't like gambling on situations when she *doesn't know* the odds. Positing this uncertainty aversion solves the Paradox. Notice that the gamble (a) has a known 1/3 chance of paying off. Gamble (b) has unknown odds (since the number of black balls is unknown). An uncertainty-averse person will therefore prefer gamble (a). Gamble (b) might have better or worse odds than gamble (a), but at least the odds of (a) are known. Gamble (c) has unknown odds, but gamble (d) has a 2/3 chance of paying off (since the number of black+yellow balls is known). Therefore, an uncertainty-averse person will prefer (d) to (c). We can thus explain people's preferences for (a) and (d) in a consistent way, by positing uncertainty-aversion.

Okay, so how does this all apply to the original position? Since the veil is one of ignorance and not of risk – since agents don't know how likely they are to end up in any given position – an uncertainty-averse person won't like the situation. She can't, however, opt out, so how can she best minimize the uncertainty? The best option would be to insist on an equal distribution. Though she would still be uncertain as to exactly which role she'd end up in, she'd at least know exactly how much stuff she'd have. The uncertainty, in other words, would have no consequence. So, as a first step, we can say that the uncertainty-averse person would prefer whatever equal distribution gives the most to each person – that is, the maximal equal distribution. Call that distribution E. Now, compare E to some other distribution that weakly dominates it. That is, suppose E' assigns at least as much to every role as does E, and assigns more to at least some role. As we saw earlier, it's plausible to suppose that any rational decision procedure will prefer E' to E. If we suppose, therefore, that the uncertainty-averse person will prefer E' to E, we've basically arrived at a maximin decision procedure. An uncertainty-averse person who also prefers dominating distributions to dominated distributions will settle on the Difference Principle from the behind the veil of ignorance. Notice that we didn't need to assume anything about risk aversion, and in particular we arrived at maximin without (implausibly) assuming that the parties to the original position are maximally risk-averse.

IV. FURTHER READING

Luce and Raiffa's *Games and Decisions* (Wiley & Sons, 1st printing 1957) is a classic text. It's a bit technical, but definitely comprehensible to non-math majors. The stuff on decisions under uncertainty is in chapter 13.

Resnik's *Choices: an Introduction to Decision Theory* (U of MN press, 1987) is written by a philosopher and goes out of its way to make these concepts comprehensible to non-mathematicians. He also explicitly discusses the philosophical consequences of various aspects of decision theory. He talks about Rawls, Harsanyi, and the veil of ignorance at 40-43.

Harsanyi first proposes something like the Veil of Ignorance in "Cardinal Utility in Welfare Economics in the Theory of Risk-taking" (*The Journal of Political Economy* 61 (5), 1953). He explicitly takes on Rawls (and argues that the veil of ignorance would lead to utilitarianism) in "Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory" (*The American Political Science Review* 69 (2), 1975). Harsanyi explains some of the more significant objections to maximin, but he also misinterprets Rawls in key respects.

Rawls, in addition to the passages in *A Theory of Justice* (see especially §28), discusses the justification for maximin in an article, "Some Reasons for the Maximin Criterion" (*The American Economic Review* 64 (2), 1974). His arguments there aren't really decision-theoretic. Rather, they focus on certain practical ideas. (For example: maximin is much easier to apply and more likely to gain universal acceptance than the principle of average utility.) The more persuasive arguments in this article seem to me to be more closely tied to his second argument for the two principles: that our natural assets are arbitrary from a moral point of view.

Hurley's argument is presented in her book *Natural Reasons* (Oxford UP, 1992) beginning at 373, and also in a shorter form in "Cognitivism in Political Philosophy" (*Well-Being and Morality*, ed. Crisp and Hooker, Oxford UP, 2000).