1 Introduction

This paper weighs in on the ongoing debate over the appropriate balance between efficiency and equity. I explore the relationship between our attitudes about risk and our beliefs about distributive justice. The notion that there is a deep relationship between risk and inequality is not a new one; discussions of the topic date at least from Harsanyi’s work on the subject in the 1950s. Harsanyi argued that the cardinal utility functions appearing in the theory of decisions under uncertainty and the cardinal utility functions in welfare economics should be one and the same (I will refer to this as the Harsanyi framework). In this paper I review Harsanyi’s argument, as well as Rawls’s theory of justice as fairness, which (despite reaching very different conclusions) also links ideas of risk and inequality. I extend this parallel by noting that wealth redistribution can be viewed as an insurance contract from the perspective of the original position, and I explore whether observed levels of inequality in America appear to be consistent with the level of insurance one would demand from the perspective of the original position. I will consider a potential objection to Harsanyi’s view—that individual utilities are additively separable across possible outcomes, while social welfare is not additively separable across individual utilities. I will argue that this objection in fact poses no theoretical difficulty for the Harsanyi framework, though it may make empirical evaluation of the appropriate amount of inequality more difficult. Finally, I will present a loose argument that despite such empirical difficulties, we can still place a lower bound on the amount of inequality consistent with observed risk preferences.

1.1 Harsanyi and Rawls

Harsanyi (1953) argued that the formal similarity between cardinal utility in the theory of risk and in welfare economics was not a coincidence, but represented a basic fact about our method of moral reasoning. According to Harsanyi, we make moral judgments by setting aside facts about our personal interests (or attempting as much) and assessing a situation from an impartial perspective. In the context of judgments concerning social welfare, this entails setting aside facts regarding one’s particular place in society (such as one’s race or gender) when making social value judgments. Harsanyi writes that a social value judgment exhibits the “highest degree” of impartiality if it is made without knowledge of one’s own position. The selection of a just distribution of income, for
instance, is best made without knowledge of one’s own place in that distribution. This desire for impartiality motivates reframing issues of distributive justice in terms of uncertainty: if one assumes an equal chance of occupying any position within society and selects an income distribution by maximizing expected utility, then the cardinal utility function used to make decisions in the face of risk is precisely the same as the function that would be used to construct an impartial social welfare function.

Rawls (1958, 1999) proposed a similar theory of social value judgments, calling the agnostic position of impartiality the “original position”; in this position, one is “behind the veil of ignorance.” However Rawls argued that one would employ a “maximin” decision rule, favoring a society that maximally benefits the least advantaged member of the income distribution, rather than maximizing expected utility.\(^1\)

Although Rawls’s maximin rule has generated a great deal of controversy—including an interchange between Harsanyi and Rawls (see Harsanyi (1975); Rawls (1974))—the general concept of the original position has proven quite popular, suggesting there is a great deal of intuitive appeal in applying feelings about risk and uncertainty to reach impartial judgments concerning social welfare and inequality.

Others have noted the formal similarity between representations of risk and inequality without positing any deeper connection. Atkinson (1970) discusses a number of alternative metrics of inequality, noting that each implicitly assumes a different social welfare function. He proposes a measure closely related to the concept of the certainty equivalent in expected utility theory (an application he calls “economically unrelated but formally similar”). Atkinson’s inequality metric takes the form

\[
I = 1 - \left[ \sum_i \left( \frac{y_i}{\bar{y}} \right)^{1-\epsilon} f(y_i) \right]^{\frac{1}{1-\epsilon}},
\]

which depends on a “coefficient of relative inequality aversion,” \(\epsilon\), parameterizing the curvature of the utility function. Since \(\epsilon\) represents a normative value, rather than a positive one, it cannot be measured empirically, and it must be selected before one can compare levels of inequality or social welfare across alternative distributions.

One cannot help but wonder about the relationship between \(\epsilon\) and the formally identical coefficient of relative risk aversion, which can be estimated empirically, at least in principle. In particular, in light of Harsanyi’s proposed correspondence between expected utility and social welfare, an obvious question is whether \(\epsilon\) should be set equal to empirical estimates of the coefficient of relative risk aversion. In section 2 I explore the implications of answering this question in the affirmative.

### 1.2 Welfare as insurance

Later in this paper I will note that if one accepts the Harsanyi framework, income redistribution can be viewed as a form of insurance. However it is also possible to motivate the analogy between redistribution and insurance by appealing to current insurance practices directly. Typically insurance markets exist when a group of risk averse individuals face idiosyncratic risks across outcomes.

\(^1\)Rawls’s theory is in fact considerably more nuanced, as preferences over resource distributions are subject to the requirement that all citizens enjoy maximal equal liberty, and the distribution in question concerns “primary goods,” of which monetary income is only a single component. Nevertheless, for the purposes of economic modeling, Rawls’s maximin rule with respect to wealth is a useful simplification.
to be realized at a later date. By entering an insurance pool, each individual can effectively transfer resources from the possible state of the world in which he or she realizes the good outcome to the bad outcome state. Of course, it is not strictly necessary that the good or bad state actually occur after the date at which insurance is purchased—the state may occur beforehand (or may be correlated with something that occurs beforehand). A key example is in the case of health insurance. Many diseases are correlated with genetic information that can be observed, via genetic testing, prior to the onset of the disease. Yet there seems to be broad agreement that information from such tests should not be used when setting insurance premiums. For example, in 2008 the Genetic Information Nondiscrimination Act, or GINA (which prevented insurance companies from using information from genetic tests when setting premiums) passed unanimously in the Senate and passed 414–1 in the House of Representatives.

I suspect that GINA received broad support for a number of reasons, but a key one is the notion that any one of us could harbor genetic predispositions for such a disease—it seems unfair to charge higher premiums to those with a specific gene (over which they had no control). Other factors, such as the fact that genetic testing is not yet widespread, or that the results of such tests are currently confidential, surely made the law easier to pass and enforce, but these factors seem less relevant from an ethical point of view. Of course, as with any type of insurance, if many individuals have private information revealing their level of risk, the market risks unraveling due to adverse selection (see Rothschild and Stiglitz (1976)). Mandated insurance can mitigate this problem; indeed, preventing unraveling in the face of preexisting conditions was a key argument in favor of the insurance mandate in the care reform bill passed in 2010 (see, for example, Starr (2010)). And while some aspects of that bill were unpopular with the public, a broad majority (76%) supported requiring coverage of preexisting conditions (see Kliff (2010)).

It seems there is a general consensus that one should be able to insure against potentially bad health outcomes, even if the traits causing those outcomes predate the purchase of insurance. A reasonable question, then, is whether one should be able to insure against bad outcomes caused by other, non-health related traits. In particular, should one be able to insure against traits that are apparent from birth? At first this seems like a strange thought—if the trait is apparent from birth, there is no uncertainty, so how could we insure against such an “outcome”? But it turns out this notion might make perfect sense.

Recall that insurance is simply a transfer of resources from possible good outcomes to possible bad outcomes. We can easily think about such a transfer in the case of genetic diseases: one is born either with or without the disease gene, and prior to revealing one’s genetic state, one can transfer resources to the possible bad state (in which one has the disease gene) from the good state (in which one lacks the disease gene). In effect, such insurance simply redistributes resources from those who are genetically fortunate to those who are genetically unlucky. This practice raises an obvious question: why not institute similar redistribution on the basis of other, non-health traits? And furthermore, why not extend such redistribution to traits that are readily apparent from birth? Of course, such an insurance market for traits evident at birth would unravel, for those with favorable traits would choose not to purchase insurance. Unlike genetic diseases, there is no period during which insurance can be purchased prior to the revelation of visible genetic traits. Nevertheless, we can certainly imagine what might occur if it were possible to enter into insurance contracts while behind the veil of ignorance, prior to learning one’s own traits. Presumably members of the original position would choose to insure against the possibility that they are revealed to have unfavorable traits, effectively redistributing away from those who will have good traits, and toward those who will have bad traits.
Finally, consider the possibility that measuring all of one’s specific traits (and the way in which those traits relate to economic outcomes) is prohibitively complex, and that as a result insurance contracts tied to such traits are impossible. Yet suppose insurance contracts based on income (which is straightforward to measure) are feasible. Would people behind the veil of ignorance choose to buy insurance? The answer would likely depend on the cost of the insurance. Since income is not genetically determined, there would be moral hazard costs associated with such insurance. But moral hazard need not rule out insurance entirely—there is moral hazard associated with homeowners insurance, for example, yet the market for such insurance is alive and well. Thus it is plausible that people behind the veil of ignorance would choose to enter income-dependent insurance contracts, even if they choose not to insure completely. The effect of such insurance contracts would be income redistribution.

2 Taking Harsanyi seriously

2.1 Estimating optimal redistribution

Suppose we adopt Harsanyi’s model of the veil of ignorance: one selects a social structure (or, more specifically, an income distribution) without knowing one’s own place in that distribution. Furthermore, suppose the selection is made by maximizing expected utility.

Of course, the domain of possible income distributions is restricted. The total level of income, regardless of distribution, is limited by factors such as capital and technology. And due to the behavioral responses to income taxes, there is likely a substantial efficiency cost to income redistribution. (See Mirrlees (1971) for a seminal work on the theory of these efficiency costs.) Following Atkinson, we can apply results from expected utility theory to reach conclusions concerning social welfare. In particular, the efficiency costs of income redistribution correspond to ordinary insurance with an actuarially unfair premium. The standard results from the theory of optimal insurance then apply: if insurance is actuarially fair then the marginal utility of consumption should be equated across possible outcomes. Analogously, if there is no deadweight loss from income redistribution, then income should be redistributed to achieve complete equality, with 100% marginal tax rates on incomes over this amount. Indeed, this is a familiar result from applying the utilitarian framework to optimal taxation theory.

These conclusions provide a starting point for an appraisal of justice in our own society. Specifically, under this framework we can say that our society is “just” if the level extent of income redistribution is precisely that which would be selected by someone insuring against the risk of occupying an unfavorable position in society, in the face of an actuarially unfair insurance premium equivalent to the efficiency cost of redistributive taxation. This analysis requires empirical estimates of two parameters—the coefficient of relative risk aversion and the efficiency cost of redistributive taxation—in addition to data on the societal income distribution.

2 Of course, there are many economic outcomes of interest besides income, and indeed, there are many other, non-economic outcomes which impact one’s quality of life, some of which could be redistributed. For simplicity, I will refer to income in this paper, but keep in mind that many of these results could be extended.

3 Although we frequently refer to welfare and other redistributive policies as “social insurance”, the meaning of “insurance” here is slightly different. Whereas we typically think of social insurance as protecting against one’s own future outcomes (e.g. unemployment or a workplace injury), from the perspective of the veil of ignorance, income redistribution insures against the possibility of having traits that are immutable once the veil is lifted, such as being a member of a subjugated class or gender.
Turning first to the coefficient of relative risk aversion (CRRA), a number of authors have attempted to measure this parameter empirically. Perhaps most useful in our context, Szpiro (1986) used data on liability and property insurance to estimate risk aversion, finding support for the constant relative risk aversion functional form, and estimating that the CRRA lies between 1.2 and 1.8. More recently, Chetty (2006) uses 33 studies of wage and income elasticities to conclude that under the assumptions of expected utility theory, the CRRA is at most two. Barro (2009) calibrates his model of rare disasters using a CRRA of four, which he notes is consistent with Chetty’s analysis, as his model has independent parameters for risk aversion and consumption-leisure substitution.

Clearly estimates of risk aversion vary considerably, depending on the context in which they are measured. I will take Szpiro’s estimate of 1.2 to 1.8 as a starting point, as it is based on revealed preferences from insurance purchases—precisely the concept we would like to apply to the original position. Note that this value is much higher than typical valuations of $\epsilon$ in (1) in the context of measuring income inequality. Indeed, the census report discussing trends in inequality from 1947 to 1998 erroneously states that $\epsilon$ must lie between zero and one (see Jones Jr and Weinberg (2000)). This discrepancy highlights the importance of determining whether the CRRA ought to equal the coefficient of inequality aversion $\epsilon$ in Atkinson’s functional form, for a value of $\epsilon$ between zero and one would have quite different normative implications.

The social welfare function implied by (1) is

$$W = \left[ \frac{1}{n} \sum_i y_i^{1-\epsilon} \right]^{\frac{1}{1-\epsilon}}.$$  

where $y_i$ is each individual’s wealth. From this we can see that if two citizens $a$ and $b$ have incomes of $y_a$ and $y_b$, then

$$\left( \frac{\partial W}{\partial y_a} \right) \left( \frac{\partial W}{\partial y_b} \right) = \left( \frac{y_a}{y_b} \right)^{-\epsilon}.$$  

From expected utility theory, we know that a consumer optimizes by buying insurance up to the point at which the cost of transferring consumption from one state to another is equal to the ratio of marginal utilities in those states. An analogous relationship holds here. Suppose (without loss of generality) that $y_a < y_b$ (so that the right side of (3) is greater than one). Let $\alpha$ denote the efficiency cost of transferring $1$ from $b$ to $a$, meaning that $b$’s taxes must go up by $1 + \alpha$ in order to pay $1$ to consumer $a$. For ease of notation, let $y_b = ky_a$, so that $y_a/y_b = 1/k$. Then at the social optimum, we must have

$$k^\epsilon = 1 + \alpha.$$  

This equation implies that for a given $\epsilon$, for any income ratio we can calculate the efficiency cost below which it becomes welfare-enhancing to redistribute. If the true efficiency cost of redistribution is below that threshold, then from the perspective of the original position, one would prefer to “buy more insurance” by transferring resources from better outcomes to worse outcomes, i.e., by redistributing wealth from rich to poor.

Of course, calculating the marginal cost of income redistribution is extremely difficult, as it is hard to find instances of plausibly exogenous variation in income redistribution. Still, some studies have attempted to estimate these costs. Using a calibrated computational general equilibrium
model, Ballard (1988) finds that “a tax-financed universal cash grant generates losses for the higher-income groups that exceed the gains of the lower-income groups by from 50 to 130 percent.” In other words, \( \alpha \in [.5, 1.3] \). This cost would not remain constant as tax levels change overall; indeed, public finance theory suggests that the costs should rise with the square of the tax rate (Harberger (1971)). Thus while we may be able to determine whether the current degree of redistribution is too high or too low, we will likely be unable to quantify the optimal level of redistribution, as we do not know how quickly the marginal cost of public funds changes.

Despite this uncertainty involving the efficiency costs of redistribution, we can calculate the cost that would be required to justify various levels of inequality, using (4). For example, Figure 1 shows the efficiency cost required to justify wealthy-to-poor income ratios between one and five for the range of CRRA values suggested by Szpiro.

**Optimal redistribution for \( \varepsilon = 1.8, \varepsilon = 1.2 \)**

![Graph showing optimal redistribution](image)

Figure 1: Optimal redistribution using Szpiro’s CRRA
This plot indicates that the cost of redistribution would have to be quite high ($\alpha > 15$) in order for a member of the original position to tolerate an income ratio of five. In other words if the efficiency cost of redistributing one dollar is less than $15$, then at the optimum, the highest income in the population must be less than five times the lowest.

To see how sensitive this result is to the chosen value for $\epsilon$, we also can use Barro’s CRRA to calculate the optimal value of $\alpha$ for a range of values of $k$, as shown in Figure 2. For $\epsilon = 4$, the required value of $\alpha$ is over 600!

The final step in this empirical analysis is to consider the actual level of income inequality, and ask whether it is plausibly optimal in light of these calculations. Importantly, since we wish to evaluate the current extent of redistribution, we are interested in the post-tax distribution of income. Moreover, we should include in-kind transfers such as food stamps and medical insurance.
where possible, as these benefits constitute a large portion of existing government aid. According to a research paper on income inequality published by the Heritage Foundation, after appropriate adjustments, the ratio of income earned by the top income quintile to that earned by the bottom was 4.21 (see Rector and Hederman Jr (2004)).

The inconsistency between this ratio and the calculations in Figure 1 is remarkable. The efficiency cost necessary to justify an income ratio of 4.21 is about 15. It is difficult to directly compare the value of $\alpha$ (which indicates the efficiency cost of transferring between any two individuals) with the parameter estimated by Ballard (which indicates an average efficiency cost of raising overall lump sum grants by increasing marginal tax rates). Nevertheless, his estimates are an order of magnitude lower than the cost that would be required to justify the gap between the first and fifth income quintiles from the perspective of the original position. Put differently, at least 40% of the U.S. income distribution has an income that appears inconsistent with our best estimate of the efficiency costs of redistribution.

2.2 Could the present level of inequality be justified?

The above result seems to suggest that, from the perspective of Harsanyi’s original position, the present level of income inequality is unjustly high. But before we accept this conclusion, it is worth considering two objections to these calculations. First, one might object that Szpiro’s value of the CRRA is simply incorrect. Some empirical studies estimating the CRRA find a much lower value—typically between .5 and 1 (see Chetty (2006) for an overview of this literature). Of course, some estimates (such as Barro’s) are substantially higher as well, so perhaps Szpiro’s estimate is a reasonable midpoint. Nevertheless, given the importance of the CRRA for our analysis, it is worth exploring the objection that our estimate is too high.

Reducing the CRRA dramatically lowers the efficiency cost required to justify a given variation in income. Chetty reports 15 different labor supply elasticity estimates, along with the CRRA implied by each. The average is a CRRA of .71. He notes that macroeconomic and trend studies tend to find higher implied CRRA estimates—the average CRRA among such studies was 1.37—while microeconomic studies yield lower estimates, often below one.

A related objection might be that the notion of evaluating risk from behind the veil of ignorance is ill-defined. Implicit in the above reasoning was the assumption that people with different incomes nevertheless have similar risk preferences (i.e., that there is a constant coefficient of relative risk aversion that is fixed across wealth levels). If risk preferences vary with income, however, it is unclear whose preferences we should favor in the original position. This objection is, in a sense, a deeper one than the worry that Szpiro’s estimate of the CRRA is incorrect, for it cannot be resolved by gathering better data on empirical risk preferences. Nevertheless, it seems safe to suppose that even if the CRRA varies with income, the risk preferences assumed in the original position should lie somewhere within the range observed empirically. So a useful starting point is to examine how our results would vary across the range of typical CRRA estimates.

Figure 3 shows the efficiency cost of redistribution corresponding to various income ratios at the social optimum for $\epsilon = 1$ and $\epsilon = .5$. Although lowering $\epsilon$ sharply reduces the efficiency cost of redistribution required to justify a given level of inequality, it appears this adjustment is still not enough to justify the extent of inequality in the current American income distribution. To consider

---

4This paper uses data from the U.S. Census’s 2002 Current Population Survey. In addition to adjusting for taxes and in-kind transfers, this figure adjusts for the unequal number of people in each income quintile (an artifact of the Census’s method of aggregating data at the level of the household, rather than the individual).
the most conservative case, suppose $\epsilon = .5$ and that $\alpha$ is 1.3 (the upper limit of the range estimated by Ballard). Then an expected utility maximizing agent behind the veil of ignorance would still choose to redistribute whenever $k > (1 + \alpha)^{1/\epsilon}$, that is, whenever $k > 5.3$. Yet the highest incomes in the American income distribution are far more than 5.3 times the lowest.\(^5\) This suggests that

\(^5\)Unfortunately, aside from the Rector and Hederman Jr (2004), there are few studies of the US income distribution that incorporate federal and state taxes as well as in-kind transfers, so it is difficult to be precise about the fraction of the American population that would be subject to greater redistribution under these parameters. Nevertheless, the results in Rector and Hederman Jr (2004) as well as in other studies of pre-tax income, such as Piketty and Saez (2003) indicate that a substantial fraction of the population could be arranged into pairs in which one income was at least 5.3 times the other. A report from the Congressional Budget Office (2007) (which adjusts only for federal taxes) finds that in 2005 the mean income in the top decile was $264,000, while the mean income in the lowest quintile was $15,000, suggesting that the post-tax mean income of the top decile is almost certainly over 5.3 times that of the bottom decile.
lowering our CRRRA estimate is not, in itself, sufficient to render current income inequality consistent with the Harsanyi framework.

Another possible objection is that Ballard’s estimates of efficiency costs are far too low. Ballard focuses on the short term efficiency impacts of redistribution. But an important component of the cost of taxation is the long-run impact on economic growth. Unfortunately it is difficult to empirically estimate the effect of taxes on growth. However, even if taxation is quite detrimental to economic growth, there is reason to think those costs still might not justify current inequality. Specifically, we are assuming that marginal utility of wealth declines as wealth rises. (This explains why a high CRRRA calls for a great deal of redistribution—the marginal utility of wealth is far lower for richer members of society than for poorer members.) The costs of lower economic growth will be greatest for people in the distant future, and so long as taxation doesn’t drive economic growth to zero, those people will be much richer than poor members of the current wealth distribution (and possibly even the rich members). Thus while we cannot set aside this critique entirely, it appears unlikely that the impact of taxation on economic growth could be large enough to justify observed inequality for Szpiro’s CRRRA.

If both objections are correct (Szpiro’s CRRRA is too high and Ballard’s efficiency cost estimates are too low), it is possible that present inequality could be justified under the Harsanyi framework. But neither objection alone seems to suffice. I now turn to the possibility that inequality is indeed too high.

2.3 Excessive inequality

Even if the above analysis is completely correct, it may still not be clear that we have too much inequality in the current income distribution. One might grant that our estimates of the CRRRA are indeed higher than than estimates of the implied parameter of inequality aversion, and that these parameters ought to be equal in a just society. But when they are not, who is to say that the inequality aversion parameter should increase, rather than the CRRRA decrease? In other words, all we have shown so far is an inconsistency between two parameters that ought to equal, but the Harsanyi framework alone does not tell us which parameter is the “right” value.

We can begin by identifying circumstances in which we would expect risk aversion and implied inequality aversion to have similar values. We would expect to see such a similarity when the range of potential personal outcomes is similar to the range of positions in society. In the extreme, if each individual risks occupying any point in the income distribution at a later date, then the distribution of one’s possible outcomes is simply the full distribution of the population, just as in the original position. Thus in highly mobile societies (with a high likelihood of occupying a future place very different from one’s present position) we should expect citizens to favor policies that redistribute wealth according to the Harsanyi framework.

Of course, in practice it is hard to envision realizing this level of social mobility, for there are typically immutable traits (such as race, gender, or intelligence) which are correlated with economic standing, and thus the range of one’s own possible outcomes is typically narrower than the full range of positions in society. In such cases, purely self-interested agents will employ the Harsanyi approach (equating risk and inequality) only across their limited range of future positions, rather than across all positions in society. Risk averse agents in a society of limited social mobility would thus favor policies that redistribute across portions of the economic distribution which they are more likely to one day occupy. And they would favor policies that tend to benefit those portions of the economic distribution, relative to portions of the distribution that they are unlikely to occupy.
This yields two key empirical implications. First, the CRRA and the implied coefficient of inequality should be closer together in societies with a great deal of social mobility. And second, to the extent that a society falls short of perfect social mobility, we should expect policy to benefit those portions of the income distribution occupied by those with the most influence over policy.

Of course, this reasoning is based on a number of unrealistic assumptions. Agents are not purely self-interested—people exhibit aversion to inequality even when their own outcomes are secure (see Kroll and Davidovitz (2003)). Nevertheless, one might expect that this concern for others would typically be weaker than concern for oneself, and as a result, that those influencing policy would still favor wealth distributions that benefit positions they are likely to occupy.

I do not intend to delve into the theory of political economy in this paper, but it is interesting to note that our conclusion from section 2.1 (that redistribution appears low relative to a just society by Harsanyi’s measure) is consistent with what we would expect in a society with imperfect social mobility and in which wealthy people have more influence on policy than the poor. Moreover, in such a society, one could (in principle) accurately measure risk aversion, since agents have no incentive to conceal their true risk preferences (they internalize all costs associated with decisions involving risk). Implied inequality aversion, by contrast, would appear lower than risk aversion, as wealthy agents (who influence policy) would fail to fully internalize the utility cost of being poor, since there is little chance of becoming poor themselves. If this reasoning is correct for our own society, then there is reason to think our current level of inequality is unjustly high, rather than that our level of risk aversion is too overestimated.

Alternatively, we could reach a similar conclusion using recent advances in the theory of inconsistent preferences. Suppose we accept Harsanyi’s claim that in a just society risk and inequality are treated similarly, yet in practice we observe that risk aversion appears stronger than inequality aversion, even when people attempt to create just policies. This would be an apparent case of inconsistent preferences, as analyzed by Bernheim and Rangel (2009): preferences appear to differ by context, even when the change in context is irrelevant. Bernheim and Rangel discuss how one can use “refinements”—the process of excluding observations in which expressed preferences are likely to be biased—in order to discern an agent’s true preferences. Chetty et al. (2009) apply this theory to their study of tax salience, arguing that agents express their true preferences when facing salient taxes. We could take a similar approach for our analysis. We might think that risk aversion appears stronger than inequality aversion because it is difficult for agents to internalize the welfare state of positions in society that they will likely never occupy. (A privileged heir to a fortune might have difficulty imagining a life in destitute poverty, for example.) In this case we could use the logic of refinements to exclude observations regarding inequality aversion when estimating agents’ true preferences, considering instead only their preferences with respect to risk (a domain in which they can better internalize their range of outcomes).

Either line of reasoning suggests we should trust our estimates of risk aversion more than estimates of implied inequality aversion. As a result, the Harsanyi framework, combined with the social welfare function underlying Atkinson’s inequality metric, suggest that the current extent of inequality in America is unjustly high.
3 Relaxing the separability assumption

3.1 Separability across outcomes, people, and time

A key assumption of the social welfare function used in the previous section is that social welfare is additively separable in utilities. In other words, the marginal social utility of wealth for each person in society is independent of the rest of the distribution. Such separability is a common assumption in welfare economics; nevertheless, it is quite a strong assumption, so it is worth examining its implications more closely.

First, separability seems quite a natural assumption in the context of decisions involving risk. (Some have objected to it even in this context, but by and large expected utility theory is the standard approach to decisions involving risk—see Schoemaker (1982) for a review.) It is not hard to see why separability has seemed quite plausible in this application. For although one might consider many possible outcomes when making a decision, ultimately the outcomes are exclusive. Only one state of the world actually occurs, and that state has no interaction with other states that did not come to pass.

As discussed in the introduction, social welfare functions often aggregate utilities across individuals in the same way that individuals aggregate expected utilities across possible outcomes. (Indeed, the social welfare function from Atkinson’s analysis has this form.) Additionally, macroeconomic models frequently aggregate an individual’s utility over time in the same way, such that one’s utility in one period is additively separable from utility in other periods.

One might worry about extending the additive separability assumption to aggregation across individuals and across time, however. For although it seems plausible that the utility from one outcome should not, in general, be dependent on other states that did not occur, it is less clear that the utility from one level of income should be independent of all other incomes in society. Indeed, the notion that utility may depend on relative income has a long history (see, for example, Boskin and Sheshinski (1978)). Broome (1995) discusses the separability assumption in each of these three contexts at length.

In this section I first consider the constraints imposed by the separability assumption. I then argue that relaxing the separability assumption does not necessarily undermine Harsanyi’s original position as a framework for evaluating an income distribution, though it may make empirical calculations more difficult. I will conclude the section by reexamining our initial calculation that the current level of inequality is too high.

Before dismissing the separability assumption outright, it is worth examining what the assumption actually entails. Separability implies that a high level utility function consists of a (perhaps weighted) summation of lower level functions, for example

$$U = \sum_i u(y_i) f(y_i).$$

The higher level function might represent social welfare, for example, in which case $i$ indexes individuals and $u(\cdot)$ represents each individual’s utility function, while $f(\cdot)$ governs the social weight associated with each individual’s welfare. Alternatively, the higher level function might represent expected utility, so that $i$ indexes possible outcomes, $f(\cdot)$ indicates the probability of each outcome, and $u(\cdot)$ represents the Bernoulli utility function for each potential outcome. Finally, $U$ could represent an individual’s lifetime utility, with $i$ indexing time periods, $f(\cdot)$ discounting future periods, and $u(\cdot)$ representing utility within each period.
Importantly, when aggregating across more than one of these levels, the lower level utility function need not be the same. For example, when aggregating utility across time periods, each of which has a number of possible outcomes, one need not weigh tradeoffs across outcomes within a given period (risk) in the same way one weighs tradeoffs in consumption levels across time (intertemporal substitution). (See Epstein and Zin (1989) for a detailed discussion.) This implies that the Harsanyi framework is not just a consequence of additive separability, but is in fact a stronger assertion, for it states that tradeoffs in wealth across individuals should be weighed in “the same way” (in some sense) as tradeoffs across possible outcomes for a given individual, while separability allows aggregating quite differently across outcomes and individuals.

3.2 Separability and the Harsanyi original position

One might worry the justification for separability across individual utilities in the social welfare function is much weaker than the justification for separability across mutually exclusive outcomes in expected utility theory. After all, individuals of various wealths interact with one another—surely my utility, varies depending on whether those around me are relatively richer or poorer than I am. Does this undermine the Harsanyi argument that risk and inequality should be treated similarly? After all, separability demands a very specific kind of utility aggregation. If we aggregate utilities separably in one context and not in another, we can hardly claim to be aggregating “in the same way” in each case.

In fact, I do not believe this difference poses a problem for the Harsanyi framework. To see why, we must examine in more depth the nature of the correspondence between risk and inequality in Harsanyi’s approach. Specifically, in Harsanyi’s original position, one evaluates a social structure by imagining that one will occupy any place within that society with equal probability. Suppose that society contains \(N\) individuals, and one wishes to evaluate a wealth distribution \(X = (x_1, x_2, \ldots, x_N)\), a vector of length \(N\) indicating the wealth of each member of society. Suppose first that utilities are additively separable, so that the utility of income \(x_i\) does not depend on any \(x_j\), \(j \neq i\). Denote this utility function \(u(x_i)\). The expected utility from the perspective of the original position, then, is

\[
U(X) = \frac{1}{N} \sum_{i=1}^{N} u(x_i). \tag{6}
\]

In this case \(U(\cdot)\) is indeed an additively separable function of each \(x_i\). If instead the utility of a given position depends not only on one’s own income, but also on the full distribution of income, then \(i\)’s utility is a function of both \(x_i\) and \(x_{-i}\) (the vector of incomes other than \(i\)’s); we could write \(u(x_i, x_{-i})\). Importantly, this alters only the function inside the summation in (6), not the summation itself:

\[
U(X) = \frac{1}{N} \sum_{i=1}^{N} u(x_i, x_{-i}). \tag{7}
\]

With this functional form, \(U(\cdot)\) is no longer an additively separable of the \(x_i\). However, the method of aggregating utilities still takes the expected value form. Thus Harsanyi’s approach is still perfectly reasonable. His framework does not require that \(U(\cdot)\) be an additively separable function of the \(x_i\). Rather, it requires that the utility function used to evaluate outcomes in the
original position be the same utility function used when making decisions involving risk. That is, the $u(x_i, x_{-i})$ in (7) should be the same function we use when buying car insurance.

It is now clear why the notion of separability is a tricky one—separability with respect to *what*? When we make decisions involving risk, we indeed maximize an expected utility function that is additively separable function of our utility across each outcome we might experience, but the lower level utility function used to evaluate each outcome may still be dependent on others’ incomes within that possible state of the world.

However, although a utility function that depends on others’ incomes is perfectly consistent with Harsanyi’s framework in theory, such a function makes empirical calculations (like that presented in section 2.1) considerably more difficult in practice. Why? Because typical insurance contracts depend only on one’s own outcomes, rather than on outcomes for others. If one’s utility is independent of others’ incomes, then observed insurance purchases allow us to estimate $u(x_i)$, which is the function required to evaluate an income distribution in (6). But if one’s utility also depends on others’ incomes—$u(x_i, x_{-i})$—then observation reveals only how $u(x_i, x_{-i})$ varies with $x_i$, not with $x_{-i}$. Note that this need not be a problem in theory—if insurance contracts truly allowed reallocating consumption across all possible states of the world, one could insure not only against variation in one’s own income, but also against variation in the income distribution as a whole. For example, in addition to insuring against one’s own income falling, one might wish to insure against the general distribution rising while one’s own income remains the same. If such insurance products were available, we could estimate the way in which $u(x_i, x_{-i})$ varies with both $x_i$ and $x_{-i}$.

Unfortunately the incompleteness of insurance markets in this respect is a major problem for empirically estimating the just amount of redistribution. If we know only how $u(x_i, x_{-i})$ varies with $x_i$, we can estimate (7) for the income distribution observed in our own society, since $x_{-i}$ is fixed; however we cannot compare this estimate to other hypothetical income distributions. Suppose we wish to compare the current distribution, $X$, to another feasible distribution $Y$. As the number of individuals grows large, the impact of $x_i$ on the distribution $X$ becomes minuscule, so we can rewrite $u(x_i, x_{-i})$ as $u(x_i, X)$—a function of one’s own income and the fixed distribution as a whole. And by observing insurance consumption, we can see how $u(x_i, X)$ varies with $x_i$, thus we can estimate

$$U(X) = \frac{1}{N} \sum_{i=1}^{N} u(x_i, X),$$

for some normalization of $u(\cdot, X)$. We run into trouble, however, when we attempt to estimate

$$U(Y) = \frac{1}{N} \sum_{i=1}^{N} u(y_i, Y),$$  \hspace{1cm} (8)

for a distribution $Y \neq X$, since we have no empirical method of estimating how $u(x_i, X)$ varies with $X$. Although this is clearly a major problem for the empirical application of Harsanyi’s approach, we are not quite at a loss yet. After all, though we cannot estimate (8) empirically, we can estimate

$$\frac{1}{N} \sum_{i=1}^{N} u(y_i, X).$$

That is, we can estimate the expected utility of having each income in the distribution $Y$ if that income were received in the distribution $X$. Moreover, it may be possible to establish at least an ordinal ranking between $u(c, X)$ and $u(c, Y)$ for a fixed personal income $c$, with the help of some
basic assumptions about the relationship between one’s utility and the societal income distribution.
Note that this is effectively the same as a calculation assuming utility is solely a function of personal
income, yet with restrictions introduced based on the assumptions involved in ranking \( u(c, X) \) and
\( u(c, Y) \).

So what can we say about the relationship between \( u(c, X) \) and \( u(c, Y) \), based on a comparison
\( X \) and \( Y \)? This is a complex question, and it is rather difficult to find even basic axioms that appear
true in all cases. One might think, for example, that one gets more utility from a given income if
one is relatively higher in the distribution, i.e., if others’ incomes are lower. On the other hand, one
presumably gets some utility from others’ utility, which would favor distributions in which other’s
incomes are higher.

One line of reasoning, which is by no means conclusive but might prove suggestive, is that one is
concerned both with one’s absolute personal income and with one’s relative position in the income
distribution, but that the former concern is dominant at low incomes, while the latter dominates at
high incomes. I find this possibility intuitively compelling when thinking about individual examples.
For example, when I imagine being very rich, I imagine a great deal of concern with my relative
position—whether the top 5%, the top %1, or the top .01%. I further imagine that much of the
utility gained from the consumption of things like luxury cars or custom architecture would be
highly dependent on what others are consuming. Moreover, I would tend to consume things in
limited supply, for which the price is likely to be strongly related to demand. If there is a fairly
fixed supply of large diamonds, for example, they could be consumed by those at the top of the
income distribution, regardless of the absolute level of income at the top of the distribution. Thus
varying the level of my income at the top of the distribution might make little difference in the
bundle of goods I actually consume. On the other hand, when I imagine being destitute, I imagine
being wholly concerned with my own level of consumption, regardless of the overall distribution.
Whether I am a member of the bottom 1% or 5% would make little difference to me if I am
struggling to survive. If most of the products I would consume (such as food and housing) come
from industries that are fairly competitive and are thus sold at nearly marginal cost, then variations
in my income would have a large impact on the bundle of goods I can afford, and thus such variation
would have a large impact on my utility. Obviously this reasoning is largely speculative, and to
be sure, I encounter salience problems when trying to imagine the behavior of my utility function
at either extreme of the income distribution. That said, I think this approach is at least plausible
enough to serve as a starting point when thinking about utility as a function of both absolute and
relative income.

What, then, would this line of reasoning suggest about the shape of \( u(\cdot, \cdot) \)? One way to roughly
capture the above reasoning is to approximate the discrete distribution \( X \) with the continuous and
differentiable CDF \( F_X(\cdot) \), and rewrite \( u(x_i, X) \) as \( v(x_i, F_X(x_i)) \). Note that this is a simplification—
it assumes that one’s utility depends on relative income only through one’s ordinal ranking in
the distribution, and not at all on the specific shape of the distribution. Still, this may be a
useful and tractable approximation. In terms of this functional form, our previous reasoning sug-
gests that the partial derivative with respect to each argument is nonnegative: \( v_1(x_i, F_X(x_i)) \geq 0 \)
and \( v_2(x_i, F_X(x_i)) \geq 0 \), and as usual I assume diminishing marginal utility of absolute income:
\( v_{11}(x_i, F_X(x_i)) \leq 0 \). This reasoning also suggests that for low incomes, only absolute income
matters:
\[
\exists x : \forall x_i < x : v_2(x_i, F_X(x_i)) = 0. \tag{9}
\]
Below the income \( x \) the derivative of \( v \) with respect to the second argument is zero, so the
cross partial derivative is clearly zero as well. Above \( x \), the cross-partial is positive, so we have: 
\[
v_{12}(x, F_X(x)) \geq 0.
\]
In other words, as income rises, the utility gained by varying one’s relative position rises, while the utility gained by varying absolute income declines.

To consider the implications of this functional form, suppose we again consider the distributions \( X \) and \( Y \). As noted earlier, we can empirically estimate the way that \( u(x_i, X) \) varies with \( x_i \), since this is the type of variation upon which typical insurance contracts are based. Such variation causes utility to change for two reasons: first, absolute income changes, and second, since the overall income distribution is fixed, any change in absolute income causes a corresponding change in one’s relative position within the income distribution. In terms of our rewritten function, we can find this expression by totally differentiating \( v(x_i, F(X)) \) with respect to \( x_i \).

\[
\frac{dv(x_i, F(x_i))}{dx_i} = v_1(x_i, F_X(x_i)) + v_2(x_i, F_X(x_i))f_X(x_i).
\]  
(10)

Now suppose instead that \( x_i \) varies as a result of smooth redistribution within society. In the extreme, we could imagine a policy that redistributed income while preserving the ordinal ranking of wealth. In other words, such an incremental policy adjustment would vary \( x_i \) while preserving \( F_X(x_i) \). Thus the variation in utility as a result of such a policy change is simply the partial derivative of \( v(x_i, F(x_i)) \) with respect to its first argument:

\[
v_1(x_i, F_X(x_i)).
\]  
(11)

Subtracting (11) from (10) yields the error generated if we use the empirically estimated utility function to approximate changes in utility from redistribution:

\[
v_2(x_i, F_X(x_i))f_X(x_i).
\]  
(12)

Below the point \( x \) from (10), the error is zero. Above that point, the error is positive. For a uniform distribution of \( X \), the error rises with wealth, since the cross-partial derivative is positive. If \( f_X(x_i) \) falls as wealth rises, as is true in the American income distribution, then the product in (12) could rise or fall as \( x_i \) increases, depending on whether \( v_2(x_i, F_X(x_i)) \) increases fast enough to compensate for the fall in \( f_X(x_i) \). I will address each of these possibilities in turn.

Suppose first that the error in (12) rises monotonically with wealth. The behavior of this error allows us to place an upper bound on the amount of inequality acceptable when the Harsanyi framework is combined with the new functional form of the utility function outlined above. Recall from (7) that from the perspective of the original position, one seeks an income distribution that maximizes

\[
U(X) = \frac{1}{N} \sum_{i=1}^{N} u(x_i, x_{-i}).
\]

In terms of our rewritten utility function, this becomes

\[
U(X) = \frac{1}{N} \sum_{i=1}^{N} v(x_i, F_X(x_i)).
\]  
(13)

As discussed earlier, if we want to compare the distribution \( X \) to the distribution \( Y \), we would ideally compare (13) to
\[ U(Y) = \frac{1}{N} \sum_{i=1}^{N} v(y_i, F_Y(y_i)). \] (14)

But since the incompleteness of insurance markets prevents empirically estimating how \( v(x_i, F_X(x_i)) \) varies with respect to \( x_i \) while holding \( F_X(x_i) \) fixed, we cannot actually compare (13) to (14). We can, however, compare (13) to

\[ \frac{1}{N} \sum_{i=1}^{N} v(y_i, F_X(y_i)). \] (15)

To select the most just income distribution, we are not just interested in comparing individual income distributions—rather, we are interested in selecting the feasible distribution that maximizes \( U(Y) \) over the domain of all feasible distributions \( Y \)—call this set of feasible \( Y \) (that is, the set of income distributions \( Y \) that can be achieved via redistributive tax policy from our actual income distribution \( X ) \( \Gamma(X) \). Thus, analogously to (14), we would ideally like to select

\[ \max_{Y \in \Gamma(X)} \frac{1}{N} \sum_{i=1}^{N} v(y_i, F_Y(y_i)), \] (16)

and analogously to (15), we cannot actually solve this problem due to incomplete insurance markets, but we can hope to solve

\[ \max_{Y \in \Gamma(X)} \frac{1}{N} \sum_{i=1}^{N} v(y_i, F_X(y_i)). \] (17)

Although estimating the components of (17) and solving for the maximum is admittedly difficult, suppose for the moment that we have done so. Can we say anything about how the solution to (16) would compare? First note that for a distribution \( Y \) to solve (17), it must be the case that the marginal cost of redistributing from any member of that distribution to any poorer member is equal to or higher than the marginal gain in utility resulting from that transfer. To see how the solution to (16) would differ, consider a marginal pair of members of the distribution \( Y \) for whom the cost of redistribution is actually equal to the marginal utility gain. Now consider the true optimal amount of redistribution between this pair, according to (16). If the error term in (12) rises continuously with wealth, then it is easy to see how the pair’s true marginal utilities differ from those estimated when solving (17). Specifically, since the error is positive, both marginal utilities will be at or below that estimated using (17). Assuming the error rises monotonically with wealth, then the error in the marginal utility for the richer member of the will be at least as large as the error for the poorer member, and thus the difference between their marginal utilities according to (17) is an underestimate. Since nothing about the costs of redistribution changes when going from (17) to (16), this indicates that the solution to (16) must require more redistribution than the solution to (17).

A key assumption in this reasoning was that the error in (12) rises monotonically with wealth. And indeed, this seems like a fairly reasonable assumption. For recall that the error is written \( v_2(x_i, F_X(x_i))f_X(x_i) \). As discussed, the first term rises with \( x_i \), so the error can only fall with wealth if \( f_X(x_i) \) falls sharply. But we are assuming that we are already beginning with the maximum of (16)—that is, we are already working with the optimum laid out in section 2.1. For the reasons
presented in that section, there is good reason to think that such an optimal distribution will be quite egalitarian; it will probably not have a small number of individuals making substantially more than everyone else. This reasoning suggests that the density function approximating that optimal distribution will not drop off to a thin right tail—that is, $f_X(x_i)$ will not fall sharply as $x_i$ rises.

Still, this is all very speculative, so it is worth exploring the implications of a nonmonotonic error. By assumption the error is zero below some $\bar{x}$, but what if the error first rises, then falls with wealth? In this case, if we return to the solution to (17) and pick one member with income below $\bar{x}$ and one member with income above that level, the same result applies. But if instead we pick two members with incomes above $\bar{x}$, it is possible that the error is actually larger for the member with the lower income, and thus that the solution to (16) would call for less redistribution within that pair than the solution to (17). As before, we would expect the maximum of (16)—which corrects for the error—to redistribute away from those whose error is very negative, and toward those whose error was small in magnitude. Thus if the error does not rise monotonically with wealth, then we would still expect redistribution away from those in the middle of the distribution (where $f_X(x_i)$ is relatively large) and toward those whose incomes are below $\bar{x}$. But without more information about the specific shape of $v(x_i, F_X(x_i))$, we cannot say whether those high in the income distribution (where the error is decreasing with wealth) will lose or gain under the solution to (16), relative to the maximum of (17).

Although this speculation is far from conclusive, it suggests a couple of key features we can infer about the optimal income distribution, relative to the distribution that optimizes (17). First, we would expect the truly optimal distribution to be more redistributive at the bottom. And second, if the density function approximating the distribution that maximizes (17) is fairly uniform, we would expect the optimal distribution to be more redistributive than the maximum of (17) throughout. If, however, the distribution optimizing (16) has a density $f_X$ that varies greatly, we would expect the optimal distribution to redistribute primarily away from those portions of the distribution where $f_X$ is large.

4 Concluding thoughts

As is evident from the previous sections, admitting relative income dependent utility functions complicates our analysis considerably. Nevertheless, such utility functions present no problem for the Harsanyi approach in principle. Due to the incompleteness of insurance markets in practice, however, we cannot empirically estimate the utility function that should be used to calculate the optimal income distribution, and we can show that our previous calculations of the optimal distribution, from section 2.1 are biased. Yet with a few (admittedly restrictive, but hopefully plausible) assumptions about the way utility relates to absolute and relative income, we can still reach an important result: we should generally expect the optimal income distribution to be more redistributive than our biased estimate from section 2.1, rather than less.

Ultimately these results suggest that the level of inequality in America is greater than that which would be chosen from the original position according to Harsanyi’s framework of expected utility maximization. This conclusion is based on a number of key assumptions, including specific estimates of the coefficient of relative risk aversion and the efficiency costs of redistributive taxation. Obviously each of these assumptions is open for debate; the theoretical simplifications necessary to reach this conclusion are very strong, and the key parameters are very difficult to estimate empirically. Nevertheless, this approach is quite useful insofar as it leverages our revealed preferences
in contexts where we fully internalize future risks (e.g., insurance purchases) to reach conclusions about our implied impartial preferences concerning distributive justice.
References


