TAKING RISKS BEHIND THE VEIL OF IGNORANCE

Lara Buchak, UC Berkeley

> "The usefulness of [the concept of the original position] depends on its being combined with a satisfactory decision rule." (John C. Harsanyi)

1. Introduction

This paper provides a new argument for a natural view in distributive ethics: that the interests of the relatively worse off matter more than the interests of the relatively better off, in the sense that it is more important to give some benefit to those that are worse off than it is to give that same benefit to those that are better off, and that it is sometimes (but not always) more important to give a smaller benefit to the worse off than to give a larger benefit to those better off. I will refer to this position as relative prioritarianism. The formal realization of this position is known as weighted-rank utilitarianism or the Gini social welfare function, and it is typically classified as an egalitarian view, though for reasons I will mention that classification may be misleading.

The argument takes as its starting point the proposal, due to Harsanyi and Rawls, that facts about distributive ethics are discerned from individual preferences in the "original position." I adopt Harsanyi's framework and draw on recent work in decision theory to argue for relative prioritarianism, which is a position intermediate between those that Harsanyi and Rawls each argue for: whereas Harsanyi holds that each individual's well-being matters equally to the evaluation of a social distribution, and Rawls holds that only the well-being of the worst off matters, relative prioritarianism holds that the well-being of the relatively worse off counts for more than that of the relatively better off but that everyone's well-being counts for something. I explain how this distribution might be justified to members of a society. Finally, I explain how this argument avoids two worries associated with theories that link intrapersonal and interpersonal decisions.

2. Decisions and Social Choices

Decision theory concerns the evaluation of gambles, where a gamble specifies the outcome or life-path a given individual gets in each possible state of the world: for example, {HEADS, a short and difficult life; TAILS, a long and happy life}. The question for decision theory is how to aggregate the values of the outcomes that are realized in each state, in order to arrive at a value for the gamble. Social choice theory

concerns the evaluation of social distributions, where a social distribution specifies the outcome a given state of the world yields for each individual in a social group: for example, {ALICE, a short and difficult life; BOB, a long and happy life}. The question for social choice theory is how to aggregate the values of the outcomes that go to each person, in order to arrive at a value for the social distribution.

There is a structural analogy between gambles and distributions: the role played by states in decision theory is played by individuals in social choice theory. And the aggregation question is analogous: once we know which values are to be found in various "positions" (states of the world or individuals), we want to know how to aggregate these values to determine a value for the whole. Thus it is unsurprising that analogous answers have been given to the two questions. Three such answers will be important to this paper. I will explain each of them using an example; the general equations can be found in the Appendix. We will assume a fixed population and no risk in the social case. The assumption of a fixed population implies that the "average" and "total" formulation of social choice rules produce equivalent rankings of distributions.

Consider the decision whether to quit one's dull but stable job and instead work at a risky start-up. If the company is amazingly successful (probability 0.01) then one will be very wealthy, feel a strong sense of personal accomplishment, have a short and pleasant workday, and be able to travel the world and enjoy the finer things in life (outcome A). If the company is very successful (probability 0.29) then one will be fairly wealthy, feel accomplished, and have an enjoyable workday (outcome B). If the company is moderately successful (probably 0.5), then one will have enough money to pay the bills, but the hours will be long and boring (outcome C). If the company fails (probably 0.2), then one will have to instead get an unpleasant, demanding job where one is merely scraping by (outcome D).

Or consider a policy (Policy X) that will lead to a particular distribution of life-paths. 1% of the population is very wealthy, fulfilled, enjoys a short and pleasant workday, and travels and dines out regularly. 29% of the population is wealthy, fairly fulfilled, and has an enjoyable workday. 50% of the population has enough money, and works long and boring hours. 20% of the population is merely scraping by at unpleasant, demanding jobs.

Utility is a measure of how valuable each outcome is to the individual for whom it obtains. There are two views of how utility is determined—on the one hand, intuitively and in advance of preferences, and on the other, from preferences themselves—but the differences between these views won't matter for this paper.[1] We will assume for the sake of our example that everyone has the same utility function. So, let us assign

u(A) = 1, u(B) = 4, u(C) = 6, u(D) = 7. We can then represent the two situations graphically, where "proportion" stands for the probability of each state in the individual case, and for the proportion of people in each group in the social case:

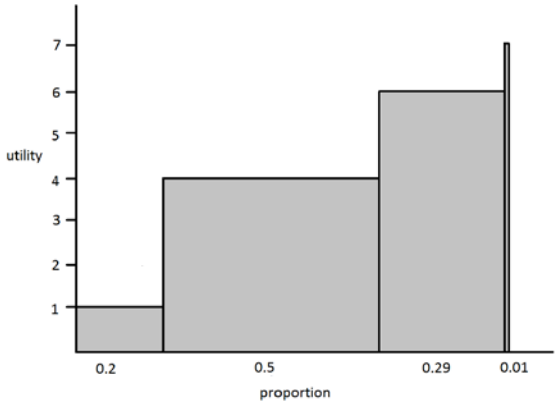

FIGURE 1: Start-up Gamble and Policy X

The first pair of aggregation rules is the individual decision rule *maximize expected utility* and the analogous social choice rule *average utilitarianism*. These rules take a weighted average of the possible utility values, each value (the height of each bar in the above graph) weighted by the proportion of positions that realize it (the width of each bar). Thus, the expected utility of working at the start-up, and the average utility of the society in which Policy X is implemented, is the area under the curve:

EU(Start-up) = (0.2)(1) + (0.5)(4) + (0.29)(6) + (0.01)(7) = 4.01

U(Policy X) = (0.2)(1) + (0.5)(4) + (0.29)(6) + (0.01)(7) = 4.01

According to these rules, equiprobable states get equal weight, and the interests of each person get the same weight as those of each other person. Furthermore, since a higher number corresponds to the gamble an individual should prefer or the distribution that is better, we can compare the gamble or the distribution with alternatives. Assume one's current job has utility 4 for certain, and Policy Y yields an outcome of utility 4 to everyone. Then one should prefer the start-up, and Policy X is socially better.

The second pair of rules is the individual decision rule *maximin* and the social choice rule *maximin* (sometimes called *maximin equity*). Both rules say to choose the distribution that maximizes the minimum utility value—the utility value in the worst state in the individual case, and the utility value to the worst-off person in the social case:

MAXIMIN(Start-up) = 1

MAXIMIN(Policy X) = 1

These equations represent the height of the first bar in the above graph, or the area under the first bar if its width were stretched to one. Thus, the worst state gets all the weight in evaluating the gamble, and the interests of the worst-off person count exclusively. According to maximin, one should prefer to stay at one's current job; and according to maximin equity, Policy Y is socially better.

Finally, we will be concerned with a decision theory and a social choice theory that both fall under the general heading of *rank-dependence*. The idea behind rank-dependence is that the weight of a given state or of a given group's interests can depend on the relative position of that state or group: where its outcome ranks relative to other outcomes of the gamble or distribution. For example, what happens in the worst-case scenario might matter twice as much as what happens in the best-case scenario, or the interests of the worst-off group of people might matter twice as much as the interests of the best-off (even if the states are equiprobable and the groups are of the same size). As a result, it may be better to improve, by a given increment of utility, the outcome in a relatively worse state or the outcome of a relatively worse-off person—and it will sometimes be better to improve the outcome in a worse state or the outcome of a worse-off person than to improve the outcome in a better state or the outcome of a better-off person, even if we can improve the latter by a smaller utility increment. But it also may sometimes be better to improve the outcome in a better state or the outcome of a better-off person, if we can improve that state or person's outcome by much more. Or vice versa—what happens in the best-case scenario, or to the best-off group of people, might matter more than what happens in the worst-case scenario or to the worst-off group of people.

3. Rank-Dependent Decision Rules

There are many examples of rank-dependent decision theories in the literature,[2] but I will concentrate on risk-weighted expected utility,[3] because it employs both subjective probabilities and subjective decision weights (in the form of a "risk function"), and because unlike other rank-dependent theories it purports to characterize rational preferences.

To understand risk-weighted expected utility (REU) maximization, recall again the above graph. The way it is drawn encourages us to conceptualize a gamble or distribution as including four considerations—four possible utility values—each of which gets a weight equal to the probability of states or the proportion of people that realize it. But we can instead conceptualize it as including considerations about which states or people realize incremental benefits. In our example of working at

the start-up, the individual will at least get utility 1; in 80% of the states, he will do better than this by at least utility 3; in 30% of the states, he will do better than this by at least utility 2; and in 1% of the states, he will do better than this by utility 1.

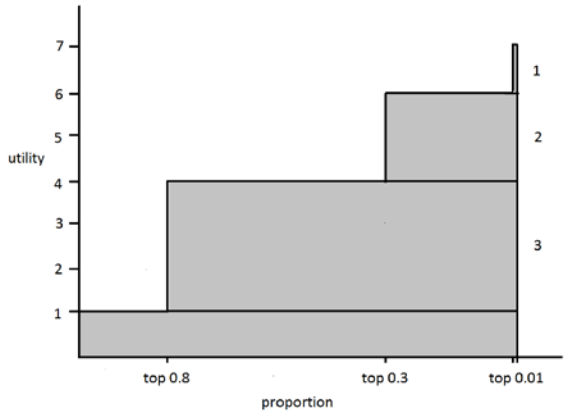

FIGURE 2: Expected Utility and Average Utilitarianism, Reconceptualized

In this reconceptualized graph, the height of each rectangle represents the difference between two adjacent utility levels (benefits that one might receive), and the width of each rectangle represents the probability of attaining *at least* the relevant utility level (the probability of receiving those benefits). As before, the area under the curve is the expected utility of the gamble. Thus, we can conceptualize EU-maximization as holding that the weight of each consideration of the form *I might obtain benefits of a certain size (in addition to whatever other benefits I obtain)* is the probability of obtaining those benefits.

REU-maximization says, on the contrary, that it is up to the individual how to weight each of these considerations. For example, in the above decision, that the individual will get at least utility 1 is guaranteed, so this consideration counts "all the way": it gets weight 1. That he will do better by at least utility 3 is a benefit realized in only the top 80% of the states, and he might care proportionately less about benefits that are only realized in some states, so he might weight this consideration only 0.64. That he will do better than this by at least utility 2 is a benefit only realized in the top 30% of states, so he might weight this consideration only 0.09. And that he will do better than this by utility 1 is a benefit only realized in the top 1% of states, so he might weight this consideration only 0.0001. (The REU of working at the start-up is therefore 3.1001, compared with an REU of staying at one's current job of 4, and one should prefer to stay at one's current job.[4]) Thus, the top states might get proportionately less and less weight in his decision-making.

Thus, the individual assigns a weight to the top *p*-portion of outcomes for each *p*, and this is his *risk function*, r(p). The risk function represents how much what happens in the top *p*-portion of outcomes matters to his practical decision making. (The aforementioned individual can be represented by $r(p) = p^2$.) Graphically, since the width of each rectangle represents the weight of attaining each utility level, the risk-function "shrinks" or "stretches" the horizontal rectangles:[5]

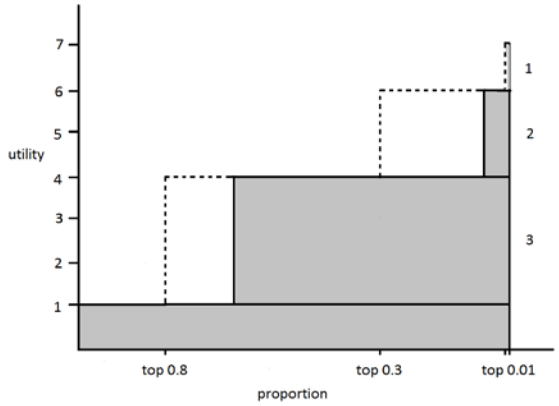

FIGURE 3: Risk-Weighted Expected Utility, Risk-Avoidant Individual

This individual is *risk-avoidant*. He is more concerned with what happens in worse states than better states, and thus holds that the value of a gamble is closer to its minimum value than the EU-maximizer holds. Risk-avoidant individuals have convex risk-functions: as benefits are realized in less likely states, these individuals care proportionately less about them. A limit case of risk-avoidance is maximin, in which benefits realized in only some states garner *no* weight.

Other individuals—*risk-inclined* individuals—might be more concerned with what happens in better states than worse states. These individuals have concave risk functions: as benefits are realized in less likely states, these individuals care proportionately more about them. (For example, $r(p) = p^{(1/2)}$ is a concave risk function, according to which the top 80% of outcomes garner 0.89 weight, the top 30% of outcomes garner 0.55 weight, and the top 1% of outcomes garner 0.1 weight. For this risk-function, REU(Start-up) = 4.88.)

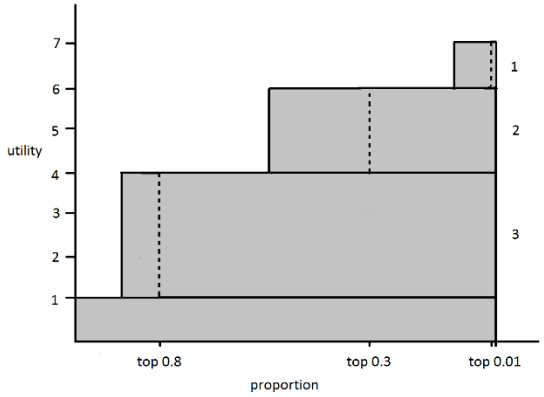

FIGURE 4: Risk-Weighted Expected Utility, Risk-Inclined Individual

Finally, some individuals—*globally-neutral* individuals—might be equally concerned with what happens in all (equiprobable) states, regardless of their relative rank, and thus will simply be expected utility maximizers. Globally neutral individuals have linear risk-functions: the top 80% of outcomes garner 0.8 weight, the top 30% of outcomes garner 0.3 weight, and the top 1% of outcomes garner 0.01 weight. These individuals are expected utility maximizers, with REU(Start-up) = 4.01.

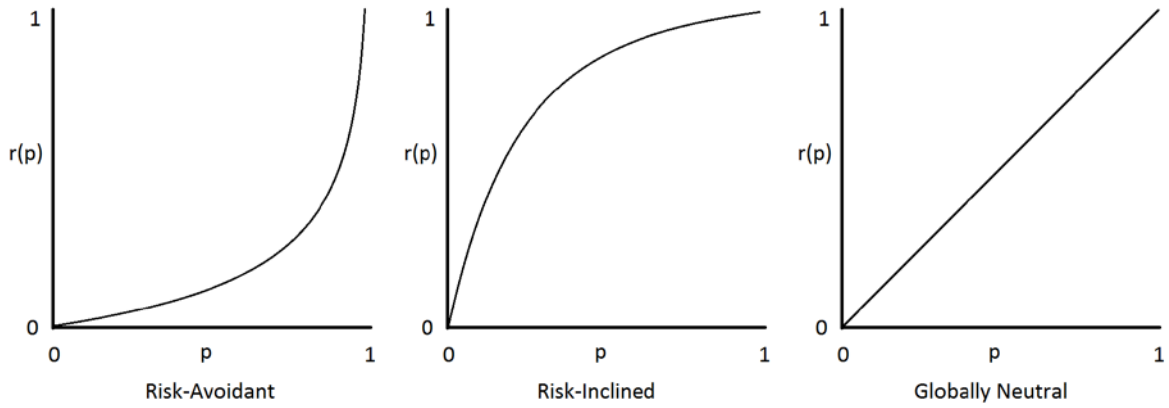

FIGURE 5: Weight of Top *p*-Portion of Outcomes

The idea behind REU-maximization is that there are actually three psychological components in preference-formation and decision-making: how much an individual values outcomes (utilities), how likely an individual thinks various states of the world are to obtain (probabilities), and the extent to which an individual is willing to trade off value in worse scenarios against value in better scenarios (the risk function). There are two different ways to think about the risk function: as a measure of distributive

justice among one's 'future possible selves'—how one trades off the interests of better-off possible selves against the interests of worse-off possible selves—and as a measure of how one trades off the virtue of prudence (making sure that the worst possibilities are not too bad) against the virtue of venturesomeness (making sure that the best possibilities are as good as possible).

The risk function does not measure how much an individual cares about some value, risk, that is not associated with any particular state. Instead, like EU-maximization, REU-maximization holds that all value is value in particular states. But contra EU-maximization, some individuals might be more concerned about what goes on in relatively worse or relatively better states.

Furthermore, REU-maximization is meant to be normative rather than descriptive: maximizing REU according to a non-linear risk function is *rational*. While space does not permit me to go through the details of the argument, one motivating thought is this. Merely determining how much an individual values outcomes and how likely he thinks various states of the world are to obtain is not enough to answer the question of how he should value a gamble that has some probability of realizing any one of a number of various outcomes—determining this is not enough to answer the question of how to aggregate the utility values of the possible outcomes to arrive at a single value for the gamble. Taking an average weighted by probabilities is just one way to aggregate, a way that corresponds to holding that the importance of what happens in the top states is just their proportion—in short, holding that the weight of a possible outcome in one's practical deliberation is just the probability of that outcome. But there are other ways to aggregate, and no reason to privilege a linear risk function over any one of a number of possible risk functions: there is to special reason to be globally neutral.

Of course, there will be some constraints on the risk function: it must set $r(0) = 0$ and $r(1) = 1$ (informally: a benefit that has no chance of being realized must get no weight, and a sure-thing benefit must get maximal weight). It must be non-decreasing—or perhaps positively increasing—in probability (informally: one must not prefer a worse chance of some benefit to a better chance of that benefit, or one must positively prefer the latter). And perhaps it must be continuous. But, holds REU theory, as long as an individual makes decisions according to a coherent utility and probability function and a risk function that has these characteristics, she will be rational.

There is more to be said about risk attitudes. Where a very wide range of risk attitudes are rational, the range of *reasonable* risk attitudes is, though still wide, slightly narrower. To understand this point, it helps to consider similar claims that are made about utility and probability, and to notice that there are

two uses of the term "rationality" which come apart: coherence and reasonableness. Coherence is what decision theory is traditionally concerned with. On the utility side, it allows preferring the destruction of the world to the scratching of your finger; and on the probability side, it allows assigning a high probability to the claim that a mad scientist is controlling the movements of your finger. As long as your utility and probability assignments are internally consistent, you count as rational in the "coherence" sense. Nonetheless, we tend to think there is something important to be said against these assignments: the person who prefers the destruction of the world fails to appropriately track value, and the person who thinks it likely that a mad scientist is controlling his finger fails to appreciate the force of the evidence. These two people are being unreasonable—though it is difficult to say with precision which values and beliefs are reasonable and which are not.

So too for risk attitudes. The person who always stays at home rather than drive because she considers the risk of an accident too great, and thereby who forgoes many good things, will count as coherent, and thus as decision-theoretically rational. But we tend to think that she places too much importance on the worst-case scenario, and fails to appreciate the importance of what happens in non-worst-case scenarios. We tend to think that she is being unreasonable. As with utilities and probabilities, we may not be able to draw a clear and precise line between reasonable risk attitudes and unreasonable ones, but we can clearly point to attitudes that fall on both sides. The key point is that while some risk attitudes are unreasonable despite being coherent, there is a wide range of reasonable risk attitudes.

4. Rank-Dependent Social Choice Rules

Rank-dependent social choice rules are directly analogous to rank-dependent decision theories: they allow us to weight the interests of individuals differently depending on their relative position in a particular distribution. The general term for these rules is *weighted-rank utilitarianism* (WRU).[6] We can understand weighted-rank utilitarianism by adapting the graphs and much of the discussion in the previous section to the social case. Recall our example distribution: {1, 0.2; 4, 0.5; 6, 0.29; 7, 0.01}. There are four considerations to evaluate when evaluating this distribution: that it offers a minimum of utility 1 to everyone; that the top 80% of our population gets at least an additional 3 utils; that the top 30% of our population gets an additional 2 utils; and that the top 1% of our population gets an additional 1 util (Figure 2).

Utiliarianism, maximin equity, and rank-weighted utilitarianism agree about the importance of the first consideration: it affects everyone, so its weight is 1. But these theories disagree about the weight of the

other three considerations. Utilitarianism holds that the value contributed by each of these considerations is the benefit multiplied by the proportion of the population who enjoy it. Maximin equity holds that the value contributed by each of these considerations is the benefit multiplied by zero (i.e. they contribute nothing). And rank-weighted utilitarianism holds that value contributed by each of these considerations is the benefit multiplied by an "importance function" of the proportion of the population who enjoy it. Just as the risk function measures the importance of what happens in the top *p*-portion of states to the value of a gamble, the importance function measures the importance of the interests of the top *p*-portion of individuals to the value of a social distribution. And just as a convex risk function corresponds to caring proportionately less and less about what happens in a smaller and smaller portion of the top states (risk-avoidance), a convex importance function (the *Gini family*[7]) corresponds to the claim that as benefits affect fewer and fewer people at the top, these benefits count for proportionately less and less in the evaluation of a distribution. So, for example, Figure 3 (interpreted for the social case) represents a weighted-rank utilitarian valuation with a convex importance function.

As should be clear from the above, and as others have noted,[8] utilitiarianism and maximin are each special cases of weighted-rank utilitarianism. Utilitiarianism holds that each individual gets equal weight, and so the weights of each group are proportionate to the size of the group—everyone's interests are counted the same. Maximin holds that all the weight accrues to the worst off, so that their interests are the only thing that matter. Utilitarianism places no special importance on the interests of the worst off, and maximin places no importance on the interests of anyone else. For utilitarianism, a consideration makes a differences in proportion to how many people it affects, and for maximin, a consideration only makes a difference if it affects everyone. The Gini family allows for a position between these two extremes: the interests of the relatively worse off count *more* than those of the relatively better off, but not exclusively. According to this distribution rule, we ought to help the worse off even if we can help them a bit less than the better off—but if we can help them much less, then we ought to help the better off instead.

One way to classify these three rules is in terms of the interpersonal tradeoffs we ought to accept: for example, if we remove some utility from the worst-off person and add some to the best-off person, what does the ratio of these two amounts need to be for the resulting distribution to be at least as good as the original? Utilitiarianism says that the amounts just need to be equal—we should be willing to accept 1:1 utility transfers from the worst-off individual to the best-off. Maximin says that the resulting distribution is always worse—we can only accept 1:∞ utility transfers from worst to best, which is to say we can't accept any such transfers. The Gini family allows for some ratio intermediate between these two. A

similar point holds in the decision theory case: when the worst and best state are equiprobable, an EU-maximizer is willing to accept 1:1 utility transfers from worst state to best state; a maximin-er is only willing to accept 1:∞ transfers from worst to best; and a risk-avoidant agent will accept something in between.

A note about how the Gini family fits into the philosophical literature on distributive ethics. Broadly speaking, philosophers identify two "aggregative" alternatives to utilitarianism: egalitarianism and prioritarianism. Egalitarianism is identified in two separate ways in the literature. It is typically identified as a philosophical view about the appropriate object of concern in evaluating a distribution: a view is egalitarian if it holds that inequality matters in itself.[9] This is in contrast to utilitarianism, which holds that the only objects of concern are what happens to individual people. Others identify egalitarianism as the purely formal view that the social welfare function is not strongly separable: roughly, the difference that some individuals' interests make to the value of a social distribution depends on what happens to other individuals.[10] Although not typically made explicit, a common thought is presumably that these two ideas go together: if the importance of what some individuals have depends on what others have, we must value something other than what happens to individuals, i.e., we must care about a "global" value, inequality.[11]

Prioritarianism was formulated by Derek Parfit because he wanted a theory to capture three criteria for evaluating distributions: (1) sometimes giving a smaller utility benefit to the worse off is better than giving a larger utility benefit to the better off, (2) *because* they are worse off, but (3) not because inequality is bad in itself—the only relevant concern is the well-being of each individual.[12] Utilitarianism falls afoul of the first criterion, and the "philosophical" formulation of egalitarianism falls afoul of the third. Prioritarianism meets all three criteria by holding that the interests of the worse-off matter more than those of the better-off, where "worse-off" and "better-off" are understood in an absolute sense; and prioritarianism is expressed formally by holding that the value of a distribution is its *total moral value*, where moral value is a concave function of utility. (For our purposes, we could instead formulate this as average moral value.)

Weighted-rank utilitarian theories are egalitarian in the *formal* sense: since an individual's rank affects the weight of her interests, and an individual's rank depends on the outcomes that other individuals receive, then the weight of each individual's interests depends on the outcomes that other individuals receive. (These theories reject strong separability.) However, weighted-rank utilitarianism needn't be egalitarian in the *philosophical* sense. In particular, just as REU-maximization holds that the only objects of concern

are outcomes in particular states, but that it is an open question how much each state counts towards the evaluation of a gamble, one could adhere to weighted-rank utilitarianism with the following motivation: the only objects of concern are the interests of each individual, but each individual's interests needn't be given the same weight in the evaluation of a distribution.[13] Indeed, the Gini family, when supported by this reasoning, meets all three of Parfit's desiderata—the formal difference between Gini and prioritarianism being that prioritarianism gives priority to the interests of those who are worse off in an absolute sense, and Gini gives priority to the interests of those who are worse off in a relative sense.

To disambiguate, I will call the view argued for in this paper—the view that accepts a member of the Gini family as its formal method for ranking distributions and accepts the philosophical claim that the interests of the relatively worse off matter more (but that inequality does not matter in itself) as its reason for doing so—*relative prioritarianism*. The key claim of relative prioritarianism as distinct from philosophical egalitarianism is about why the rank of each individual matters. It is not because we want to reduce inequality in itself, as if equality were some value over and above the well-being of each individual. Nor is it because individuals care about what other individuals have—they are not motivated by envy. Rather, it is because the claims of those who are relatively worse off take priority over the claims of those who are relatively better off. The key claim of relative prioritarianism as distinct from prioritarianism is that it is relative standing, rather than absolute standing, that determines priority.

5. Social Gambles

So far we have seen that there is an analogy between decision rules and social choice rules: the role played by states in decision theory is played by groups of individuals in social choice theory. But John Harsanyi and John Rawls make a more direct connection.[14] To use Rawls's terminology, individuals consider their preferences about institutional arrangements in the "original position," in which decisions are made behind a "veil of ignorance," where no one knows ahead of time their "place in society, their class position or social status, their place in the distribution of natural assets and abilities, their deeper aims and interests, or their particular psychological makeup."[15] Thus, individuals consider their preferences about gambles which correspond to social distributions and in which the possible 'states of the world' specify which place each of them will occupy in society—about *social gambles* such as:

> Social Gamble X = {I AM IN SOCIAL CLASS A, wealthy and fulfilled with short and pleasant workday; I AM IN SOCIAL CLASS B, wealthy and fairly fulfilled with fairly enjoyable workday; I AM IN SOCIAL CLASS C, enough money with long and not unpleasant hours; I AM IN SOCIAL CLASS D, scraping by with an unpleasant and demanding job.}

Harsanyi considers the situation in which social gambles also include risky prospects,[16] but for purposes of this paper I will only consider the case in which social gambles include sure-thing outcomes. Rawls considers social gambles not over outcomes themselves but over "primary goods" (rights, liberties, opportunities, power, income, wealth, and the bases of self-respect[17]), which help to determine what an individual can expect over his lifetime. While I will continue to use the term 'outcomes', this is meant to apply to whatever is to be distributed.

Preferences over social gambles are used to choose institutional arrangements (Rawls) or, more abstractly, to compare distributions and determine which is better (Harsanyi). For example, we can determine whether Policy X or Policy Y should be chosen, or which of their resulting distributions is better.

What preferences do rational individuals form about social gambles? Here is a crucial point where the assumptions that Harsanyi and Rawls make about the original position differ from each other. Harsanyi holds that individuals assign equal probability to being each individual in the society. He also holds that everyone knows the utility that each individual would derive from each outcome, so that outcomes (such as "wealthy-and-fulfilled-as-a-member-of-social-class-A") can be replaced by a single utility value for all individuals making the decisions—which is to say, a member of social class B assigns the same utility to \$0-as-a-member-of-A as a member of social class A assigns to \$0-as-a-member-of-A.[18] These two assumptions together imply that everyone is considering their preferences over gambles with agreed-upon utilities and probabilities. The effect is to transform social gambles into utility lotteries; for example, the above social gamble becomes the utility lottery {1, 0.2; 4, 0.5; 6, 0.29; 7, 0.01}. Harsanyi also assumes that rational individuals maximize expected utility, and the result is an argument for utilitarianism. If Social Gamble X has a higher expected utility than Social Gamble Y, then it follows both that Policy X should be chosen over Policy Y *and* that Policy X has a higher average utility than Policy Y.

Rawls, on the other hand, holds that individuals cannot assign probability to occupying each social role.[19] And as for assigning utility to outcomes realized by particular people, he holds only the weak assumption that we can know how outcomes are ordinally ranked—or at least that we can easily determine which group is worst off.[20] And since he holds that an individual facing a decision under uncertainty without subjective probabilities should employ maximin, the result is an argument for the maximin equity criterion.[21] If we assume that scraping-by-as-a-member-of-class-D is worse than any of the other outcomes, then maximin selects Social Gamble Y over Social Gamble X, and so Policy Y should be socially chosen over Policy X *and* the former is selected by maximin equity.

Notice that taking our social recommendations from preferences in the original position allows us to link together the intrapersonal tradeoffs each individual is willing to accept—the amount by which an individual is willing to make one state worse in order to make another state better—and the interpersonal tradeoffs we ought to accept in the case of social distributions. Harsanyi holds that individuals in the original position maximize expected utility, which is to say they are willing to accept 1:1 utility transfers from worst state to best state; correspondingly, our society should be willing to accept 1:1 utility transfers from the worst-off *person* to the best-off *person*. Rawls holds that individuals in the original position maximize the minimum value, which is to say they are not willing to accept *any* utility transfer from worst state to best state; correspondingly, our society should not accept utility transfers of any ratio from the worst-off person to the best-off person.

Harsanyi and Rawls differ on four main points. Harsanyi holds that individuals assign common cardinal utility values to each outcome realized by each person (the "interpersonal comparability assumption") whereas Rawls holds that they cannot. Harsanyi holds that individuals assign equal subjective probability to being each person (the "equiprobability assumption"), whereas Rawls holds that they are unable to assign probabilities at all. Harsanyi holds that individuals in the original position form preferences according to EU-maximization, whereas Rawls holds that they form preferences according to maximin. And each holds that his own resulting distributive rule—Harsanyi's utilitarianism and Rawls's maximin equity criterion—is superior to that derived by the other.

Whether the rule derived from the original position is utilitarianism or maximin equity depends on whether the setup makes expected utility maximization or maximin appropriate. And, of course, if the setup makes a different rule appropriate, then the result will be something else. As the reader no doubt anticipated, we will use the assumption that individuals in the original position maximize risk-weighted expected utility to argue for relative prioritarianism.

Given the obvious parallel between rank-dependent rules in social choice theory and those in decision theory, it is unsurprising that some work has already been done to connect the two. Axiomatizations of weighted-rank utilitarianism have been proposed that make use of analogous axiomatizations for rank-dependent decision rules.[22] However, individual preferences about social gambles do not figure into these results, and in particular none of them assume that the preferences of individual decision-makers are captured by a rank-dependent rule. Indeed, non-EU or non-utilitarian approaches that do make assumptions about individual preferences appear to come in two types: derivations of non-utilitarian rules

that assume individuals maximize expected utility, and derivations that relax the expected utility assumption and derive a rule other than weighted-rank utilitarianism (sometimes utilitarianism itself).[23]

The only author to consider what follows from the Harsanyi/Rawls approach if we assume that individuals use a rank-dependent rule is John Quiggin, who assumes that individuals maximize anticipated utility.[24] Anticipated utility maximization is formally equivalent to REU-maximization with an objective rather than subjective probability function, but it is interpreted differently: anticipated utility theory's "weighting function" (the equivalent of REU theory's risk function) is interpreted as a measure of optimistic or pessimistic beliefs—beliefs that are different from known probabilities. Quiggin notes that if in the original position we make the standard assumption that individuals overweight small probabilities of both good and bad outcomes, then we will arrive at policies which benefit both the very wealthy and the very poor. However, he does not think that this approach would be convincing, since "there is no reason why the weighting function which would be adopted in choosing between risky prospects should be the same one which would be used in social choice".[25] Since anticipated utility interprets decision weights as incorrect attributions of probability, it is unsurprising that we would have no reason to use these decision weights in the original position. Indeed, it is only when we have reason to think that rank-dependent utility maximization is normative, and to hold that decision weights are a necessary "third component" of instrumental rationality that we have reason to hold that decision makers in the original position maximize some form of rank-dependent utility.

Before we employ REU-maximization to argue for relative prioritarianism, we need three assumptions: two to make REU-maximization appropriate in the original position, and one to deal with the fact that a plurality of risk attitudes are rationality permissible, whereas we want to derive a single social choice rule. For purposes of this paper, I will simply accept without argument the first two elements of Harsanyi's setup: equiprobability and interpersonal comparability. The basic motivation for accepting equiprobability is that it represents the idea of giving "the same *a priori* weight to the interests of all members of the society."[26] Like Harsanyi, I take equiprobability to be an assumption necessary in order for choices in the original position to represent the idea that all people are treated equally, rather than an epistemic assumption. And while there are known worries about interpersonal comparability—particularly Rawls's worry that we cannot know from behind the veil each individual's conception of the good—I will simply assume that these can be overcome. (Even if they cannot, bracketing them will help to focus our attention on the ideal case in which we do have interpersonal comparability, so that we can see the structural features of distributive ethics.) The assumptions of equiprobability and interpersonal comparability together imply that individuals treat social gambles in the original position as utility

lotteries, and indeed that all individuals agree on which utility lotteries they are. The third key assumption will be the subject of Section 6.

A few caveats before we get to the heart of the argument. First, I am concerned here with the distributional rule to employ at the point at which distribution is the relevant question. Many utilitarians hold that the only moral question is a distributional question, and thus that utilitarianism is meant to apply at the most fundamental level. Rawls, on the other hand, adopts two rules—first ensure that everyone is to have adequate basic liberties and then employ maximin—and only the second is a distributional rule. We needn't take a stand on which of these two pictures, the one in which a distributional rule is the most basic kind of rule or the one in which a distributional rule is secondary to some other kind of rule, is correct. We are simply asking, when we get to the stage which requires a distributional rule, what should that rule be?[27] Second, I take no stand on whether that which is to be distributed is opportunities, happiness, primary goods, fulfilled desires, satisfied preferences, or something else—for purposes of this paper, outcomes are whatever entities we've decided the distributional question applies to. What we're interested in is the *structure* of distributive ethics, rather than the scope or content of distributive ethics. Even if this question is too abstract for real-world application, I take it that knowing what the criteria are for judging distributions—whether equality is a separate good, for example, or whether an increase in the well-being of the middle class contributes substantially to the overall good—will help determine where we should direct our attention in real-world cases.

Finally, although we will make use of points originating with Harsanyi and Rawls, we are adapting their insights to our purposes, rather than doing exegesis of either author. As should be clear, although their motivations and the traditions in which they are working are very different, we are arguing at a level of abstraction according to which they are addressing the same basic structural question. Indeed, although we will draw substantive conclusions about the structure of distributive ethics, one of the purposes of this paper is even more abstract: to provide a framework for relating risk and inequality.

6. Taking Risks for Others

So far, we've assumed that individuals who face social gambles in the original position assign common probabilities to occupying social positions and assign common utilities to the outcomes realized in each. We next assume that all of our citizens have preferences for social gambles that maximize risk-weighted expected utility. However, there is a plurality of acceptable risk attitudes; therefore, our actual citizens might have different preferences with respect to these gambles. For example, an actual member of social

class D might be risk-avoidant and thus prefer Social Gamble Y to Social Gamble X, whereas an actual member of social class A might be risk-inclined and thus prefer Social Gamble X.

How, then, should we think about preferences in the original position, given that there is a plurality of acceptable risk attitudes? Let us start by observing how we make decisions for other people whose risk-attitudes are unknown to us.

Imagine your acquaintance hurts his shoulder and is in moderate pain, and you do not know whether it is a muscle spasm or a pulled muscle. For simplicity, imagine these two possibilities are equally likely. Applying heat will help greatly if it is a muscle spasm, but will lead to intense pain if it is a pulled muscle; on the other hand, applying ice will do nothing for a muscle spasm and will provide mild relief for a pulled muscle:

Apply Heat = {muscle spasm, relief; pulled muscle, intense pain}

Apply Ice = {muscle spasm, moderate pain; pulled muscle, mild pain}

Applying heat is the risky but possibly rewarding course of action, and applying ice is the relatively safe course of action. It seems reasonable for an individual to prefer either choice for *himself*. However, whatever you would prefer for yourself, it seems you should choose ice for your acquaintance: without knowing someone's preferences, you can't subject him to a risk you're not sure he would take. But only to a point: if a pulled muscle is incredibly unlikely, then intuitively it seems like you should apply heat.

Thus, we seem to operate using:

> *Rule 1:* When making a decision for another individual, if I don't know which risks he is willing to take, err on the side of caution and choose the less risky option, within reason.

Importantly, you don't simply make the choice that is in line with your own risk-attitude. Nor do you pick haphazardly or arbitrarily. You would be criticizable if you picked the risky act, even if it turns out that this act is the one the acquaintance himself would have chosen and even if his injury turns out to be a muscle spasm. That you would be criticizable points to the fact that we treat making the less risky choice as normative.

Exactly how risk-avoidant do we think we ought to be in choices for others? I submit that the default risk-attitude we should adopt when making choices for others is the most risk-avoidant of the reasonable risk-attitudes. When we make a decision for another person, we consider what no one could fault us for, so to speak: if no reasonable person would reject an option on the grounds that it is too risky, then we are

justified in choosing that option. Conversely, if a reasonable person could reject it on these grounds, then we are not justified in choosing it.

These observations contrast with how we approach decision-making for another person when what is at issue is not someone else's risk-attitude but their basic desires. For example, if you are picking up ice cream for your acquaintance and you do not know whether he prefers chocolate or vanilla, you have no rule to guide you. There is nothing to do but to choose arbitrarily, or perhaps, lacking another way to make the choice, to choose in line with your own preferences. Notice further that finding out that a majority of people would prefer chocolate could sway me, even if I know a sizeable minority would prefer vanilla; but in the risk case, finding out a majority would take the risk could not sway me, if I knew a sizeable minority would not take the risk. Different reasonable utility assignments are on a par in a way that different reasonable risk assignments are not: we default to risk-avoidance, but there is nothing to single out any utility values as default.

A final observation about taking risks for another person. If we know the person's risk-attitude, we tend to defer to it:

> *Rule 2:* When making a decision for another individual, if I know which risks he is willing to take, choose for him as he would choose for himself.

I leave it open whether Rule 2 requires that we know a person's risk-attitude in general or in the relevant domain or in the choice at hand. I also leave it open whether we defer when the person's risk-attitude is unreasonable. Finally, I leave it open whether the factor that allows me to choose the risky option for someone else is my *knowledge* of his preferences or his *consent* to the choice, though for ease of exposition I will speak as if knowledge of his preferences is the relevant factor.

Putting these observations together, we have the following general normative principle:

> *Risk Principle*: When making a decision for an individual, choose under the assumption that he has the most risk-avoidant attitude within reason unless we know that he has a different risk-attitude, in which case, choose using his risk-attitude.

According to the Risk Principle, the default distribution among states of a person is one in which worse states are given significantly more weight than better states. In order to transfer utility between states, relative to the default distribution, the person who would be made worse off in some state by the transfer must desire this.

If, following the Risk Principle, we make the assumption that all individuals in the original position have the default risk-attitude, then they can reach unanimous agreement: everyone will choose social gambles that prioritize what happens in worse states, and we will arrive at a distribution which prioritizes the interests of individuals who are worse off. Specifically, the weight that we give to the interests of each rank-ordered group will be equivalent to the weight we give to rank-ordered states of the same proportion, according to the most risk-avoidant reasonable risk-attitude.[28] If Social Gamble Y has a higher risk-weighted expected utility than Social Gamble X relative to the default risk-attitude, then it follows both that Policy Y should be chosen over Policy X *and* that the former has a higher weighted-rank utility when the importance function corresponds to the default risk function. Call this importance attitude the *designated* importance attitude.

Thus, we will arrive at relative prioritarianism, which says to weight the interests of those relatively worse off more than the interests of those relatively better off but to give everyone's interests some weight—and to weight these interests according to the designated importance attitude. When the veil of ignorance is lifted and individuals become actual people with particular characteristics, they may instead adopt any risk attitude they wish. They may choose to take gambles which are riskier than the default risk attitude would recommend—they may move utility between states of themselves—precisely because it is they themselves who will be made worse off if worse states obtain. But we may not choose social policies that privilege the better-off more than the designated importance attitude recommends.

7. Justifying an Ethic of Distribution

We are now most of the way to an argument for relative prioritarianism with the designated importance attitude: rational decision makers—risk-weighted expected utility maximizers—with the default risk-attitude behind the veil of ignorance will prefer social gambles that maximize weighted-rank utility with this importance attitude. But we still need to say *why* preferences in the original position dictate what our distributive ethical principle should be, and in particular what role the Risk Principle plays in the argument. There are two different routes to take here, and we can differentiate them by how they justify, to each citizen, the choice of one distribution rather than another.

The first line holds that the conditions imposed on individuals in the original position are the conditions under which individual preferences reflect moral judgments about distribution. (This is how Harsanyi makes use of the original position, and why he calls preferences in the original position *moral preferences*.) Thus, for example, we assume that individuals form preferences under the condition of

anonymity—each does not know ahead of time which social position he will occupy—because this condition is necessary to ensure that each individual's preferences are not unduly sensitive to her own interests. We assume that individuals assign equal probability to their being each person because this condition is necessary to ensure that each person's interests count exactly as much as each other person's interests.

In addition to these structural assumptions, there must be assumptions about the content of preferences if they are to reflect moral judgments. One such assumption is *mutual disinterest*.[29] For example, assume that actual Alice is happiest when Bob is happy, and so her utility is higher when things go well for Bob. If we allow these to be Alice's utilities behind the veil of ignorance—the utilities assigned in the Alice position—then we will give undue weight to Bob's interests. The assumption of mutual disinterest is thus needed even if it is in fact false of most people. It is not that we think that most individuals are mutually disinterested. And it is certainly not that we think individuals *ought* to be mutually disinterested. Rather, it is inappropriate to take actual concern for others into account when we are asking about the correct "first-order" distribution: fairness requires taking into account the interests each individual would have in the absence of other-regarding preferences. Once we have determined what claims an individual has to various resources, each individual is free to choose to give up his claims for the sake of others—but we must first determine what would be fair in the absence of these choices.

Following this line, we add that assigning individuals the default risk-attitude—abstracting away from their actual risk-attitudes—is necessary to ensure that individual preferences reflect moral judgments. There are two ways to take this line. The first route is to hold that the assumption of the default risk attitude is an assumption like anonymity or equiprobability. It reflects a structural requirement of fairness: the way we balance multiple individuals' competing interests must be the same way we balance a single individual's competing interests in the absence of special knowledge about her. Here the important fact about the default risk attitude is that it is the attitude we ought to adopt in decisions for strangers. The second route is to hold that the assumption of the default risk-attitude is an assumption like mutual disinterest. We are not assuming that most people have the default risk attitude, or that individuals ought to be this risk-avoidant. We are instead assuming that "first-order" distributive ethics is a matter of what we would prefer if we all had the default risk-attitude; once we have settled this first-order question, each individual is free to adopt whatever risk-attitude he wants for his own choices. Here the important fact about the default risk attitude is that it is default: it is the attitude which an individual must specially choose to move away from.

This line builds a normative assumption into the original position, an assumption about which risk attitude is the one from which we derive an ethics of distribution. But that we start with a normative assumption does not mean that we have made little progress in arguing for a normative theory. The conclusion that we arrive at—that distribution should be relative prioritarian with a particular importance attitude—is substantially different from the assumption. What we have shown is that if we start with a normative assumption about how we must act on behalf of an individual absent knowledge of what he prefers, we can derive a normative conclusion about how we ought to rank social policies or distributions.

According to this line, the way we justify our policy or distribution to each citizen is this: *if you didn't know the things that cloud your moral judgment—namely, which individual you are, what characteristics you have, and what your actual risk-attitude happens to be—then you yourself would have chosen this policy or distribution.*

The second line for holding that preferences in the original position dictate our distributive ethical principle posits a more indirect connection between preferences in the original position and distributive ethics. This route starts with Rawls's point that *reflective equilibrium* is called for: we ought to be willing to modify both our description of the original position and the resulting principles until we find a reasonable description of the original position that yields principles which match our considered judgments of distributive justice.[30] More generally, the whole edifice is justified holistically: each piece is partially supported by its cohering with each other piece. Adapting this point to the justification of principles about our judgments rather than the judgments themselves, we look at the reasons to adopt the Risk Principle, see what they imply about the principles of distributive ethics if the Risk Principle is adopted in the original positon, and see whether these principles can be given a considered foundation along the lines of our reasons to adopt the Risk Principle.

Let us examine, then, what might justify the Risk Principle. Why should we default to a fairly risk-avoidant attitude? To answer this question, we must explain both why we don't default to a less risk-avoidant attitude (e.g., global neutrality) and why we don't default to an even more risk-avoidant attitude (e.g. maximin). The explanation for both lies in the fact that if an individual ends up in a state where a different choice would have been better, then the actual choice (made by her or someone else) requires justification. For example, ending up with intense pain because heat was applied to a pulled muscle requires justification for why ice wasn't instead applied; and ending up with moderate pain because ice was applied to a muscle spasm requires justification for why heat wasn't instead applied.

Thus, justification is required when there is a discrepancy between how things turned out and how they could have turned out if a different choice were made. The required justification will be supplied by some advantage of the current choice in a different state. ("Why didn't you apply heat instead—given that I have a muscle spasm, I would have relief instead of moderate pain?" "Because it might have been a pulled muscle, and then heat would have caused intense pain!") And the size of the advantage in that other state, as compared with the size of the disadvantage in the current state, will be relevant. And so too the size of advantages and disadvantages in other possible states. But—and this is the crucial point—ending up in a relatively worse state requires stronger justification than ending up in a relatively better state. ("Why didn't you apply ice instead—given that I have a pulled muscle, I would have mild pain instead of intense pain?" "Because it might have been a muscle spasm, and then the ice would have caused relief" seems like an irresponsible answer.) Thus, we don't default to global neutrality. However, a discrepancy between how things turned out and how they could have turned out always requires some justification, even if one ends up in a relatively good state—if an alternative would have been better, we still need to say why this alternative wasn't chosen. Thus, we don't default to maximin.

In the absence of any additional considerations, we default to being more concerned with what happens in worse states because of the relative amount of justification needed should these states obtain. This is why, for example, we ought to choose ice over heat when we don't know anything else about the individual's preferences. However, if it is worth it to the individual to risk intense pain in exchange for the possibility of relief, then we have a new justification that supplants the ordinary justification in every state: that it was worth it to the individual to trade off between states in the relevant way. ("Why didn't you apply ice instead?" "Because the risk associated with applying heat was worth it to you!")

What is it that makes it worth it to the individual himself to accept a gamble that is riskier than the default? It is that the amount of (prudential) justification he needs in worse states is less than that needed by someone with the default risk-attitude. This presents another reason to identify the default risk-attitude with the most risk-avoidant attitude within reason: although some people might need less justification in worse states than the default requires in these states, no one could reasonably need *more*.

To bring this point to the social case: a particular social distribution (or a policy that results in a particular social distribution) must be justified to each citizen, and each citizen has a potential complaint if she would have done better according to an alternative arrangement. Thus, we have to answer each citizen's complaint. Or each citizen has a claim to potential benefits that might be realized by various arrangements, and we have to adjudicate these competing claims.[31] Notice that what needs to be justified

here is not inequalities themselves—the difference between what I have and what you have—but rather what we might call disappointments—the difference between what I have and what I could have had under some alternative arrangement. And *all* disappointments require justification, but disappointments experienced by those who are worse off require more justification.

We can compare the story here to that given by Rawls. Rawls notes that a social distribution must be justifiable to all of the citizens, and he holds that maximin meets this criterion but utilitarianism does not. He points out that the most disadvantaged will have a hard time accepting "deep and pervasive" inequalities. Utilitarianism asks them to accept their situation with reference to the fact that their having less allows the well-off to have still more advantages than they otherwise could have. For example, when we choose a policy that gives the worse-off 10 utils less so that the better-off can have 20 utils more, the worse-off are expected to accept this on the basis of the fact that they are giving up something smaller so that others can have something greater. But, says Rawls, this is an extreme psychological demand. Maximin does not impose this burden on the least advantaged, because the well-off are only allowed to have further advantages if these benefit the worse-off as well. Rawls notes that maximin faces a symmetrical problem: the better-off must accept less than they would receive under utilitarianism. But, Rawls argues, two things make this easier to accept: they are "more fortunate and enjoy the benefits of this fact; and insofar as they value their situation relatively in comparison with others, they give up that much less."[32]

The view here agrees with Rawls's point that we have a particular duty to justify a distribution to the most burdened, but adds that this is not our only justificatory duty. While it will be *easier* for the better-off to accept their situation than it will be for the worse-off to accept theirs, this is a matter of degree rather than kind. For example, the better-off can easily accept a policy that gives them 10 utils less so that the worse-off may have 20 utils more, and can somewhat easily accept a policy that gives them 20 utils less so that the worse-off may have 10 utils more—but they may not be able to easily accept a policy that gives them 2,000 utils less so that the worse-off may have 10 utils more. This is particularly true when the worse-off and better-off under consideration are not the two extreme groups (the worst-off and best-off), but are instead two closely ranked groups: the worst-off and the second-worst-off, for example, or the second-best-off and the best-off.

When we choose policies that give the better-off less so that the worse-off can have more, these policies are not automatically accepted by the better-off, and when we choose policies that give the worse-off less so that the better-off can have more, these policies are not automatically rejected by the worse-off.

Instead, the requirement that less advantaged individuals be given special consideration should be this: policies that give the worse-off less so that the better-off can have more must be much more advantageous to the better-off than they are disadvantageous to the worse-off, whereas policies that give the better-off less so that the worse-off can have more needn't be much more advantageous to the worse-off—and can indeed be much less advantageous—than they are disadvantageous to the better-off. And, furthermore, the needed size of these relative advantages depends on the rank of each group within society as a whole.

According to this line, the way we justify our policy or distribution to each citizen is this: *distributive ethics is a matter of adjudicating the competing claims of individuals, and this is the policy or distribution that takes your interests into account as much as is fair, keeping in mind that the less advantaged are owed special consideration.*

If one of these lines is successful, then we have an argument for relative prioritarianism with the designated importance attitude. To sum up the argument, we've started with five premises:

(1) Individuals in the original position assign common utilities to outcomes they would experience as other individuals. [assumed without argument]
(2) Individuals in the original position assign equal subjective probability to being each individual. [assumed without argument]
(3) Individuals in the original position maximize risk-weighted expected utility. [section 3]
(4) We should ascribe to individuals in the original position the most risk-avoidant reasonable risk-attitude. [section 6]
(5) Preferences in the original position dictate what our distributive ethical principle should be. [section 7]

And the result is an argument for the following conclusion:

> We ought to choose policies that maximize weighted-rank utility with respect to the importance attitude $I(p) = r'(p)$, where $r'(p)$ is the most risk-avoidant attitude within reason.

In other words, we ought to give more weight to the interests of the relatively worse off than to those of the relatively better off; specifically, we ought to give them as much weight as we default to giving to relatively worse states in individual decision-making.

It might be that the reader is unconvinced by either of the lines taken in this section. In this case, what one can take from this paper is a template for deriving rules of distributive ethics. If one can supply an alternative justification for the Risk Principle, then this will lead to an alternative story upon which we can base relative prioritarianism with the designated importance attitude. Alternatively, perhaps a milder

form of relative prioritarianism is attractive. If one instead argues for a version of the Risk Principle according to which the default risk-attitude we use in choices for others is risk-avoidant but less so, then this will lead to a less inequality-avoidant relative prioritarianism. Or one might argue for other rules for assigning risk-attitudes in the original position. Two obvious possibilities have already been discussed: if individuals in the original position are globally neutral, then we have an argument for utilitarianism, and if individuals in the original position are maximally risk-avoidant, then we have an argument for maximin. But there are a number of additional possibilities. For example, one might argue that individuals in the original position must have the risk-attitude *typical* of those found in our actual society, or the *median* risk-attitude of members of our actual society. Interestingly enough, both of these possibilities will result in giving more weight to the interests of the very worst off, but also giving more weight to the interests of the very best off.[33] Finally, it might be that different "risk rules" are appropriate for distributive ethical questions in different circumstances, or that some policies can be accepted via overlapping consensus among different risk-attitudes. In any case, what we have shown is that there is a natural argument from assumptions about risk-attitudes to conclusions about distributive ethics, and we have provided a way to connect the two.

8. The Separateness of Persons and the Shift

I close by briefly suggesting that the argument in this paper can escape two worries that typically plague theories linking distributive ethics to individual preferences among social gambles. The first is that such arguments do not take seriously the distinction between persons.[34] When I trade off utility between different possible states of an individual, the individual losing utility is the same as the individual gaining utility, and this is what makes the tradeoff justified to the individual losing utility. However, this is clearly not the case when trading off utility between different individuals: distributive ethics is not rational prudence. Thus, any theory that derives the social case from the individual case by treating social gambles as individual gambles appears not to respect this distinction.

On the view here, we can state clearly the sense in which tradeoffs between states of a person and tradeoffs between different individuals are not the same. While an individual is free to venture away from the default—he is free to move utility between states—by giving either more or less weight to particular ranks than the default risk-attitude would, we are not free to move utility between persons. (We are also not free to move utility between two states of a person without his choosing this.) On the view here, the separateness of persons is summed up in the fact that there is a plurality of acceptable risk-attitudes, but a single correct importance attitude. Furthermore, the view here does not treat distributive ethics directly as

a matter of rational prudence: the social choice is not the choice that every man would make for himself if he were equally likely to be each person, or if he were to experience all lives. It is instead the choice he would make for himself if constrained to have the default risk-attitude, the attitude that we must have when making choices for others. Distributive ethics, we might say, is rational prudence on behalf of a stranger whose preferences you don't know.

The second worry concerns a phenomenon that Michael Otsuka and Alex Voorhoeve call "the shift."[35] Data suggest that we do not treat interpersonal decision-making and intrapersonal decision-making analogously. Specifically, we are unwilling to make the same tradeoffs between the worst-off person and the best-off person in the interpersonal case as we are willing to make between the worst-case scenario and the best-case scenario in our own intrapersonal case. For example, Eric Nord found that experimental subjects evaluating distributions of health outcomes counted the interests of the worst-off roughly five times as much as the interests of the best-off—but most subjects were not as correspondingly risk-avoidant.[36] Thus, there is a shift between the attitude we take towards distribution among people and the attitude we take towards distribution among states of ourselves: a typical person is willing to accept more personal risk than she holds societal inequality to be justified.

According to the view here, we can explain the shift as follows. Individual decisions and social decisions are analogous, in that a rank-dependent view captures them both. However, the weight we give to the interests of the worst-off people—the importance attitude we adopt for social decisions—is the highest reasonable weight one can give to the worst state in individual decision-making, whereas the weight each individual gives to her own worst state is up to her. And, almost by definition, most reasonable individuals will give less weight to the worst state than the highest reasonable weight. Thus, most individuals will be less risk-avoidant than the default, which is to say that they will accept more personal risk than we accept inequality. (Again, although I hold that the default risk-attitude is the most risk-avoidant within reason, one only need to hold that most individuals are not nearly as risk-avoidant as the default to explain the shift.)

Notice that the cited data also lend support to the Risk Principle itself, via the method of reflective equilibrium. For we have now shown that from the Risk Principle in the original position we can derive some of our basic intuitions about particular cases, where alternatives such as expected utility maximization in the original position do not accord with these intuitions.

There is more to be said about the relationship between individual decision-making and social choice. But what I have done, in giving principled reasons for a plurality of acceptable risk attitudes but a single importance attitude, is to open up a conceptual possibility for thinking about the relationship between individuals' actual risk-attitudes and the importance we ascribe to the interests of the worse-off relative to those of the better-off. Individuals' actual choices don't display levels of risk-avoidance analogous to intuitive levels of inequality-avoidance in social choices, but nonetheless a certain type of individual choice—choice assuming the default risk attitude—does.

9. Conclusion

I have provided an argument for relative prioritarianism, starting from preferences in the original position. If we are willing to grant the equiprobability and interpersonal comparability assumptions; and if we accept risk-weighted expected utility maximization as characterizing the preferences of rational agents; and if we hold that the default risk-attitude is fairly risk-averse; and if we hold that distributive ethics can be derived from preferences in the original position, then we can conclude that we ought to give more weight to the interests of the relatively worse off than those of the relatively better off.

ACKNOWLEDGEMENTS

This paper was greatly improved by discussions with Kenny Easwaran, Alan Hájek, Niko Kolodny, and Seth Lazar; as well as with the audiences at the Australian National University, the Arizona Ranch Metaphysics Conference, the Princeton University Center for Human Values, Stanford University, and the University of Pennsylvania.

APPENDIX: Individual Decision Rules and Social Choice Rules

In the individual case, the basic unit of evaluation is an *ordered gamble*. Let $\leq$ represent some individual's preference relation. Then let $g = \{E_1, x_1; \ldots; E_n, x_n\}$ be an ordered gamble that yields outcome x_i in event E_i, where $x_1 \leq \cdots \leq x_n$. Let $p(E_i)$ be the subjective probability of E_i. (Notice that there may be multiple ordered representations of the same unordered gamble, due to ties. All ordered representations of a given unordered gamble will yield the same values for all of the rules below.)

Let $u_k(x_i)$ be the utility of x_i for individual k. Where we have just one individual, we will omit the personal subscript and write $u(x_i)$.

For the social case, we assume that individual utilities are given (or already derived from individual preferences) and are interpersonally compatible. The basic unit of evaluation is an *ordered distribution* to groups of individuals, where groups are individuated such that every individual in a group shares the same utility function and receives the same outcome on the distribution—that this is always possible for a finite population is shown by the fact that we can simply let each group contain exactly one individual.

Let $d = \{P_1, x_1; \ldots; P_n, x_n\}$ be an ordered social distribution in which each individual in group P_i has utility function u_i and receives outcome x_i, and where $u_1(x_1) \leq \cdots \leq u_n(x_n)$. Let $p(P_i)$ be the proportion of the population that P_i represents. (Again, notice that there may be multiple ordered representations of the same unordered social distribution—again, all ordered representations of a given unordered distribution will yield the same values for all of the rules below.)

I. Expected Utility and Average Utilitarianism

The expected utility of *g* is:

$$EU(g) = \sum_{i=1}^{n} p(E_i)\, u(x_i)$$

Alternatively,

$$EU(g) = \sum_{i=1}^{n} \left[\left(\sum_{j=i}^{n} p(E_j) \right) (u(x_i) - u(x_{i-1})) \right]$$

The average utility of *d* is:

$$U(d) = \sum_{i=1}^{n} p(P_i)\, u_i(x_i)$$

Alternatively, $U(g) = \sum_{i=1}^{n} \left[\left(\sum_{j=i}^{n} p(P_j) \right) (u(x_i) - u(x_{i-1})) \right]$

II. Maximin and Maximin Equity

The utility of *g* according to maximin is:

$$M(g) = u(x_1)$$

The utility of *d* according to maximin equity is:

$$M(d) = u_1(x_1)$$

III. Risk-Weighted Expected Utility and Weighted-Rank Utilitarianism

The risk-weighted expected utility of g is:

$$REU(g) = \sum_{i=1}^{n}\left[r\left(\sum_{j=i}^{n} p(E_j)\right)(u(x_i) - u(x_{i-1}))\right]$$

where r is "risk function" from [0, 1] to [0, 1], with r(0) = 0, r(1) = 1, and r non-decreasing.

(Quiggin's anticipated utility is obtained by replacing p(E) with a given, objective probability *p*, and Gilboa's Choquet expected utility is obtained by replacing r(p(E)) with a weight w(E).)

To see the standard formulation of weighted-rank total utilitarianism, assume that each group P_i includes only one individual. Then the weighted-rank total utility of a distribution is:[37]

$$W(d) = \sum_{k=1}^{n} \lambda_k\big(u_k(x_k)\big)$$

where λ_k corresponds to the weight that the k^{th}-worst individual gets in the evaluation of the distribution.

Using an "average" rather than a "total" formulation:

$$W(d) = \sum_{k=1}^{n} \frac{\lambda_k}{\sum_{i=1}^{n}\lambda_i}\big(u_k(x_k)\big)$$

The constraints that $\lambda_1 > 0$ and (for k < n) $\lambda_k \geq \lambda_{k+1} \geq 0$ correspond to the *generalized Gini family*.[38]

In this paper, I've assumed that groups can be any size and I've stated weighted-rank utilitarianism as:

$$W(d) = \sum_{k=1}^{n}\left[I\left(\sum_{j=k}^{n} p(P_k)\right)(u_k - u_{k-1})\right]$$

where I is an "importance function" from [0, 1] to [0, 1], with I(0) = 0, I(1) = 1, and I non-decreasing.

These formulations are equivalent when we set we set $I\left(\frac{n-(k-1)}{n}\right) = \frac{\sum_{i=k}^{n}\lambda_i}{\sum_{i=1}^{n}\lambda_i}$.
The generalized Gini family is then given by the constraint that *I* is weakly convex, and strictly convex if the above inequalities are strict.

Utilitarianism is given by I(p) = p.
Maximin is given by I = 0 everywhere except I(1) = 1.

[1] What is at issue is which aggregation method is correct. Two aggregation methods here—EU-maximization and REU-maximization—each have a "representation theorem" that allows us to determine utilities from preferences. The remaining method—maximin—only relies on an ordering of outcomes, so doesn't rely on cardinally meaningful utility assignments.

[2] Examples include anticipated utility (John Quiggin, "A Theory of Anticipated Utility," *Journal of Economic Behavior and Organization* 3 (1982): 323-343), dual theory (Menahem E. Yaari, "The Dual Theory of Choice under Risk," *Econometrica* 55.1 (1987): 95-115), Choquet expected utility (David Schmeidler, "Subjective Probability and Expected Utility without Additivity," *Econometrica* 57.3 (1989): 571-587; Itzhak Gilboa, "Expected Utility with Purely Subjective Non-Additive Probabilities," *Journal of Mathematical Economics* 16 (1987): 65-88), and cumulative prospect theory (Daniel Kahneman and Amos Tversky, "Prospect Theory: An Analysis of Decision under Risk," *Econometric*a 47 (1979): 263-291; Amos Tversky and Daniel Kahneman, "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty* 5 (1992): 297-323).

[3] Lara Buchak, *Risk and Rationality* (Oxford: Oxford University Press, 2013).

[4] REU(Start-up) = (1)(1) + (0.64)(3) + (0.09)(2) + (0.0001)(1) = 3.1001. REU(Current job) = (1)(4).

[5] Graphical representation of REU-maximization roughly follows Richard Pettigrew, "Risk, Rationality and Expected Utility Theory," *Canadian Journal of Philosophy* 45.5-6 (2015): 798-826.

[6] Typically, these rules are stated as follows: rank the individuals from worst-off to best-off and multiply the utility of each individual by the weight of her rank; then sum these values. See Claude d'Aspremont and Louis Gevers (2002), "Social Welfare Functionals and Interpersonal Comparability," in *Handbook of Social Choice and Welfare* vol. 1, eds. K.J. Arrow, A.K. Sen, K Suzumura (Elsevier), 471. If the weights are decreasing, then the interests of worse-off individuals are weighted more heavily than the interests of better-off individuals. The Gini family is the general name for weighted-rank utilitarian rules with decreasing weights. The generalized Gini family was first introduced by John Weymark, as a generalization of the ranking induced by the Gini measure of inequality (John A. Weymark, "Generalized Gini Inequality Indices," *Mathematical Social Sciences* 1 (1981): 409-430); further work is due to David Donaldson and John Weymark, who consider variable populations (David Donaldson and John A. Weymark, "A Single-Parameter Generalization of the Gini Indices of Inequality," *Journal of Economic Theory* 22.1 (1980): 67-86), and John Weymark, who considers distributions over sets of opportunities rather than over outcomes

(John A. Weymark, "Generalized Gini Indices of Equality of Opportunity," *Journal of Economic Inequality* 1.1 (2003): 5-24). We will state these rules slightly differently, as an average rather than a total sum, and following the conceptualization represented by Figure 2. The technical formulation is found in the Appendix, along with a simple proof that the two statements of the rules are equivalent.

[7] See previous footnote.

[8] Weymark, "Generalized Gini Inequality Indices"; Udo Ebert, "Measurement of Inequality: An Attempt at Unification and Generalization," *Social Choice and Welfare* 5 (1988):147-169; Udo Ebert, "Rawls and Bentham Reconciled," *Theory and Decision* 24 (1988): 215-233; Han Bleichrodt, Enrico Diecidue, and John Quiggin, "Equity weights in the allocation of health care: the rank-dependent QALY model," *Journal of Health Economics* 23 (2004): 157-171.

[9] See Derek Parfit, "Equality or Priority?," in *The Ideal of Equality*, eds. Matthew Clayton and Andrew Williams (MacMillan Press Ltd, and St. Martin's Press, Inc., 2000), 81-125. Originally delivered as the Lindley Lecture at the University of Kansas (1991). See also Larry Temkin, "Equality, Priority, and the Levelling Down Objection," in *The Ideal of Equality*, eds. Matthew Clayton and Andrew Williams (MacMillan Press Ltd, and St. Martin's Press, Inc., 2000), 126-161; Marc Fleurbaey, "Equality Versus Priority: How Relevant is the Distinction?" *Economics and Philosophy* 31 (2015): 203–217.

[10] See John Broome, *Weighing Goods: Equality, Uncertainty, and Time* (Oxford: Oxford University Press, 1991); David McCarthy, "Utilitarianism and Prioritarianism II," *Economics and Philosophy* 24.1 (2008): 1-33. Karsten Klint Jensen holds that egalitarianism should be identified with the formal view that the social welfare function is not additively separable (Karsten Klint Jensen, "What is the Difference Between (Moderate) Egalitarianism and Prioritarianism?" *Economics and Philosophy* 19 (2003): 89-109). Note that additive separability implies strong separability, and the converse holds under certain conditions (see Broome, *Weighing Goods*, 82-86).

[11] But see Jensen, "What is the Difference?" and Iwao Hirose, "Reconsidering the Value of Equality," *Australasian Journal of Philosophy* 87.2 (2009): 301-312.

[12] Parfit, "Equality or Priority?"

[13] Note that strong separability in persons is analogous to the sure-thing principle for gambles (see Broome, *Weighing Goods*, 94-95); thus, WRU's rejection of strong separability is the analogue of REU's rejection of the sure-thing principle. I argue elsewhere that rejection of the sure-thing principle needn't imply a concern for risk 'in

itself' (Buchak, *Risk and Rationality*, Chapter 5). Analogous considerations can be put forward to argue that rejection of strong separability needn't imply a concern for inequality 'in itself'.

[14] John C. Harsanyi, "Cardinal Utility in Welfare Economics and in the Theory of Risk-taking," *Journal of Political Economy* 61.5 (1953): 434-435; John C. Harsanyi, "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility," *Journal of Political Economy* 63.4 (1955): 309-321; John C. Harsanyi, "Can the Maximin Principle Serve as a Basis for Morality?: A Critique of John Rawls's Theory," American Political Science Review 69.2 (1975): 594-606; John C. Harsanyi, "Morality and social welfare," *Rational Behavior and Bargaining Equilibrium in Games and Social Situations* (Cambridge: Cambridge University Press, 1977): 48-83; John C. Harsanyi, "Morality and the Theory of Rational Behavior," *Social Research* 44.4 (1977): 623-56; John Rawls, "Some Reasons for the Maximin Criteria," *The American Economic Review* 64.2 (1974): 141-146; John Rawls, *A Theory of Justice* (Cambridge: Belknap, 1999). Original edition 1971; John Rawls, *Justice as Fairness: A Restatement*, ed. Erin Kelly, Belknap Press of Harvard University Press (Cambridge: Belknap, 2001).

[15] Rawls, "Some Reasons for the Maximin Criteria," 141.

[16] Harsanyi, "Cardinal Welfare."

[17] Rawls, *Justice as Fairness*, 57-61.

[18] For simplicity, I am assuming that everyone in a given social class receives the same utility from each outcome, but this assumption is unnecessary, since we can always divide a given class into smaller classes. Harsanyi himself considers the finest-grained division, i.e., n classes for the n distinct individuals in society.

[19] Rawls, *Theory of Justice*, 134-35.

[20] Rawls, *Theory of Justice*, 123, 281-85; Rawls, "Some Reasons for the Maximin Criteria," 143.

[21] Rawls, *Theory of Justice*, 132-39; Rawls, *Justice as Fairness*, 97-99. Rawls prefers the term "the difference principle" for the resulting criterion. Furthermore, he refines the principle to leximin (*Theory of Justice*, 72), which only differs from maximin in particular situations that are not relevant here. An inability to assign probabilities is the first of three conditions under which he holds that individuals would use maximin in the original position (*Theory of Justice*, 132-39; *Justice as Fairness*, 97-104, 115-19)). The second is that it "must be rational for the parties…not to be much concerned for what might be gained above what can be guaranteed…by adopting the alternative whose worst outcome is better than the worst outcomes of all the other alternatives."(*Justice as Fairness*, 98); and the third is that "the worst outcomes of all the other alternatives are significantly below the guaranteeable

level"(*Justice as Fairness*, 98). As we will see, the key question in this paper may be posed as the choice between, on the one hand, his two principles of justice and, on the other, what he calls 'the mixed conception', which agrees with his first principle but rejects the difference principle in favor of some other aggregative principle. For this choice, Rawls (*Justice as Fairness*, 119-24) stresses the second condition. The rule I ultimately argue for can accord with to this condition, as long as "not much concerned" means "substantially less concerned" rather than "not concerned at all."

[22] See, e.g., d'Aspremont and Gevers, "Social Welfare Functionals and Interpersonal Comparability," 512-13; Bleichrodt et al, "Equity weights."

[23] For examples of the former, see Larry G. Epstein and Uzi Segal, "Quadratic Social Welfare Functions," *Journal of Political Economy* 100.4 (1992): 691-712; and Simon Grant, Atsushi Kajii, Ben Polak, and Zvi Safra, "Generalized Utilitarianism and Harsanyi's Impartial Observer Theorem," *Econometrica* 78.6 (2010): 1939-71. For examples of the latter, see Charles Blackorby, David Donaldson, and Philippe Mongin, "Social Aggregation without the Expected Utility Hypothesis," Discussion Paper No. 00-18, Department of Economics, University of British Columbia, 2000; Edi Karni and Zvi Safra, "Individual Sense of Justice: A Utility Representation," *Econometrica* 70.1 (2002): 263-84; and Jens Leth Hougaard and Hans Keiding, "Rawlsian maximin, Dutch books, and non-additive expected utility," *Mathematical Social Sciences* 50 (2005): 239-51.

[24] John Quiggin. "Extensions", *Generalized Expected Utility Theory: the Rank-Dependent Model* (Kluwer, 1993), 186.

[25] Quiggin, "Extensions," 187.

[26] Harsanyi, "Can the Maximin Principle?," footnote 10, 598.

[27] A generalization of Rawls's two-rule picture is one on which possible outcomes are divided into strata—e.g., that in which people lack basic liberties, that in which people have basic liberties but not enough money to meet their basic needs, and that in which people have more than enough—and in which only the needs of those in the lowest occupied strata are to be taken into account. Here, the distributional question will be relevant within the lowest occupied stratum. For example, if there are people who lack basic liberties, we can ask whether it is more important to secure a few more liberties for those who have the least or to secure many more for those who have more liberties but still lack some; or if everyone has basic liberties but some don't have enough money, we can ask whether it is

more important to increase the well-being of the worst off in this group somewhat or to increase the well-being of the somewhat-better-off substantially.

[28] Formally: if $R = \{r(p) \mid r \text{ is a reasonable risk function}\}$ is the set of reasonable risk attitudes, then the most risk-avoidant of the reasonable risk-attitudes is $r'(p) = \min_{r \in R} r(p)$. (We will tentatively assume that r' itself has the properties of a risk function, and is convex.) r'(p) is the default risk attitude, and we ought to maximize weighted-rank utility with $I(p) = r'(p)$.

[29] See, e.g., Rawls, *Theory of Justice*, 111-12, 123-25.

[30] Rawls, *Theory of Justice*, 17-19.

[31] Proponents of the Competing Claims View hold that distributive ethics is a problem of determining how to trade off competing claims (Michael Otsuka and Alex Voorhoeve, "Why it Matters that Some are Worse Off than Others: An Argument Against the Priority View," *Philosophy and Public Affairs* 37.2 (2009): 171-199; Alex Voorhoeve and Marc Fleurbaey, "Egalitarianism and the Separateness of Persons," Utilitas 24.3 (2012): 381-398; Michael Otsuka, "Prioritarianism and the Separateness of Persons," Utilitas 24.3 (2012): 365-380). Relative prioritarianism bears some similarity to this view. The key difference is that the Competing Claims View holds that the strength of an individual's claim depends on how well-off he is relative to other individuals who might be affected by a particular choice, whereas relative prioritarianism (spelled out in terms of competing claims) holds that the strength of an individual's claim depends on how well off he is relative to other individuals in society. Because of this feature, the competing claims view is non-aggregative in the sense that we cannot assign a single value to a distribution that is independent of which other distributions are available.

[32] Rawls, "Some Reasons for the Maximin Criteria," 144. See also Rawls, *Justice as Fairness*, 127.

[33] Kahneman and Tversky discuss of the S-shaped pattern of weighting (Kahneman and Tversky, "Prospect Theory" and Tversky and Kahneman "Advances in Prospect Theory"). Recall from section 5 that Quiggin mentioned this possibility—we can now see how one might defend it.

[34] This objection was raised by Rawls against total utilitarianism, though the above argument focuses on an adaptation to average utilitarianism. See Rawls, *Theory of Justice*, 20-24, 159-67.

[35] Otsuka and Voorhoeve, "Why it Matters."

[36] See Eric Nord, "The trade-off between severity of illness and treatment effect in cost-value analysis of health care," *Health Policy* 24 (1993): Table 1, lns. 1-5, 231; notice that line 4 involves a shift in which group is worst off. Exactly how risk-avoidant subjects were is complicated, since experimenters assume expected utility maximization

and attribute deviations from it to experimental error (See Figure 2, p. 230, where the "standard gamble" method is the method that assumes expected utility maximization)—but it is safe to say that they were not nearly as risk-avoidant as they were inequality-avoidant. More work would need to be done to show that subjects make the same recommendations as relative prioritarianism, though these results are suggestive. Nord et al and Bleichrodt et al propose a more general method for eliciting the weights to be placed on the interests of each group; the latter specifically assume WRU (Eric Nord, Jose Luis Pinto, Jeff Richardson, Paul Menzel, Peter Ubel, "Incorporating Societal Concerns for Fairness in Numerical Valuations of Health Programmes," *Health Economics* 8 (1999): 25-39; Bleichrodt et al, "Equity weights."

[37] See d'Aspremont and Gevers, "Social Welfare Functionals and Interpersonal Comparability," 471.

[38] Ibid.