Journal of the American Philosophical Association (2023) 670–704
DOI:10.1017/apa.2022.31

Prospect Utilitarianism and the Original Position

ABSTRACT: *Suppose we assume that the parties in the original position took Kahneman and Tversky's prospect theory as constituting their general knowledge of human psychology that survives through the veil of ignorance. How would this change the choice situation of the original position? In this paper, I present what I call 'prospect utilitarianism'. Prospect utilitarianism combines the utilitarian social welfare function with individual utility functions characterized by Kahneman and Tversky's prospect theory. I will argue that, once prospect utilitarianism is on the table, Rawls's original arguments in support of justice as fairness as well as his arguments against utilitarianism are, at best, inconclusive. This shows that how implausible a choice for utilitarianism in the original position is heavily depends on what one assumes to be general knowledge of human psychology that the original contracting parties know.*

KEYWORDS: utilitarianism, Rawls, difference principle, original Position, Prospect theory, Kahneman and Tversky

1. The Knowledge Assumption of Rawls's Original Position

It is well-known that one of the major aims of John Rawls, when he wrote *A Theory of Justice*, was to present a superior alternative to what he deemed to be the predominant moral philosophy of his time: utilitarianism (Rawls 1999: xvii). What Rawls aimed to do was to present an alternative theory of distributive justice that was just as systematic as utilitarianism and at the same time avoided the inherent problems with which Rawls thought all forms of utilitarianism are generally fraught. The resulting theory is what Rawls calls and what is now widely known as *justice as fairness*.

The specific contents of Rawls's theory of distributive justice (i.e., justice as fairness) can be summarized by the following three principles:

1. *The Principle of Equal Basic Liberties:* Each person is to have an equal right to a fully adequate scheme of equal basic liberties compatible with a similar scheme of liberties for others.
2. *The Principle of Fair Equal Opportunity:* Positions and offices that generate social economic inequalities should be made open to all under conditions of fair equal opportunity.
3. *The Difference Principle:* Social and economic inequalities should be arranged in a way that provides the greatest benefit to the least advantaged members of society.

Rawls thought that these three principles would be chosen in serial order by the rational parties of 'the original position' behind 'the veil of ignorance'. The veil of ignorance is a theoretical device that was designed to guarantee impartiality of the resulting agreement made by the people by blinding them from knowing any contingent information about themselves that was considered morally arbitrary—such as one's particular place in society (such as one's class position or social status); one's natural assets and talents; one's conception of the good; specific features of one's psychology (including one's attitude toward risk); the particular circumstances of one's society; to which generation one belongs, and so on (Rawls 1999: 118–19). The parties of the original position are allowed to know only:

> general facts about human society. They understand political affairs and the principles of economic theory; they know the basis of social organization and the laws of human psychology. Indeed, the parties are presumed to know whatever general facts affect the choice of the principles of justice. (Rawls 1999: 119)

Rawls did not specify the specific contents of the type of general knowledge he assumed the rational parties of the original position to know. Yet, it is clear that he assumed that the parties of the original position knew and relied on their general knowledge of human psychology in choosing what they believed to be the most normatively compelling principles of justice to regulate the basic structure of their society. Presumably, relying on different psychological theories to guide one's deliberation would have different effects on what principles of justice the original contracting parties would choose. I believe this aspect of the original position has not been sufficiently discussed in the literature as most of the literature so far has instead focused on what is now known as the famous Rawls-vs-Harsanyi debate (Rawls [1999, 1974]; Harsanyi [1953, 1955, 1975, 1977]; see Gaus and Thrasher [2015] for an excellent survey of the Rawls-vs-Harsanyi debate; Moehler [2015] argues that there is no clear winner of the Rawls-vs-Harsanyi debate as each author attempts to model different moral ideals). This paper invites the reader to consider what would happen if the parties of the original position took Kahneman and Tversky's *Prospect Theory* (Kahneman and Tversky 1979, 1992) as constituting their general knowledge of human psychology and based their subsequent choice of principles of justice on such knowledge. I will try to show that once the parties of the original position take prospect theory as a part of their general knowledge of human psychology and characterize each individual's utility function accordingly, the choice of utilitarianism becomes far less implausible than Rawls had initially argued.

2. Rawls's Criticism of Utilitarianism

As briefly mentioned, one of Rawls's major aims was to present a systematic theory of distributive justice that was superior to utilitarianism in its various forms. The main reason for doing this was that Rawls believed that utilitarianism, regardless of its many variations, is problematic at the most fundamental level. Rawls's

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

criticisms against various forms of utilitarianism and the reasons why he thinks his justice as fairness is superior to utilitarianism are spread throughout his *A Theory of Justice* (1999); he mentions issues related to the distinction between separate persons (section 5), strains of commitment (section 29), stability and the publicity condition (section 29), the social basis for self-respect (section 29), and so on. All of these are distinct considerations but, nonetheless, stem from what is perceived to be a characteristic feature that all forms of utilitarianism share, namely, that utilitarianism can result in vastly unequal distributions that would, for the purpose of maximizing aggregate or average social welfare, sacrifice society's lesser advantaged groups and put them far below what they could reasonably expect to achieve under different distributional arrangements. The traditional objection that utilitarianism may justify the institution of slavery derives from exactly the same concerns (see Rawls 1999: 135). In other words, the main problem with utilitarianism has to do with its potential to generate extreme inequalities that may sacrifice the lesser advantaged for the sake of benefiting society as a whole.

3. Social Welfare Functions and Individual Utility Functions

Compared to justice as fairness, how much inequality does utilitarianism really generate? Let us try to understand this more precisely. Let $N = \{1, \ldots, n\}$ be the set of n individuals who are members of a given society, and let X be the set of all feasible social alternatives. Let $u_i{:}X \to \mathbb{R}$ be individual i's cardinal utility function representing his/her welfare level. Given social alternative $x \in X$, $u_i(x)$ denotes individual i's welfare level when social alternative x is realized. A social welfare function $W{:}X \to \mathbb{R}$ represents the social preferences over the different social alternatives in X such that social alternative $x \in X$ is socially preferred to social alternative $y \in X$ if and only if $W(x) > W(y)$.

From this, we may define the utilitarian social welfare function $U{:}X \to \mathbb{R}$ as follows:

$$U(x) = \sum_{i=1}^{n} u_i(x).$$

According to the utilitarian social welfare function, social alternative $x \in X$ is socially preferred to social alternative $y \in X$ if and only if $U(x) > U(y)$ if and only if $\sum_{i=1}^{n} u_i(x) > \sum_{i=1}^{n} u_i(y)$. That is, according to the utilitarian social welfare function, social alternative $x \in X$ is socially preferred to social alternative $y \in X$ if and only if the sum total of individual utility that is generated by social alternative x is greater than that generated by alternative y.

We may define the Rawlsian social welfare function $R{:}X \to \mathbb{R}$ as follows:

$$R(x) = \min\{u_1(x), \ldots, u_n(x)\}.$$

According to the Rawlsian social welfare function, social alternative $x \in X$ is socially preferred to social alternative $y \in X$ if and only if $R(x) > R(y)$ if and only if $\min\{u_1(x),$

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

$\ldots, u_n(x)\} > \min\{u_1(y), \ldots, u_n(y)\}$. That is, according to the Rawlsian social welfare function, social alternative $x \in X$ is socially preferred to social alternative $y \in X$ if and only if the welfare level of the individual who obtains the lowest welfare level under social alternative x is greater than that obtained under social alternative y.

With these social welfare functions, we may now formally characterize what social alternatives utilitarianism and the difference principle would each respectively recommend for a given distribution problem. Given the set of social alternatives X, the social alternative(s) that would be prescribed by utilitarianism would be the solutions to the following maximization problem:

$$\max_{x \in X} \sum_{i=1}^{n} u_i(x).$$

We can think of this expression as a formal characterization of utilitarianism. Similarly, given the set of social alternatives X, the social alternative that would be prescribed by the difference principle would be the solutions to the following maximization problem:

$$\max_{x \in X} \min \{u_1(x), \ \ldots, u_n(x)\}.$$

We can think of this expression as a formal characterization of Rawls's difference principle applied to people's welfare levels. Of course, strictly speaking, Rawls intended his difference principle to be applied to the index of primary social goods (Rawls 1999: sec. 15) rather than to people's welfare levels. There are four main reasons why, in this paper, I am applying Rawls's difference principle to people's welfare levels instead of to the index of primary social goods as Rawls had initially proposed.

The first reason relates to what is known as *the index problem*. Primary social goods, according to Rawls, are the kind of all-purpose means that any rational person would want and, if possible, want more of rather than less, regardless of his/her specific ends. The primary social goods include 'rights, liberties, and opportunities, and income and wealth' (Rawls 1999: 79). Rawls's basic thought was that we could assign numbers to different bundles of primary social goods in such a way that bundles of primary social goods that were assigned higher numbers are universally preferred by everybody regardless of his/her particular aims. The question is whether such indexing of primary social goods is practically possible.

The short answer is that it would be virtually impossible once we recognize a multitude of primary social goods and assume that people's aims are sufficiently different. For instance, let (x, y) denote a bundle of primary social goods, where x denotes the amount of money and y denotes the amount of freedom. Consider two bundles of primary social goods: $A = (10, 5)$ and $B = (5, 10)$. Suppose that John values money more than his freedom, while Mary values her freedom more than money. John would prefer bundle A to bundle B, while Mary would prefer bundle B to bundle A. In such cases, it is impossible to assign numbers to the two

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

bundles of primary social goods in such a way that a bundle that gets assigned a higher number is universally preferred by both individuals. Assigning a higher number to *A* would be inconsistent with Mary's preferences, while assigning a higher number to *B* would be inconsistent with John's preferences. Assigning the same number to both *A* and *B* would be inconsistent with both person's preferences. According to John Roemer, the index problem cannot be solved unless we assume that there is a single primary social good, say, money (Roemer, n.d.). This would be inconsistent with Rawls's presumption that there are multitudes of primary social goods such as rights, liberties, and opportunities and income and wealth.

A related problem with the primary social goods approach is what Amartya Sen (1980) views as its potential 'resource fetishism'. When Rawls presented the notion of primary social goods, his intention was to use the index of primary social goods as a simplified measure of advantage: one person is better off relative to another person if and only if s/he enjoys a greater index of primary social goods than the other person. However, according to Sen, this ignores that two bundles of primary social goods that have the same index could very well generate different values for different people. As Sen writes:

> The primary goods approach seems to take little note of the diversity of human beings. . . . If people were basically very similar, then an index of primary goods might be quite a good way of judging advantage. But, in fact, people seem to have very different needs varying with health, longevity, climatic conditions, location, work conditions, temperament, and even body size (affecting food and clothing requirements). . . . Judging advantage purely in terms of primary goods leads to a partially blind morality. Indeed, it can be argued that there is, in fact, an element of 'fetishism' in the Rawlsian framework. Rawls takes primary goods as the embodiment of advantage, rather than taking advantage to be a relationship between persons and goods. (Sen 1980: 215–16)

The second reason for reinterpreting Rawls's difference principle in terms of people's welfare levels is to avoid this kind of resource fetishism.

The third reason for reinterpreting Rawls's difference principle in terms of people's welfare levels relates to why Rawls had initially proposed using primary social goods as a measure of advantage in the first place, which was to find a more objective basis for making interpersonal comparisons than utility or welfare. Note that the applications of the difference principle as well as of utilitarianism both require making interpersonal comparisons. However, Rawls thought that making interpersonal comparisons in terms of people's welfare levels was problematic for various theoretical reasons (Rawls 1999: 282–85), and, hence, he proposed to use the index of primary social goods as an alternate way to make interpersonal comparisons in an objective way. However, utilitarianism, by its very definition, aggregates people's utility/welfare, which presumes that utility/welfare is interpersonally comparable. If one wishes to compare the distributional consequences of utilitarianism to that of the difference principle, one has no

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

choice but to assume that interpersonal comparisons of utility/welfare can be adequately performed. Otherwise, the debate between utilitarianism and justice as fairness cannot even start.

Actually, the type of interpersonal comparison that is minimally required for utilitarianism to make theoretic sense is not that demanding as it might at first appear. It requires what is known as *unit comparability*—namely, that people's welfare gains and losses can be compared in the same unit. In contrast, the difference principle requires *level comparability*—namely, that two people who are at the same level of advantage can be deemed to be equally well-off. There is no clear sense in which one informational requirement is more stringent than the other: assuming that people's welfare gains and losses can be compared in the same units does not imply that their levels can be compared; conversely, assuming that people's advantage levels can be compared does not imply that their welfare gains and losses can be compared in the same units. As a matter of fact, when comparing the distributional consequences of utilitarianism and the difference principle, Rawls, for illustrative purposes, frequently invokes utility functions that are intended to be comparable both in terms of unit and level. When doing so, Rawls explains that 'justice as fairness does not deny that the idea of a utility function can be used to formulate justice as fairness' (Rawls 2001: 107), which suggests that Rawls was open to reinterpreting his difference principle from a welfarist framework. The point is that if Rawls himself is willing to grant that people's welfare or utility can be interpersonally comparable in a way that both utilitarianism and the difference principle make theoretical sense, then the very reason why Rawls had initially proposed to use the index of primary social goods as a basis for interpersonal comparisons vanishes.

Lastly, I have previously shown, formally, that the difference principle, once it is applied to primary social goods, performs counterintuitively and offers decisive reasons for the parties in the original position to choose utilitarianism instead of justice as fairness under Rawls's own assumptions (Chung 2020). There, I have explained that many of these counterintuitive results may be overturned if the difference principle is instead applied to people's welfare levels (Chung 2020: sect. 8). For these reasons, let us proceed by reinterpreting Rawls from a welfarist framework and apply the difference principle to people's welfare levels.

When examining the distributional consequences of utilitarianism and Rawls's difference principle, the following toy examples will be quite illustrative.

Example 1 Let us consider a simple society with just two individuals: $N = \{1, 2\}$. Suppose there is a fixed amount of social resources (say, 10 units) that can be distributed to each individual. Let x_1 be the amount distributed to individual 1, and let x_2 be the amount distributed to individual 2. The set of social alternatives, then, is the set of all possible distributions of 10 units of resources to the two individuals: $X = \{(x_1, \ x_2) \in \mathbb{R}^2 | \ \ x_1 \geq 0, \ \ x_2 \geq 0, \ \ x_1 + x_2 \leq 10\}$. Figure 1 represents set X, the set of feasible social alternatives, graphically. Each point on the plane represents a given distribution of resources to the two individuals 1 and 2. In figure 1, all the

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

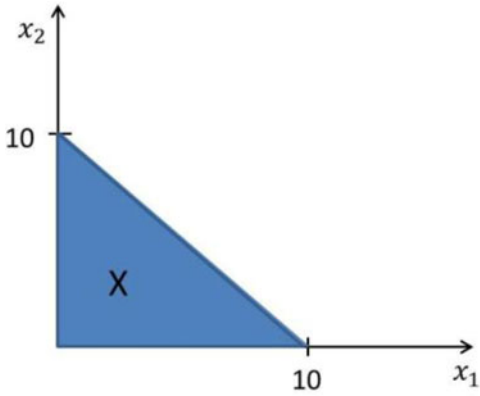

Figure 1. The set of feasible social alternatives.

social distributions represented by points inside the shaded triangle are those that are socially feasible.

Suppose $u_1(x_1) = x_1$ and $u_2(x_2) = x_2$. That is, suppose individual welfare levels increase linearly in the amount of assets s/he enjoys and that the same amount of assets generates the same welfare level for each individual. Figure 2 represents the shape of each individual's utility function when it is linear in assets:

Under these conditions, what distributions would the difference principle and utilitarianism, respectively, recommend? Figure 3 shows the indifference curves of the Rawlsian social welfare function (SWF). Among the socially feasible distributions, the difference principle will choose a distribution that would put society on the highest indifference curve of the Rawlsian social welfare function. We can find this by overlapping figure 3 on top of figure 1. From this, we can verify that the difference principle will choose (5, 5), a perfectly equal distribution of the social resources. (See Figure 4.)

Let us consider what type of distributions utilitarianism will choose. With linear individual utility functions, the indifference curves of the utilitarian social welfare function will be straight lines as depicted in Figure 5.

Again, the social distributions that utilitarianism will recommend will be those within the set of socially feasible distributions that put society on the highest indifference curve of the utilitarian social welfare function. We can find such distributions by overlapping figure 5 on top of figure 1. (See Figure 6.)

We can verify that all the social distributions on the hypotenuse of the triangular region formed by the set of socially feasible alternatives are compatible with utilitarianism. Let S denote the set of social distributions that are compatible with utilitarianism. Here, $S = \{(x_1, 10 - x_1) \in \mathbb{R}^2 | x_1 \in [0, 10]\}$. Note that the distributions (10, 0) and (0, 10) (i.e., distributions in which one person gets the entire social wealth) are both in S. However, also note that (5, 5) (i.e., a perfectly equal distribution) is also in S. The problem here is not that utilitarianism will

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

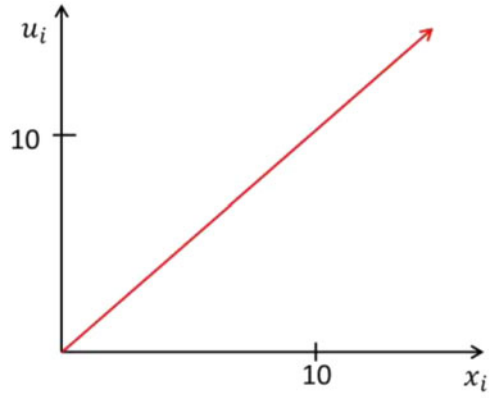

Figure 2. Linear individual utility function.

necessarily generate vastly unequal distributions (it might very well generate a perfectly equal distribution), but rather that utilitarianism is *compatible with* vastly unequal distributions and may possibly generate them. Thus, we can at least say that Rawls's worry about the distributional consequences of utilitarianism is valid when individual utility functions are linear.

Example 2 Consider the same example, but now suppose individual utility functions are *strictly concave*. Specifically, suppose $u_1(x) = \ln x_1$ and $u_2(x_2) = \ln x_2$. For any given amount of assets, each individual receives a welfare level equivalent to the natural log of that wealth. Figure 7 shows the general shape of our logarithmic (strictly concave) individual utility function.

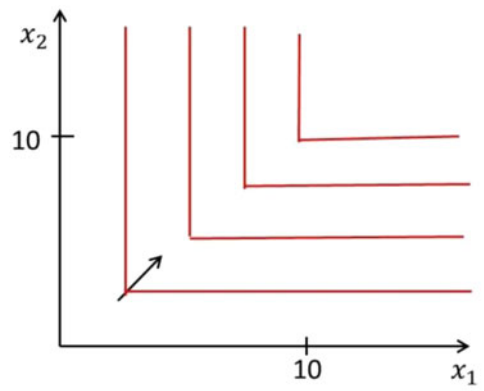

Figure 3. Indifference curves of Rawlsian SWF.

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

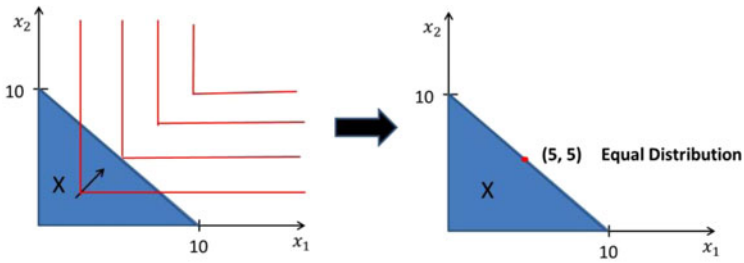

Figure 4. Indifference curves of Rawlsian SWF on top of the set of feasible social alternative, Social alternative chosen by Rawlsian SWF.

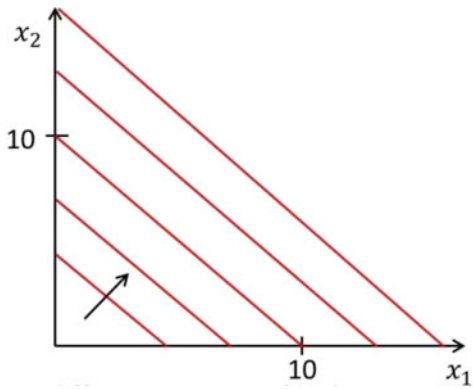

Figure 5. Indifference curves of utilitarian SWF (when individual utility functions are linear).

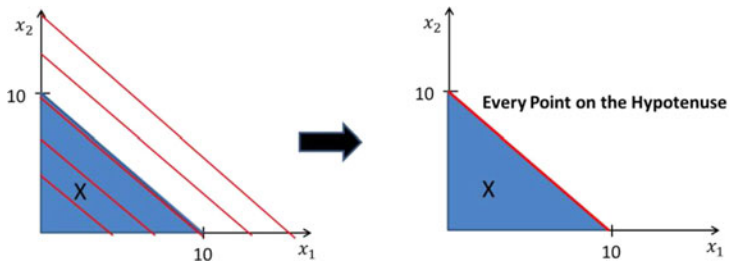

Figure 6. Indifference curves of utilitarian SWF on top of set of feasible social alternatives, Social alternatives chosen by utilitarian SWF (when utility functions are linear).

One may think of this as a situation in which the law of diminishing marginal utility holds. Here, nothing really changes the shape of the indifference curves of the Rawlsian social welfare function; as a result, (5, 5) is the unique distribution that will be chosen by the difference principle as before.

However, when individual utility functions become strictly concave (or display the law of diminishing marginal utility) an interesting change happens to the

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

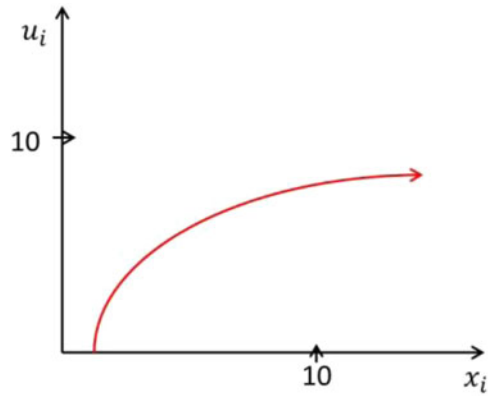

Figure 7. Logarithmic (strictly concave) individual utility function.

general shape of the indifference curves of the utilitarian social welfare function. The indifference curves of the utilitarian social welfare function start to bend and become convex to the origin. Figure 8 depicts the indifference curves of the utilitarian social welfare function when individual utility functions are strictly concave.

We can find the social distributions chosen by utilitarianism as before by overlapping figure 8 on top of figure 1. (See Figure 9.) Unlike what has happened when individual utility functions were linear, the only social distribution of assets that is compatible with utilitarianism, when individual utility functions are strictly concave, is now (5, 5), a perfectly equal distribution!

The main point that I want to emphasize with these toy examples is the following: The specific distribution of assets or resources that a given theory of distributive justice prescribes depends not simply on the distributional principle (or its corresponding social welfare function), but also on the specific characterization or shape of each individual's utility function. Utilitarianism becomes more inequality-averse (resp.

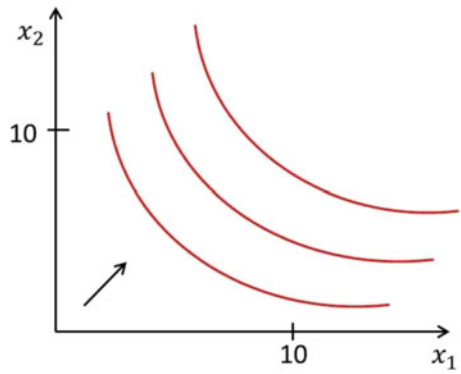

Figure 8. Indifference curves of utilitarian SWF (when individual utility functions are strictly concave).

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

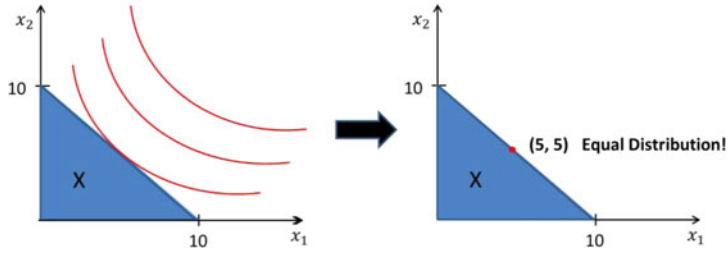

Figure 9. Indifference curves of utilitarian SWF on top of set of feasible social alternatives, Social alternatives chosen by utilitarian SWF (when individual utility functions are strictly concave).

inequality-prone) as individual utility functions become more concave (resp. convex.) And the specific characterizations of each individual's utility function depend on what general theory of human psychology we adopt—for instance, if we believe that people's psychology is subject to the law of diminishing marginal utility, then their utility functions will have a concave shape. If we believe that people are generally risk-neutral, then their utility functions will be linear; if we believe that people generally enjoy risky prospects, then their utility functions will be convex (i.e., they will display an increasing slope). This implies that even if a society adopts utilitarianism, the distributional consequence that this will entail is underdetermined unless we properly specify the utility functions of its individuals by our general theory of human psychology. Depending on which psychological theory we adopt, utilitarianism may or may not result in extreme inequalities. If it does not do so, then Rawls's criticisms against utilitarianism and his arguments for justice as fairness become far less forceful.

4. Kahneman and Tversky's Prospect Theory

As briefly explained in section 1, the veil of ignorance, despite depriving the original contracting parties of morally irrelevant, contingent information about themselves, allows them to retain general knowledge related to society, politics, economics, and human psychology. That the parties of the original position retain such general knowledge is not a mere placeholder. As a matter of fact, one of Rawls's arguments against classical utilitarianism is that classical utilitarianism, in effect, requires people to be perfect altruists, and this, according to Rawls, is impossible from what we know about the laws of general human psychology (Rawls 1999: 164–65).

Suppose that the parties in the original position took Kahneman and Tversky's *Prospect Theory* (1979, 1992), which represents one of the most prominent theoretical developments made in the area of decision under risk, as constituting their general knowledge of human psychology. How would this influence their resulting choice of principles of justice in the original position?

To answer this question, let us first try to understand prospect theory. It has long been shown by experimental evidence that when faced with risky prospects, people do not behave in ways that conform to the standard axioms of

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

expected utility theory (see Allais 1953; Ellsberg 1961). What Kahneman and Tversky did was to propose an alternative theory of choice under risk that was designed to explain these seemingly anomalous behaviors better and to provide a more adequate descriptive model of choice under risk. The result is what they call prospect theory.

The main point of departure of prospect theory from standard expected utility theory is that instead of assuming that decision makers assign values (or utilities) to final assets, prospect theory assumes that decision makers assign values (or utilities) to changes (i.e., gains and losses) assessed from a perceived reference point. One psychological effect by which Kahneman and Tversky have found people to be influenced is what they call the *certainty effect*. Based on the results of many laboratory experiments, Kahneman and Tversky have found that people tend to *overweight* outcomes they perceive to be certain, relative to outcomes they perceive to be merely probable (Kahneman and Tversky 1979: 265). Not only does certainty increase the desirability of gains, but it also increases the averseness of losses. This phenomenon of overweighting outcomes that are certain has the mirror effect of making people risk-averse toward gains, while at the same time making people risk-seeking toward losses. Note that the fact that people are risk-seeking toward losses does not mean that people enjoy experiencing losses. Rather, it implies the exact opposite; that is, it implies that people hate certain losses so much that they are willing to take a gamble that has a greater expected loss but that gives them some chance of experiencing no loss.

It is generally understood in the theory of rational choice that one's attitude toward risk is reflected in the curvature of one's utility function; risk-averseness is represented by concavity, and risk-lovingness is represented by convexity of utility functions. This means that the phenomenon of overweighting certainty, which has an effect of making people risk-averse toward gains and risk-seeking toward losses, will be reflected in the utility function (or what Kahneman and Tversky call the value function) being concave above the given reference point and convex below the given reference point. Furthermore, Kahneman and Tversky have also discovered that 'the aggravation that one experiences in losing a sum of money appears to be greater than the pleasure associated with gaining the same amount' which implies that 'the value function for losses is steeper than the value function for gains' (Kahneman and Tversky 1979: 279).

Putting all of these components together, we now arrive at a general characterization of individual utility functions based on Kahneman and Tversky's prospect theory:

> In summary, we have proposed that the [utility function] is (i) defined on deviations from the reference point; (ii) generally concave for gains and commonly convex for losses; (iii) steeper for losses than for gains. (Kahneman and Tversky 1979: 279)

Figure 10 is a picture of a utility function that meets all of these three characteristics and gives us a sense of what individual utility functions would generally look like according to prospect theory.

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

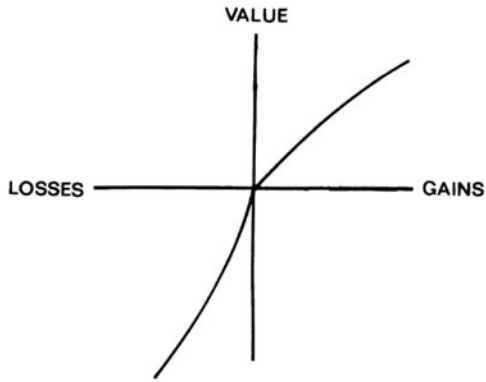

Figure 10. Shape of utility function according to prospect theory (Kahneman and Tversky 1979: 279).

5. Prospect Utilitarianism and the Difference Principle: A Comparison

Let us return to our original discussion. As we have seen, how much inequality utilitarianism allows depends on the specific characterization of individual utility functions; utilitarianism becomes more prone toward an equal distribution as individual utility functions become more concave. According to prospect theory, when faced with risk, individual utility functions generally take a particular shape; they are convex below a reference point, concave above that reference point, and the curve's slope is steeper below the reference point than it is above that point.

Suppose that the parties in the original position took prospect theory as a part of the general knowledge of human psychology that survives through the veil of ignorance and used their knowledge of prospect theory to characterize individual utility functions for the purpose of comparing the distributional consequences of different distributive principles. Of course, there still remains a great deal of controversy over which decision theory—between von Neumann and Morgenstern's standard expected utility theory and Kahneman and Tversky's prospect theory—is, on the whole, most descriptively accurate. The purpose of this paper is not to settle this issue. Rather, the main aim of this paper is to examine what distributional implications both utilitarianism and the difference principle (applied to welfare levels) would have if the parties in the original position were to use prospect theory as their general theory of human psychology to characterize individual utility functions and to see how this would affect the selection of principles of justice in the original position.

5.1 The Model

Having this in mind, let us now examine the different distributional consequences of utilitarianism and the difference principle when individual utility functions are characterized by prospect theory. Before we do this, it should be noted that

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

Kahneman and Tversky have never intended their prospect theory to be a normative theory (Kahneman and Tversky 1979: 277; 1992: 317). Instead, their main aim was to provide what they believed to be a more descriptively accurate model of individual choice under risk that could better explain the type of behaviors that were regarded as failures of standard assumptions of rationality.

By contrast, the original position is primarily a normative device—it tries to figure out what principles of justice can be normatively justified from the choices made by fully rational agents behind the veil of ignorance. It is important to understand that assuming that the parties in the original position take prospect theory as a part of their general knowledge of human psychology, which they use to characterize individual utility functions, does *not* mean that we are regarding the parties in the original position to be irrational in ways that prospect theory assumes. Prospect theory is not a *decision procedure* on which the original contracting parties rely to derive normatively compelling principles of justice—for instance, the parties in the original position do not go through separate editing and evaluation phases, nor are they influenced by the certainty effect or other kinds of framing effects that prospect theory posits.

Rather, what I am assuming is that the parties in the original position, while being fully rational, will use prospect theory as *background knowledge* that helps them characterize individual utility functions for the normative purpose of comparing and eventually choosing among different principles of justice as regulative for the basic structure of their society. In other words, the parties in the original position are fully rational agents but know (from their knowledge of prospect theory) that their real-world clients are not as fully rational as they are, and hence, they try to choose the best principles of justice that will regulate the basic structure of a society composed of less than perfectly rational individuals whose individual utility functions are shaped in the way suggested by prospect theory.

We will follow the basic formal setup of my 'Rawls's Self-Defeat: A Formal Analysis' (Chung 2020; see also Roemer 2002 and Moreno-Ternero and Roemer 2008 for alternate models of the original position.) and consider a hypothetical society in which the first principle of justice (i.e., the principle of equal basic liberties) is formally satisfied by its constitution. Just like Rawls, we assume that our model society consists of two representative groups (Rawls 1999: 66–67; 2001: 62–63): the more advantaged group (MAG) and the less advantaged group (LAG). Following Rawls, we assume that the members of MAG and LAG both have 'physical needs and psychological capacities within the normal range' (Rawls 1999: 83–84) that allow them to be fully participating members of mutually beneficial social cooperation. An important implication of this assumption is that there are neither extraordinarily efficient utility-producing machines nor extremely poor translators of wealth-to-welfare 'so that the questions of health care and mental capacity do not arise' (Rawls 1999: 83–84). However, let us assume that, within the normal range, the members of LAG are disadvantaged in their overall natural abilities and/or social circumstances relative to members of MAG.

Let $u_M{:}\mathbb{R}_+ \to \mathbb{R}$ be the utility function of MAG, and let $u_L{:}\mathbb{R}_+ \to \mathbb{R}$ be the utility function of LAG, which conform to the general characteristics of individual utility functions proposed by Kahneman and Tversky's prospect theory. Accordingly, we

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

would need to specify the reference points of MAG's and LAG's utility functions. How should we do this? In Kahneman and Tversky's prospect theory, each person's reference point denotes what that person perceives to be the status quo—that is, his/her current or expected asset level. As the veil of ignorance deprives the parties of knowing their starting positions in society, it deprives the contracting parties of this information. Hence, the reference points would have to be assigned by the fully rational original contracting parties themselves. What would then be the most normatively appealing way to assign each representative group's reference point that would best serve the normative purpose of the original position?

It seems there are broadly two possibilities. The first is to assign equal distribution of the initial stock of the noncooperatively available resources that our society starts with before MAG and LAG fully engage in mutually beneficial productive cooperation as each group's baseline reference point. In describing what particular baseline it would be reasonable for the parties in the original position to assume, Rawls explains that 'since it is not reasonable for [the parties] to expect more than an equal share in the division of social primary goods, and since it is not rational for [them] to agree to less, the sensible thing is to acknowledge as the first step a principle of justice requiring an equal distribution' (Rawls 1999: 130). This is the textual ground for assigning equal distribution of the initially available resources as each representative group's reference point.

Alternatively, the parties in the original position may instead assign as their reference points the amount of resources each representative group needs to enjoy fully the *fair worth* of the basic rights and liberties constitutionally guaranteed by the first principle of justice. Recall that one of Rawls's main reasons for proposing the difference principle in addition to the liberty principle and the principle of fair equal opportunity was to secure best the fair worth of the equal basic rights and liberties guaranteed by the first liberty principle. For example, in our model society, in which the first principle of justice is formally satisfied by the constitution, everybody will have the right to freely choose his/her occupation. However, the worth of this formal right will be different for different individuals; for instance, having a formal right to choose one's occupation freely will have *less value* to a person who is too poor to afford, say, a college education than to a person who is born in a wealthy family that can afford the best private education in the nation. One important purpose of the difference principle was to distribute social primary goods in a way that would allow the least advantaged members of society to enjoy the best worth of their basic rights and liberties guaranteed by the first principle of justice. As Rawls writes,

> Freedom as equal liberty is the same for all; the question of compensating for a lesser than equal liberty does not arise. But *the worth* of liberty is not the same for everyone. Some have greater authority and wealth, and therefore greater means to achieve their aims. The lesser worth of liberty is, however, compensated for . . . . Taking the two principles together, the basic structure is to be arranged to maximize *the worth* to the least advantaged of the complete scheme of equal liberty shared by all. (Rawls 1999: 179, emphasis added)

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

Such is the textual ground for assigning the amount of resources each group needs to enjoy fully the fair worth of their basic rights and liberties constitutionally guaranteed by the first principle of justice as each representative group's reference point.

Let $W_\circ > 0$ be our society's initially available resources prior to social cooperation and let $r_\circ = \frac{W_\circ}{2}$ denote the amount of resources the two representative groups, MAG and LAG, can each expect to receive under equal division of the initially available default social resources. Let $r_M \in \mathbb{R}_+$ (resp. $r_L \in \mathbb{R}_+$) denote the amount of resources required for the members of MAG (resp. LAG) to enjoy fully the fair worth of their basic rights and liberties. We assume $0 < r_\circ = \frac{W_\circ}{2} < W_\circ < r_M < 2r_M < r_L$, which basically indicates three things.

First, the fact that both r_M and r_L are greater than $r_\circ = \frac{W_\circ}{2}$ (i.e., $r_\circ = \frac{W_\circ}{2} < r_M,\ r_L$) means that to enjoy the fair worth of their basic rights and liberties fully, both MAG and LAG require *more* resources than what they would expect to receive under equal division of the initially available noncooperative default social resources.

Second, the assumption that $W_\circ < r_M, r_L$ implies that mutually beneficial social cooperation between MAG and LAG is *necessary* for any group to enjoy fully the fair worth of their basic rights and liberties. Together, the two assumptions are intended to reflect Rawls's condition of 'moderate scarcity', which means that in our model society, 'natural and other resources are not so abundant that schemes of cooperation become superfluous, nor are conditions so harsh that fruitful ventures must inevitably break down' (Rawls 1999: 110). If we had $r_M,\ r_L \leq r_\circ = \frac{W_\circ}{2} < W_\circ$, that is, if the amount of resources required for both MAG and LAG to enjoy fully the fair worth of their basic rights and liberties was less than or equal to what each group could expect to receive under equal division of the initial stock of the noncooperative default social resources, then there would be little reason for the two groups to engage in mutually beneficial social cooperation in the first place. Under such resource situations, it would be possible for both groups to enjoy fully the fair worth of their basic rights and liberties from the initial stock of social resources without engaging in any sort of mutually beneficial social cooperation whatsoever, which would render schemes of cooperation 'superfluous'. In this sense, we may think of the condition $r_\circ = \frac{W_\circ}{2} < W_\circ < r_M,\ r_L$ as reflecting what Rawls calls the 'objective circumstances of justice' (Rawls 1999: 109–10).

Third, the fact that $r_M < 2r_M < r_L$ implies that the members of LAG require more resources (in particular, more than 2 times of r_M) to enjoy the fair worth of their basic rights and liberties fully, which the members of MAG can fully enjoy at a lower resource level of r_M. Such an assumption is how our model tries to represent that the members of LAG, while within the normal range, are relatively disadvantaged in their overall natural abilities and/or social circumstances relative to the members of MAG. For instance, if compared to the members of MAG the members of LAG come from families with relatively poor educational backgrounds and reside in regional districts where the public education system is not so great, then they might require additional financial resources (that may, perhaps, be used to subsidize additional training and education for the members of LAG) to enjoy fully the fair worth of the various rights and opportunities that

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

are formally guaranteed by the constitution. Such is the intuition behind our assumption that $r_M < 2r_M < r_L$.

Let $\underline{r_M} \in \mathbb{R}_+$ denote the reference point of MAG and let $\underline{r_L} \in \mathbb{R}_+$ denote the reference point of LAG. We assume that the utility function of MAG, $u_M(x)$ and the utility function of LAG, $u_L(x)$ satisfy the following properties:

1. Utility is both *level* and *unit* comparable. Formally, let X be the set of social alternatives and let $U^1 = (u_M^1,\ u_L^1)$ and $U^2 = (u_M^2,\ u_L^2)$ be any two profiles of utility functions of MAG and LAG. We say that utility/welfare is both level and unit comparable if and only if the social ordering on X induced by U^1 and U^2 are the same whenever there exists an $\alpha > 0$ and a $\beta \in \mathbb{R}$ such that for all $x \in X$, $u_M^2(x) = \alpha u_M^1(x) + \beta$ and $u_M^2(x) = \alpha u_M^1(x) + \beta$. (For more on informational constraints on utility functions, see Roemer [1996: sect. 1.1] and Bossert and Weymark [2004]).
2. $u_M(x)$ is twice differentiable in the left and right regions of its reference point, i.e., $u_M(x)$ is twice differentiable for all $x \in \mathbb{R}_+ \backslash \{\underline{r_M}\}$.
3. For all $x \in \mathbb{R}_+ \backslash \{\underline{r_M}\}$, $u'_M(x) > 0$, i.e., u_M is strictly increasing in resources.
4. For all x, x' such that $0 \leq x < \underline{r_M} < x'$, $u''_M(x) > 0$, $u''_M(x') < 0$, and $u'_M(x) > u'_M(x')$, i.e., u_M is strictly convex below, strictly concave above its reference point, and the slope of u_M is steeper below its reference point than above.
5. For simplicity, we will assume that u_L is a horizontal translation of u_M. Specifically, for all $x \in \mathbb{R}_+$, define $u_L(x) = u_M(x - (\underline{r_L} - \underline{r_M}))$. If we assume $\underline{r_M} = \underline{r_L} = r_0 = \frac{W_0}{2}$ (i.e., if we assume equal division of the initially available resources to be each group's reference point), then we have $u_L(x) = u_M(x)$ (i.e., both MAG and LAG share the same utility functions). If we assume $\underline{r_M} = r_M < r_L = \underline{r_L}$ (i.e., if we assume each group's reference point as the amount of resources each group needs to enjoy fully the fair worth of their basic rights and liberties), then we have $u_L(x) = u_M(x - (r_L - r_M))$, i.e., u_L is obtained by horizontally shifting u_M an increment of $(r_L - r_M)$ to the right.

Property 1 is assumed to render both utilitarianism and Rawls's difference principle applicable under a welfarist framework. As already explained, utilitarianism requires unit comparability, while Rawls's difference principle (applied to people's utility/welfare levels) requires level comparability. Hence, when utility is both unit and level comparable, both utilitarianism and Rawls's difference principle applied to people's utility/welfare levels make theoretical sense. Properties 2, 3, and 4 are imposed to make both u_M and u_L display the general characteristics of individual utility functions described by prospect theory with respect to each group's reference points $\underline{r_M}$ and $\underline{r_L}$. The differentiability assumptions are added to allow us to use calculus techniques to identify solutions to subsequent distributional problems. Property 5 is, strictly speaking, not necessary and can be significantly weakened: specifically, it can be weakened to

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

$\forall x > 0$, $\min\{u'_M(r_M - x),\ u'_L(r_L - x)\} > \max\{u'_M(r_M + x),\ u'_L(r_L + x)\}$. Property 5 is included mainly for the purpose of simplifying the proofs without losing too much generality. Nonetheless, one may derive the property's justification from Rawls's assumption that both social groups have 'physical needs and psychological capacities within the normal range'—which in this case means that they enjoy similar (or not too divergent) welfare levels for any incremental asset gains (or losses) above (or below) their reference points.

5.2 Results of the Model

Let us now examine and compare the distributional prescriptions of utilitarianism and Rawls's difference principle inside our model under different degrees of social affluency. As already noted, from the condition of moderate scarcity (which, in our model, is represented by the condition $r_0 = \frac{W_0}{2} < W_0 < r_M,\ r_L$) in order for at least one representative group (either MAG or LAG or both) to enjoy the fair worth of their basic rights and liberties fully, the two groups must mutually cooperate to produce social surplus.

Let $\bar{W}$ denote the amount of social surplus that MAG and LAG have jointly produced by their mutual social cooperation, which in our model is assumed to be determined exogenously. We assume that $\bar{W} \geq W_0$—that is, we assume that MAG and LAG's social cooperation does not, at the very least, result in destroying the initial stock of resources that their society starts with. Hence, all the resource situations that we will subsequently analyze and discuss will be those that do not fall below the condition of moderate scarcity. We then consider how utilitarianism and Rawls's difference principle distribute the cooperative surplus $\bar{W}$ to the two representative groups, MAG and LAG. For the remaining discussion, let (x_M, x_L) denote a distribution in which MAG gets x_M and LAG gets x_L amount of assets. All proofs are relegated to the appendix.

Proposition 1: Suppose $\underline{r_M} = \underline{r_L} = r_0 = \frac{W_0}{2}$. Then, for all $\bar{W} \geq W_0$, the distributional prescriptions of utilitarianism and Rawls's difference principle are the same—namely, $(x_M, x_L) = \left(\frac{\bar{W}}{2}, \frac{\bar{W}}{2}\right)$.

Proposition 1 claims that if the parties in the original position assign equal division of the initial stock of noncooperative default social resources as each representative group's reference point (i.e., $\underline{r_M} = \underline{r_L} = r_0 = \frac{W_0}{2}$), then both utilitarianism and Rawls's difference principle will completely coincide in their distributional prescriptions: specifically, both utilitarianism and Rawls's difference principle will divide the cooperative surplus $\bar{W}$ into half (viz., $\frac{\bar{W}}{2}$) and distribute it equally to MAG and LAG (viz., $(x_M, x_L) = \left(\frac{\bar{W}}{2}, \frac{\bar{W}}{2}\right)$). Given that the social surplus jointly produced by MAG's and LAG's mutual social cooperation is positive (i.e., $\bar{W} > W_0$), this implies that, under both distributional principles, both MAG and LAG are guaranteed to secure an amount of resources that exceeds what each group would expect to receive under equal division of the initial stock of noncooperatively available default social resources, which constitutes each group's reference point (viz., $(x_M, x_L) = \left(\frac{\bar{W}}{2}, \frac{\bar{W}}{2}\right) > \left(\frac{W_0}{2}, \frac{W_0}{2}\right) = (\underline{r_M}, \underline{r_L})$).

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

Under both distributive principles, the resulting distribution is equal, and both representative groups end up securing an amount of resources that meet or exceed their reference points—so far so good. However, we cannot say that the resulting distribution is *ideal* in all relevant aspects. First of all, dividing the social surplus into half and distributing it equally to each group does not attend to the fact that members of LAG are relatively disadvantaged compared to the members of MAG and would require more resources to enjoy fully the fair worth of their basic rights and liberties than what the members of MAG could fully enjoy at a lower resource level. Second, when $\underline{r_M} = \underline{r_L} = r_\circ = \frac{W_\circ}{2} < r_M < \frac{\bar{W}}{2} < r_L$, then an equal distribution allows *only* the members of MAG and *not* the members of LAG to enjoy the fair worth of their basic rights and liberties fully.

However, this is not a defect of either distributive principle; rather, the defect stems from assigning equal division of the initial stock of noncooperative default social resources as each representative group's reference point. What is relevant for our current purpose (which is to compare the distributional consequences of utilitarianism and Rawls's difference principle) is to understand that when the parties in the original position assign equal division of the initial stock of noncooperative default social resources as each representative group's reference point, the distributional prescriptions of both utilitarianism and Rawls's difference principle completely coincide, and therefore, the parties have no reason to prefer one distributional principle over the other.

Suppose that the parties in the original position, instead, assume each group's reference point as the amount of resources each group needs to enjoy fully the fair worth of their basic rights and liberties—that is, $\underline{r_M} = r_M < 2r_M < r_L = \underline{r_L}$. Would the distributional prescriptions of utilitarianism and Rawls's difference principle still coincide? Or will they now differ? And if they do differ, how?

Proposition 2: Suppose $r_\circ = \frac{W_\circ}{2} < \underline{r_M} = r_M = \bar{W} < 2r_M < r_L = \underline{r_L}$. Then, utilitarianism prescribes $(x_M, x_L) = (r_M, \circ)$, and Rawls's difference principle prescribes $(x_M, x_L) = (\circ, r_M)$.

According to Proposition 2, given that the parties in the original position define MAG's and LAG's reference points as the amount of resources each group needs to enjoy fully the fair worth of their basic rights and liberties and given the amount of social surplus that has been jointly produced by MAG and LAG's mutual social cooperation is such that there is just enough social surplus to allow only MAG (and not LAG) to enjoy the fair worth of their basic rights and liberties fully, utilitarianism recommends giving all the resources to MAG, while Rawls's difference principle recommends giving all the resources to LAG. Both distributions are unattractive; both distributions focus exclusively on one representative group while completely ignoring the other representative group. As a result, under both distributive principles, one of the representative groups (viz., LAG in the case of utilitarianism and MAG in the case of Rawls's difference principle) ends up receiving nothing and falls below what it would expect under equal division of the initial stock of noncooperative default social resources.

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

Despite their overall unattractiveness, there is at least *one* reason that supports the distribution recommended by utilitarianism in such a resource situation; any other distribution will put *both* groups below their reference points. In other words, the utilitarian distribution is the only distribution that makes it possible for the members of at least some group to enjoy the fair worth of their basic rights and liberties fully. The distribution that Rawls's difference principle recommends gives everything to LAG (which results in minimizing the welfare gap between the two groups), but it is still insufficient to make LAG enjoy the full worth of their basic rights and liberties. Thus, how plausible (or repugnant) one finds the respective distributions prescribed by utilitarianism or Rawls's difference principle in this resource situation will partly depend on how much importance one puts on securing each group's reference point (defined here as the amount of resources required to enjoy the fair worth of their basic rights and liberties fully). If one thinks that allowing the members of at least some representative group to enjoy fully the fair worth of their basic rights and liberties whenever possible is very important (even when this implies that the other group may literally receive nothing), then one has no choice but to opt for utilitarianism as it will secure at least MAG's reference point while Rawls's difference principle will secure neither group's reference point. If one is disturbed by the fact that under such a distribution, one representative group (viz., LAG) will end up receiving nothing for the sake of allowing the other group (viz., MAG) to enjoy the fair worth of their basic rights and liberties fully, then this would be a reason to reject defining each group's reference point as the amount of resources required to enjoy the fair worth of their basic rights and liberties fully; it would *not* be a reason to reject utilitarianism in favor of Rawls's difference principle as Rawls's difference principle also results in distributing nothing to one representative group, namely, MAG.

Proposition 3: Suppose $r_o = \frac{W_o}{2} < \underline{r_M} = r_M < 2r_M < \underline{r_L} = r_L < \bar{W} < r_M + r_L$. Then, utilitarianism prescribes $(x_M, x_L) = (r_M, \bar{W} - r_M)$, and the difference principle prescribes $(x_M, x_L) = \left(\frac{\bar{W} - r_L + r_M}{2}, \frac{\bar{W} + r_L - r_M}{2}\right)$.

Proposition 3 concerns another resource situation in which there is enough cooperative social surplus to put the members of one group (either MAG or LAG) above its reference point (defined as the amount of resources each group needs to enjoy the fair worth of their basic rights and liberties fully), but not enough social surplus to make both groups reach their reference points. According to Proposition 3, in such resource situations, utilitarianism prescribes to give just enough resources to MAG so that its members could meet its reference point and then give all of the remaining resources to LAG, while Rawls's difference principle distributes the social surplus in a way that minimizes the welfare gap between MAG and LAG but in a way in which *neither* group meets its reference point. We can make a similar criticism against Rawls's difference principle as before. Rawls's difference principle is not making the most efficient use of the limited amount of resources available; it puts every group below its reference point *unnecessarily*.

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

Furthermore, we can observe that utilitarianism does *not* entail extreme inequality; once MAG reaches its reference point, utilitarianism *prioritizes* LAG and gives all the remaining social surplus to its members.

One argument that Frankfurt (1987) presents against egalitarianism in defense of his doctrine of sufficiency is that when resources are scarce, an egalitarian distribution may result in putting everybody below their critical threshold level (Frankfurt 1987: sect. 4; see Chung [2016] for a critical analysis of Frankfurt's views). We can see that Rawls's difference principle generates similar distributive consequences in our model; that is, when the distributable social surplus is moderately scarce but not too abundant, distributing the available social surplus according to Rawls's difference principle results in making *everybody* fall below his/her reference point, and hence, this distribution keeps everybody from fully enjoying the fair worth of his/her basic rights and liberties.

Proposition 4: Suppose $r_\circ = \frac{W_\circ}{2} < \underline{r_M} = r_M < 2r_M < \underline{r_L} = r_L < \bar{W} = r_M + r_L$. Then, both utilitarianism and the difference principle prescribe $(x_M, x_L) = (r_M, r_L)$.

Proposition 4 concerns a resource situation where there is exactly enough cooperative social surplus to give both MAG and LAG an amount of resources that meets their respective reference points. When such a situation arises, Proposition 4 shows that both utilitarianism and Rawls's difference principle, again, completely coincide and prescribe exactly the same distribution; specifically, both distributional principles distribute the social surplus in such a way that both groups reach their respective reference points, enabling both groups to enjoy the fair worth of their basic rights and liberties fully.

What if the distributable social surplus is abundant—that is, what if the cooperative social surplus is more than enough to put both groups above their reference points?

Proposition 5: Suppose $r_\circ = \frac{W_\circ}{2} < \underline{r_M} = r_M < 2r_M < \underline{r_L} = r_L < r_M + r_L < \bar{W}$. Then, both utilitarianism and the difference principle prescribe $(x_M, x_L) = \left(\frac{\bar{W}+r_M-r_L}{2}, \frac{\bar{W}-r_M+r_L}{2}\right)$.

Proposition 5 says that when there is an abundance of cooperatively produced social surplus, both utilitarianism and justice as fairness, again, recommend exactly the same distribution. In particular, both distributional principles first prioritize making sure that both MAG and LAG get enough resources to reach their respective reference points, and then they distribute the remaining resources equally to each group. This results in a distribution that puts both MAG and LAG at their *highest attainable equal welfare level*.

Note that as a result of putting both MAG and LAG at their highest attainable equal welfare level, both utilitarianism and Rawls's difference principle give $r_L - r_M$ more resources to the members of LAG to compensate for their relative natural and social disadvantages. This is one critical difference that results from assigning each group's reference point as the amount of resources they need

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

to enjoy fully the fair worth of their basic rights and liberties (as opposed to assigning equal division of the initially available noncooperative default social resources as each group's reference point). What this further shows is that, unlike what many people think, utilitarianism does not sacrifice the welfare of the lesser advantaged group, LAG, for the sake of maximizing aggregate social welfare; on the contrary, once the members of MAG meet the group's reference point, utilitarianism gives priority to increasing the welfare levels of LAG so that both MAG and LAG can not only meet their reference points, but ultimately enjoy the highest attainable equal welfare levels whenever society's available resource levels are sufficiently abundant.

From Propositions 1 to 5, we now have a general sense of what sort of distribution utilitarianism and Rawls's difference principle each prescribes when individual utility functions are characterized in accordance with the general characteristics described in prospect theory. The distributional prescriptions of both distributional principles crucially depend on two factors: (a) how the reference point of each representative group is defined and (b) how much cooperative social surplus (jointly produced by MAG and LAG) is left for distribution. Generally speaking, Rawls's difference principle distributes the available social surplus so that the relative welfare gap between MAG and LAG is minimized; whenever possible, it attempts to put both MAG and LAG at their highest attainable equal welfare level from the available resources. On the other hand, the distributional prescriptions of utilitarianism can practically be realized by sequentially implementing the following two rules in serial order:

(1) **Rule 1:** As a first step, maximize the number of individuals reaching their individual reference points.
(2) **Rule 2:** Next, once everybody's reference point is fully reached, equalize everybody's welfare at the highest attainable equal welfare level given the available resources.

It should be clearly understood that utilitarianism only aims to maximize total social welfare and that these two rules are only *by-products* of such a process, which results from the individual utility functions having the particular shapes (characterized in prospect theory) that they are assumed to have. In other words, utilitarianism in this setting does not deliberately attempt to implement the two rules directly, but rather the two rules are achieved indirectly via the implementation of the utilitarian social welfare function.

Let us call the resulting utilitarianism (which results from combining the utilitarian social welfare function with individual utility functions characterized by prospect theory) *prospect utilitarianism* (Chung 2017). We can see that prospect utilitarianism practically turns out to be a hybrid principle, which combines some elements of *sufficientarianism* (Frankfurt 1987, 1997; Crisp 2003) when the social surplus is moderately scarce but not abundant (i.e., when we do not have enough social surplus to allow everybody to enjoy the fair worth of their basic rights and liberties fully) and some elements of (welfare) *egalitarianism* when social surplus is abundant (i.e., when we have enough social surplus to put everybody above their

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

reference points). The important insight of Roemer (2004) was that the most reasonable distributional principle may very well depend on the specific resource context; that is, depending on the resource situation, the most reasonable distributional principle can be utilitarianism or sufficientarianism in some cases, the difference principle or prioritarianism in other cases, and so on. We can see that the specific distributional prescriptions of prospect utilitarianism are sensitive to the resource context in a similar spirit.

In particular, as stated in Proposition 1, when each representative group's reference point is defined as the amount of resources each group expects to receive under equal division of the initial stock of the noncooperatively available default social resources, the distributional prescriptions of prospect utilitarianism completely coincide with those of Rawls's difference principle. When each representative group's reference point is defined as the amount of resources each group needs to enjoy the fair worth of their basic rights and liberties fully, the only difference between prospect utilitarianism and the difference principle in our model is the distribution each prescribes when the distributable social surplus is moderately scarce and not abundant. In such resource scenarios, prospect utilitarianism attempts to maximize the incidences of individuals who meet their individual reference points and guarantees that at least some representative group will fully enjoy the fair worth of their basic rights and liberties whenever possible. By contrast, justice as fairness minimizes the relative welfare gap between MAG and LAG, but in doing so, no group is able to meet its reference point and enjoy the fair worth of their basic rights and liberties fully.

6. Prospect Utilitarianism and the Original Position

Let us go back to the decision problem faced by the representative parties in the original position. Suppose that the representative parties, after deriving the first principle of justice (i.e., the principle of equal basic liberties), conduct a pairwise comparison between prospect utilitarianism and the difference principle.

As we have seen, whenever the parties in the original position assign each representative group's reference point as the amount of resources each group expects to receive under equal division of the initial stock of noncooperatively available default social resources *or* whenever there is an abundant amount of the cooperatively generated social surplus that is more than enough to secure the fair worth of every group's basic rights and liberties, the distributional prescriptions of prospect utilitarianism and Rawls's difference principle are exactly the same—that is, both distributive principles distribute the available social surplus so that everybody enjoys the highest attainable equal welfare level. In these cases, the original contracting parties do not have any reason to favor one distributive principle over the other.

The deciding factor must stem from the differences in how prospect utilitarianism and the difference principle distribute the distributable social surplus when it is moderately scarce and not too abundant. However, here, we have conflicting intuitions that pull us in opposite directions. On the one hand, we have an intuition that says that it is important to provide enough resources so that as

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

many people as possible can enjoy the fair worth of their basic rights and liberties. On the other hand, we have an intuition that says that people should not suffer a loss of welfare due to factors for which they cannot properly be held responsible. Remember that the distribution prescribed by prospect utilitarianism when the social surplus is moderately scarce and not abundant tends to be sufficientarian, while the one prescribed by Rawls's difference principle tends to be welfare egalitarian. Hence, we can say that the distribution prescribed by prospect utilitarianism attends to the former intuition, while the distribution prescribed by the difference principle attends to the latter. Unless the parties in the original position have clear grounds to reject one intuition over the other, they lack a clear reason to favor Rawls's difference principle over prospect utilitarianism.

I would like to note that there is one aspect of our model that has, in effect, handicapped utilitarianism's welfare performances relative to those of Rawls's difference principle; the distributable social surplus (which we assumed to have been jointly produced by MAG's and LAG's social cooperation) was assumed to be given *exogenously*. Surely, how much social wealth is created and made available for distribution (i.e., the size of the 'social pie') would, in the real world, be *endogenously* determined by MAG's and LAG's productive efforts and cooperation. Likewise, the extent of MAG's and LAG's productive contribution to the creation of social wealth would very likely be affected by the productive incentives provided by the basic structure of their society, which, in turn, is determined by the particular principle of justice their society adopts. If we assume that the total distributable social surplus is determined endogenously by MAG's and LAG's productive contributions and take into account the different productive incentives that different basic structures provide to the members of each group, we may expect that the total social wealth produced under a utilitarian basic structure would be different from that produced under a Rawlsian basic structure. In particular, because utilitarianism by its very definition aims to maximize (either total or average) social welfare and because individual welfare is assumed to be increasing in wealth, we would normally expect that the total social wealth produced under a utilitarian basic structure would be *greater* than that produced under a Rawlsian basic structure, which does not directly aim to maximize social wealth. In that case, even if both prospect utilitarianism and Rawls's difference principle completely coincide in their distributional prescriptions by attempting to provide the highest attainable equal welfare level to everybody from the available resources, it could still be possible for both MAG and LAG to be strictly better off under prospect utilitarianism than they are under Rawls's difference principle mainly because there is simply *more* cooperatively produced social surplus under a utilitarian basic structure than under a Rawlsian basic structure.

Hence, if the parties in the original position take the issue of productive incentives into account, the parties in the original position may find additional reasons to favor (prospect) utilitarianism over Rawls's difference principle. This is precisely what I have shown in 'When Utilitarianism Dominates Justice as Fairness: An Economic Defense of Utilitarianism from the Original Position' (Chung 2022). There, by taking into consideration the different productive incentives provided by different basic structures of a productive economy, I formally show that when the

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

differences in people's productive abilities are sufficiently great, utilitarianism Pareto dominates Rawls's two principles of justice by providing a higher level of overall well-being to every member of society.

Let us now consider Rawls's other substantive reasons that he thinks would make the original contracting parties prefer justice as fairness over utilitarianism and see how they fare against prospect utilitarianism. Consider the criticism that utilitarianism does not take seriously the distinction between persons. The basic point of this criticism was that utilitarianism could require unreasonable sacrifices for some members of society for the sake of maximizing aggregate social welfare and that unless these people are perfect altruists who could successfully identify their welfare with the welfare of other people, these people will not be sincerely able to endorse utilitarianism. However, we have seen that prospect utilitarianism does not require such unreasonable sacrifices from anybody. It makes sure that as many people as possible have enough resources to meet their particular reference points, and when society has enough distributable wealth to put everybody above his/her reference point, utilitarianism thereafter makes everybody achieve the highest attainable equal welfare given the resource situation. The criticism that utilitarianism does not take seriously the distinction between persons no longer applies.

We may tackle Rawls's other criticism against utilitarianism in a similar way. The argument from 'strains of commitment' (Rawls 1999: sect. 29) no longer bites because by choosing prospect utilitarianism, the original contracting parties are not entering into agreements that may have consequences they cannot accept. Furthermore, there is also a sense in which the public recognition of prospect utilitarianism will generate its own support and provide a public basis for self-respect because by knowing that their society endorses prospect utilitarianism, people will know that their society will give utmost priority in securing the fair worth of everybody's basic rights and liberties whenever possible and then help everybody reach their highest attainable equal welfare level given the available social resources. Therefore, prospect utilitarianism will be no less stable than justice as fairness (at least under what Rawls calls 'the first level' of publicity characteristic of a 'well-ordered society' in which 'everyone accepts, and knows that everyone else accepts, the very same principles of justice' [Rawls 2005: 35).

Of course, it is still debatable whether prospect utilitarianism can further meet the second and third levels of publicity and achieve what Rawls calls 'the full publicity condition' (Rawls 2005: 67). According to Rawls, when society satisfies the full publicity condition, 'the full justification' of its public conception of justice is 'to be publicly known, or better, at least, to be publicly available' (Rawls 2005: 67). Here, the 'full justification' of the public conception of justice includes both social scientific facts, theories concerning human nature and the operations of social and political institutions as well as the philosophical assumptions and arguments that are employed to justify the public conception of justice regulating the society. The basic worry here is that the social scientific theories (such as Kahneman and Tversky's prospect theory, utility theory and decision theory, utilitarian and Rawlsian social welfare functions, etc.) that were used in this paper to justify prospect utilitarianism from the original position may be too complicated for ordinary citizens to understand properly let alone endorse. Brian Kogelmann calls

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

this 'the demandingness problem' (Kogelmann 2018: 353) or the 'the complexity dilemma' (Kogelmann 2021: 198). According to Christopher Bertram, there is 'a strong presumption against principles which, though transparent in themselves, require for their justification arguments available only to those with specialized theoretical knowledge (such as economists, lawyers, and political philosophers' (Bertram 1997: 564). In this sense, prospect utilitarianism, as a public conception of justice, may fail to meet the full publicity condition.

To respond to this worry, it might be instructive, following Kogelmann (2021), to distinguish three different interpretations of full publicity:

> Accessible Full Publicity only requires access to what we as philosophers say. No actual knowledge of or acceptance of these considerations is required. Known Full Publicity is a bit stronger in that it demands that persons in society *S* know what we philosophers say, whereas Endorsable Full Publicity is even stronger in that it also requires that persons accept these things as true. (Kogelmann 2021: 194)

My short response to the worry is that prospect utilitarianism will at least be able to satisfy Accessible Full Publicity. And although we cannot reasonably expect prospect utilitarianism to satisfy the more demanding Known Full Publicity and Endorsable Full Publicity conditions, the same holds true for Rawls's justice as fairness, and hence, we may say that prospect utilitarianism is no less plausible than Rawls's justice as fairness in regard to the full publicity condition.

We can see that with prospect utilitarianism as a viable option, most of the reasons Rawls presents in favor of justice as fairness against utilitarianism lose their force, and the debate between which principles of justice the parties in the original position would choose becomes, at best, inconclusive.

7. Conclusion and Remarks on the Principle of Restricted Utility

In this paper, I have tried to examine the distributional consequences of utilitarianism and the difference principle (applied to people's welfare levels) when the parties in the original position take Kahneman and Tversky's prospect theory as a part of their general knowledge of human psychology and characterize individual utility functions accordingly. I have called the resulting version of utilitarianism prospect utilitarianism and argued that most of Rawls's substantive criticisms against utilitarianism lose their force. This shows that the implausibility of choosing utilitarianism in the original position heavily depends on what one takes to be the general knowledge of human psychology that the original contracting parties are assumed to know. In this sense, there could be other theories of human psychology that may generate other versions of utilitarianism (or alternate principles of distributive justice) that could be more or less plausible than prospect utilitarianism (see Buchak 2017; Stefansson 2021).

Before ending this paper, I would like to gesture toward the potential application of prospect utilitarianism in defending what Rawls called, 'mixed conceptions' (Rawls 1999: sect. 49) or 'the Principle of Restricted Utility' (Rawls 2001:

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

sect. 38). According to Rawls, 'mixed conceptions . . . are defined by substituting the standard of utility and other criteria for the second principle of justice' (Rawls 1999: 277). In this paper, I have assumed that our model society formally satisfies the first principle of justice. Hence, one may understand my exercise as that of comparing the distributional consequences of justice as fairness to those of a specific 'mixed conception' called 'prospect utilitarianism'. Under this interpretation, prospect utilitarianism may be presented as follows:

Prospect Utilitarianism

1. *The Principle of Equal Basic Liberty*. Each person is to have an equal right to a fully adequate scheme of equal basic liberties compatible with a similar scheme of liberties for others.
2. *Principle of Prospect Utility*. Society should maximize average social welfare assuming that each individual's utility function conforms to the general characteristics proposed by Kahneman and Tversky's prospect theory with its reference point appropriately defined.

Similar to justice as fairness, the two principles of prospect utilitarianism are ordered serially; the principle of prospect utility operates only after the principle of equal basic liberty is fully satisfied. In this paper, I have examined and compared the distributional consequences of prospect utilitarianism under various resource scenarios when each representative group's reference point was defined either as the amount of resources one expects to receive under equal division of the initial stock of noncooperatively available default social resources or as the amount of social primary goods needed to enjoy the fair worth of one's basic rights and liberties. But there could be other more appropriate ways of defining each individual's reference point depending on the overall social and resource context.

One of Rawls's major objections to proposing any such mixed conceptions of justice was that 'the guidelines it suggests do not specify a very definite minimum' (Rawls 2001: 129; see also Rawls 1999: 278–79). In other words, according to Rawls, the notion of a social minimum that any mixed conception purports to guarantee is vague. This is not the case for prospect utilitarianism once the reference points of individual utility functions are appropriately defined. For instance, if we define each individual's reference point in the second way, the basic social minimum prospect utilitarianism aims to guarantee for everybody is the specific amount of material resources that is required to secure the fair worth of each person's basic rights and liberties guaranteed by the first principle of justice. This particular amount of material resources is identified in prospect utilitarianism with the specific reference point of each person's utility function. The problem of vagueness no longer applies. In any case, I hope that this paper will trigger future research that explores the possibilities of this particular mixed conception, which I have called prospect utilitarianism.

HUN CHUNG
WASEDA UNIVERSITY
hun.chung@waseda.jp; hunchung1980@gmail.com

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

References

Allais, M. (1953) 'Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'école Américaine'. *Econometrica*, 21, 503–46.

Bertram, Christopher. (1997) 'Political Justification, Theoretical Complexity, and Democratic Community'. *Ethics,* 107, 563–83.

Bossert, W., and John Weymark. (2004) 'Utility in Social Choice'. In Barbera, Hammond, and Seidl (eds.), *Handbook of Utility Theory.* Vol. 2 (Dordrecht: Kluwer Academic Publishers), 1099–1178.

Buchak, L. (2017) 'Taking Risks behind the Veil of Ignorance'. *Ethics,* 127, 610–44.

Chung, Hun. (2016) 'Is Harry Frankfurt's 'Doctrine of Sufficiency' Sufficient?'. *Organon, F* 23, 50–71.

Chung, Hun. (2017) 'Prospect Utilitarianism: A Better Alternative to Sufficientarianism'. *Philosophical Studies,* 174, 1911–33.

Chung, Hun. (2020) 'Rawls's Self-Defeat: A Formal Analysis'. *Erkenntnis,* 85, 1169–97.

Chung, Hun. (2022) 'When Utilitarianism Dominates Justice as Fairness: An Economic Defense of Utilitarianism from the Original Position.' *Economics & Philosophy*, 1–26. doi:10.1017/S0266267122000098.

Crisp, Roger. (2003) 'Equality, Priority, and Compassion'. *Ethics,* 113, 745–63.

Ellsberg, Daniel. (1961) 'Risk, Ambiguity, and the Savage Axioms'. *Quarterly Journal of Economics*, 75, 643–69.

Frankfurt, Harry. (1987) 'Equality as a Moral Ideal'. *Ethics*, 98, 21–43.

Frankfurt, Harry. (1997) 'Equality and Respect'. *Social Research,* 64, 3–15.

Gaus, Gerald, and John Thrasher. (2015) 'Rational Choice and the Original Position, The (Many) Models of Rawls and Harsanyi'. In Timothy Hinton (ed.), *The Original Position* (Cambridge: Cambridge University Press), 39–58.

Harsanyi, John. (1953) 'Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking'. *Journal of Political Economy,* 61, 434–35.

Harsanyi, John. (1955) 'Cardinal Welfare, Individualistic Ethics, and the Interpersonal Comparisons of Utility'. *Journal of Political Economy,* 63, 309–21.

Harsanyi, John. (1975) 'Review, Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory'. *American Political Science Review,* 69, 594–606.

Harsanyi, John. (1977) *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.

Kahneman, Daniel, and Amos Tversky. (1979) 'Prospect Theory: An Analysis of Decision under Risk'. *Econometrica*, 47, 263–91.

Kahneman, Daniel, and Amos Tversky. (1992) 'Advances in Prospect Theory: Cumulative Representation of Uncertainty'. *Journal of Risk and Uncertainty,* 5, 297–323.

Kogelmann, Brian. (2018) 'The Supreme Court as The Fountain of Public Reason'. *Legal Theory,* 24, 345–69.

Kogelmann, Brian. (2021) *Secret Government: The Pathologies of Publicity*. Cambridge: Cambridge University Press.

Moehler, Michael. (2015) 'The Rawls-Harsanyi Dispute: A Moral Point of View'. *Pacific Philosophical Quarterly,* 99, 82–99.

Moreno-Ternero, Juan, and John Roemer. (2008) 'The Veil of Ignorance Violates Priority'. *Economics and Philosophy,* 24, 233–57.

Rawls, John. (1974) 'Some Reasons for the Maximin Criterion'. *American Economic Review,* 64, 141–46.

Rawls, John. (1999) *A Theory of Justice*. Rev. ed. Cambridge, MA: Harvard University Press.

Rawls, John. (2001) *Justice as Fairness*. Cambridge, MA: Belknap Press of Harvard University Press.

Rawls, John. (2005) *Political Liberalism*. Ex. ed. Columbia University Press.

Roemer, John. (1996) *Theories of Distributive Justice*. Harvard University Press.

Roemer, John. (2002) 'Egalitarianisms against the Veil of Ignorance'. *Journal of Philosophy,* 99, 164–84.

Roemer, John. (2004) 'Eclectic Distributional Ethics'. *Politics, Philosophy, & Economics,* 3, 267–81.

Roemer, John. (n.d.) 'The Original Position'. Unpublished manuscript.

Sen, Amartya. (1980) 'Equality of What?' In Sterling M. McMurrin (ed.), *The Tanner Lecture on Human Value*, vol. 1 (Cambridge: Cambridge University Press), 195–220.
Stefansson, H. O. (2021) 'Ambiguity Aversion behind the Veil of Ignorance'. *Synthese*, 198, 6159–182.

Appendix – Proofs of Main Propositions

Proposition 1: *Suppose* $\underline{r_M} = \underline{r_L} = r_0 = \frac{W_0}{2}$. *Then for all* $\bar{W} \geq W_0$, *the distributional prescriptions of utilitarianism and Rawls's difference principle are the same—namely,* $(x_M, x_L) = \left(\frac{\bar{W}}{2}, \frac{\bar{W}}{2}\right)$.

Proof of Proposition 1. We first derive the distribution prescribed by utilitarianism. Since both u_M and u_L are strictly increasing in resources, any distribution (x_M, x_L) prescribed by utilitarianism must distribute *all* the available social surplus $\bar{W}$ to MAG and LAG; that is, we must have $x_M + x_L = \bar{W}$. Otherwise, if we had $x_M + x_L < \bar{W}$, then we could always distribute the remainder of the resources $\bar{W} - (x_M + x_L)$ to either MAG or LAG in addition to what they were initially distributed, and in this way we would strictly increase the total sum of utilities, contradicting that the initial distribution (x_M, x_L) was the utilitarian distribution that maximizes the total sum of utilities.

We will now divide the distributional problem into two stages: in the first stage, we will initially distribute W_0; in the second stage, we will distribute the remainder $\bar{W} - W_0$. Consider the first stage distribution problem, in which we distribute W_0. Here, utilitarianism faces the following problem:

$$\max_{(x_1, x_2) \in \mathbb{R}^2} u_M(x_1) + u_L(x_2)$$

$$\text{subject to } x_1 + x_2 \leq W_0$$

By Property 5, given $\underline{r_M} = \underline{r_L} = r_0 = \frac{W_0}{2}$, we have $u_M = u_L$. By Property 3, for all x, x' such that $0 \leq x \leq \underline{r_M} = \underline{r_L} = r_0 = \frac{W_0}{2} < x'$, we have $u'_M(x) = u'_L(x) > u'_M(x') = u'_L(x')$ and $u''_M(x) = u''_L(x) > 0$. Here, $u'_M(x) = u'_L(x) > u'_M(x') = u'_L(x')$ implies that distributing any additional resources to any group whose resource level is *below* its reference point will increase the total sum of utilities of the two groups better than distributing those resources to any group whose resource level is *above* its reference point. And $u''_M(x) = u''_L(x) > 0$ implies that the utilities of the two groups will increase at a higher rate when their resource levels approach their reference

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

points from below. Together, this implies that dividing the available resources W_0 into half and giving MAG and LAG each $r_0 = \frac{W_0}{2}$ will maximize the total sum of MAG's and LAG's utilities for the first stage distributional problem. Thus, $(x_M,\ x_L) = \left(\frac{W_0}{2},\ \frac{W_0}{2}\right)$ will be the initial distribution prescribed by utilitarianism for the first stage distributional problem. Now, we move on to the second stage distributional problem, where the problem has been reduced to:

$$\max_{(x_1,\, x_2)\in\mathbb{R}^2} u_M\left(\frac{W_0}{2} + x_1\right) + u_L\left(\frac{W_0}{2} + x_2\right)$$

$$\text{subject to } x_1 + x_2 \leq \bar{W} - W_0 \qquad (.)$$

Again, by Property 5, given $\underline{r_M} = \underline{r_L} = r_0 = \frac{W_0}{2}$, we have $u_M = u_L$. Also, again, any distribution prescribed by utilitarianism must use all the available resources: that is, we must have $x_1 + x_2 = \bar{W} - W_0$, which, by rearrangement, gives us: $x_2 = \bar{W} - W_0 - x_1$. Plugging this into LAG's utility function and rewriting LAG's utility function as MAG's utility function the problem can be further reduced to:

$$\max_{x_1\in\mathbb{R}^2} u_M\left(\frac{W_0}{2} + x_1\right) + u_M\left(\bar{W} - \frac{W_0}{2} - x_1\right),$$

which is a simple (strict) concave program. Since u_M is strictly concave in x_1, the objective function $u_M\left(\frac{W_0}{2} + x_1\right) + u_M\left(\bar{W} - \frac{W_0}{2} - x_1\right)$ is strictly concave in x_1. Hence, the first-order condition will be sufficient to give us the unique maximizer x_1^* that maximizes the objective function. Taking the first-order condition, we have:

$$u_M'\left(\frac{W_0}{2} + x_1\right) = u_M'\left(\bar{W} - \frac{W_0}{2} - x_1\right),$$

which implies $\frac{W_0}{2} + x_1 = \bar{W} - \frac{W_0}{2} - x_1 \Rightarrow x_1^* = \frac{\bar{W}}{2} - \frac{W_0}{2}$. Plugging this into $x_2 = \bar{W} - W_0 - x_1$, we get: $x_2^* = \frac{\bar{W}}{2} - \frac{W_0}{2}$. Combining the results of the first and second stage distributional problems, the final distribution that utilitarianism prescribes is: $(x_M^*,\ x_L^*) = \left(\frac{W_0}{2} + x_1^*,\ \frac{W_0}{2} + x_2^*\right) = \left(\frac{\bar{W}}{2},\ \frac{\bar{W}}{2}\right)$ as desired.

Let us now derive the distribution prescribed by Rawls's difference principle. The distributional problem that Rawls's

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

difference principle solves is:

$$\max_{(x_1, x_2)\in\mathbb{R}^2} \min\{u_M(x_1), u_L(x_2)\}$$

$$\text{subject to } x_1 + x_2 \leq \bar{W}.$$

I claim that in the distribution prescribed by Rawls's difference principle, we must have $x_1 = x_2$. For suppose not. That is, suppose $(x_M, x_L) = (x_1, x_2)$ is the distribution prescribed by Rawls's difference principle where $x_1 \neq x_2$. Without loss of generality, suppose $x_1 < x_2$. By Property 5, given $\underline{r_M} = \underline{r_L} = r_0 = \frac{W_0}{2}$, we have $u_M = u_L$. By Property 3, u_M is strictly increasing. Hence, we have $u_M(x_1) < u_M(x_2) = u_L(x_2)$. Thus, $\min\{u_M(x_1), u_L(x_2)\} = u_M(x_1)$. Since $u_M(x_1) < u_M(x_2)$ and since u_M is continuous, there exists a small enough $\epsilon > 0$ such that $u_M(x_1) < u_M(x_1 + \epsilon) < u_M(x_2 - \epsilon) = u_L(x_2 - \epsilon) < u_M(x_2) = u_L(x_2)$. Then, $(x'_M, x'_L) = (x_1 + \epsilon, x_2 - \epsilon)$ is another feasible distribution such that $\min\{u_M(x_1), u_L(x_2)\} = u_M(x_1) < u_M(x_1 + \epsilon) = \min\{u_M(x_1 + \epsilon), u_L(x_2 - \epsilon)\}$, contradicting that $(x_M, x_L) = (x_1, x_2)$ is the distribution prescribed by Rawls's difference principle. Hence, if $(x_M, x_L) = (x_1, x_2)$ is the distribution prescribed by Rawls's difference principle, we must have $x_1 = x_2 = x$. Next, I claim that if $(x_M, x_L) = (x, x)$ is the distribution prescribed by Rawls's difference principle, then we must have $x_M + x_L = 2x = \bar{W}$. For suppose not. That is, suppose that $(x_M, x_L) = (x, x)$ is the distribution prescribed by Rawls's difference principle, but we have $2x < \bar{W}$. Define $\Delta = \frac{\bar{W} - 2x}{2}$. Then, $(x'_M, x'_L) = (x + \Delta, x + \Delta)$ is another feasible distribution such that $\min\{u_M(x), u_L(x)\} = u_M(x) = u_L(x) < u_M(x + \Delta) = u_L(x + \Delta) = \min\{u_M(x + \Delta), u_L(x + \Delta)\}$, which contradicts that $(x_M, x_L) = (x, x)$ is the distribution prescribed by Rawls's difference principle. Hence, we must have $2x = \bar{W} \Rightarrow x = \frac{\bar{W}}{2}$. Hence, we conclude that $(x_M, x_L) = \left(\frac{\bar{W}}{2}, \frac{\bar{W}}{2}\right)$ is also the distribution prescribed by Rawls's difference principle. ∎

Proposition 2: Suppose $r_0 = \frac{W_0}{2} < \underline{r_M} = r_M = \bar{W} < 2r_M < \underline{r_L} = r_L$. Then utilitarianism prescribes $(x_M, x_L) = (r_M, 0)$, and Rawls's difference principle prescribes $(x_M, x_L) = (0, r_M)$.

Proof of Proposition 2. Note that any $x \in [0, r_M]$ can be expressed as a convex combination of 0 and r_M. Let $\alpha \in (0, 1)$. Then, because both u_M and u_L are strictly convex below their

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

reference points, we have:

$$u_M(\alpha \cdot 0 + (1-\alpha)r_M) + u_L((1-\alpha)\cdot 0 + \alpha r_M)$$

$$\leq \alpha u_M(0) + (1-\alpha)u_M(r_M) + (1-\alpha)u_L(0) + \alpha u_L(r_M) \quad (1)$$

$$\text{I claim that } (1) < u_M(r_M) + u_L(0) \quad (2)$$

To show this, note that $(2) - (1)$

$$= \alpha[u_M(r_M) - u_M(0)] - \alpha[u_L(r_M) - u_L(0)]$$

$$= \alpha\left[\int_0^{r_M} u'_M(x)\,dx - \int_0^{r_M} u'_L(x)\,dx\right]$$

$$= \alpha \int_0^{r_M} [u'_M(x) - u'_L(x)]\,dx$$

$$= \alpha \int_0^{r_M} [u'_L(x + (r_L - r_M)) - u'_L(x)]\,dx \text{ (since } u_M(x) = u_L(x + (r_L - r_M)))$$

$$> 0.$$

(Because $u''_L(x) > 0$ for all $x \in [0, r_M]$, u'_L is strictly increasing and, hence, $u'_L(x + (r_L - r_M)) - u'_L(x) > 0$.) Thus, we have, for all $\alpha \in (0, 1]$,

$$u_M(\alpha \cdot 0 + (1-\alpha)r_M) + u_L((1-\alpha)\cdot 0 + \alpha r_M) < u_M(r_M) + u_L(0),$$

which means that $(r_M, 0)$ is the solution to

$$\max_{(x_1, x_2)\in\mathbb{R}^2} u_M(x_1) + u_L(x_2)$$

$$\text{subject to } x_1 + x_2 \leq r_M.$$

Hence, $(x_M, x_L) = (r_M, 0)$ is the utilitarian solution.

Now, note $u_M(0) = u_L(r_L - r_M)$.

Because by assumption $2r_M < r_L$, this implies $r_M < r_L - r_M$.

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

Thus, for all $x \in [0, r_M)$, we have:

$$\min\{u_M(0),\ u_L(r_M)\} = \min\{u_L(r_L - r_M),\ u_L(r_M)\} = u_L(r_M)$$

$$> \min\{u_M(r_M - x),\ u_L(x)\} = \min\{u_L(r_L - x),\ u_L(x)\}$$
$$= u_L(x).$$

Therefore, $(0, r_M)$ is the solution to:

$$\max_{(x_1, x_2) \in \mathbb{R}^2} \min\{u_M(x_1), u_L(x_2)\}$$

$$\text{subject to } x_1 + x_2 \leq r_M.$$

That is, $(x_M, x_L) = (0, r_M)$ is the distribution prescribed by Rawls's difference principle. ■

Proposition 3: Suppose $r_0 = \frac{W_0}{2} < \underline{r_M} = r_M < 2r_M < \underline{r_L} = r_L < \bar{W} < r_M + r_L$. Then utilitarianism prescribes $(x_M, x_L) = (r_M, \bar{W} - r_M)$, and the difference principle prescribes $(x_M, x_L) = \left(\frac{\bar{W} - r_L + r_M}{2}, \frac{\bar{W} + r_L - r_M}{2}\right)$.

Proof of Proposition 3. Note that for all $x \in [0, r_M)$, $u'_M(x) > u'_L(x) > 0$. Therefore, giving all the resources to MAG until her resources level reaches r_M is the distribution that maximizes the sum of MAG's and LAG's utilities. Once r_M is distributed to MAG, we have $\bar{W} - r_M$ of wealth left to distribute to the two individuals. Note that for all $x \in (r_M, \bar{W}]$ and for all $y \in [0, \bar{W} - r_M]$, we have $u'_L(y) > u'_M(x) > 0$. Therefore, giving all the remaining $\bar{W} - r_M$ of wealth to LAG (after MAG has received r_M), will be the distribution that would maximize the sum of MAG's and LAG's utilities. Hence, $(r_M, \bar{W} - r_M)$ is the utilitarian solution.

Now, note that $u_L(0) < u_M(0) = u_L(r_L - r_M)$. Initially, giving $r_L - r_M$ amount of resources to LAG would equalize the utilities between MAG and LAG. After giving $r_L - r_M$ amount of resources to LAG, we have $\bar{W} - (r_L - r_M)$ of wealth left. Once the utilities between MAG and LAG are equalized, distributing the remaining resources in a way that retains the equality in utility between the two individuals would be the only way to maximize the minimal utility between the two individuals. Or else, suppose not. Then there exists a distribution $(x, \bar{W} - x)$ (with $0 \leq x \leq \bar{W} - r_L + r_M$) such that $(x, \bar{W} - x)$ solves the

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

maximization problem,

$$\max_{(x_1, x_2)\in\mathbb{R}^2} \min\{u_M(x_1), u_L(x_2)\} \text{ subject to } x_1 + x_2 \leq \bar{W}$$

and either $u_M(x) > u_L(\bar{W} - x)$ or $u_M(x) < u_L(\bar{W} - x)$ hold. If the former holds, then there exists a small enough $\epsilon > 0$ such that $u_M(x) > u_M(x - \epsilon) > u_L(\bar{W} - x + \epsilon) > u_L(\bar{W} - x)$. If the latter holds, then there exists a small enough $\epsilon > 0$ such that $u_M(x) < u_M(x + \epsilon) < u_L(\bar{W} - x - \epsilon) < u_L(\bar{W} - x)$. Both cases contradict that $(x,\ \bar{W} - x)$ solves the constrained maximization problem above. Therefore, in order to maximize the minimal utility between the two individuals, we would need to distribute the remaining resources in a way that retains the equality in utilities between the two individuals. Now, note that for all $x \in [0,\ \bar{W} - r_L + r_M]$ we have $u'_L(x + (r_L - r_M)) = u'_M(x)$. Therefore, giving equal amounts of the remaining $\bar{W} - r_L + r_M$ of resources (i.e., $\frac{\bar{W} - r_L + r_M}{2}$) to each individual will equalize their utility levels. Hence, $(x_M, x_L) = \left(0 + \frac{\bar{W} - r_L + r_M}{2},\ r_L - r_M + \frac{\bar{W} - r_L + r_M}{2}\right) = \left(\frac{\bar{W} - r_L + r_M}{2}, \frac{\bar{W} + r_L - r_M}{2}\right)$ is the distribution prescribed by Rawls's difference principle. ■

Proposition 4: Suppose $r_0 = \frac{W_0}{2} < \underline{r_M} = r_M < 2r_M < \underline{r_L} = r_L < \bar{W} = r_M + r_L$. Then both utilitarianism and the difference principle prescribe $(x_M, x_L) = (r_M, r_L)$.

Proof of Proposition 4. The result follows from the proof of Proposition 3 by setting $\bar{W} = r_M + r_L$. ■

Proposition 5: Suppose $r_0 = \frac{W_0}{2} < \underline{r_M} = r_M < 2r_M < \underline{r_L} = r_L < r_M + r_L < \bar{W}$. Then both utilitarianism and the difference principle prescribe $(x_M, x_L) = \left(\frac{\bar{W} + r_M - r_L}{2},\ \frac{\bar{W} - r_M + r_L}{2}\right)$.

Proof of Proposition 5. By Proposition 4, when the available social surplus is $r_M + r_L$, $(r_M,\ r_L)$ is the distribution that maximizes the sum of individual utilities of MAG and LAG. Thus, as a first step, distribute r_M to MAG and r_L to LAG. After such distribution, we have $\bar{W} - (r_M + r_L)$ of resources left for further distribution. Now the problem reduces to:

$$\max_{(x_1, x_2)\in\mathbb{R}^2} u_M(r_M + x_1) + u_L(r_L + x_2)$$

$$\text{subject to } x_1 + x_2 \leq \bar{W} - (r_M + r_L).$$

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press

Note that for all $x \in [0, \ \bar{W} - (r_M + r_L)]$, we have $u_M(r_M + x) = u_L(r_L + x)$. Thus, $u_M(r_M + x_1) + u_L(r_L + x_2) = u_M(r_M + x_1) + u_M(r_M + x_2)$. Again, any utilitarian solution requires $x_1 + x_2 = \bar{W} - (r_M + r_L)$. By substituting $\bar{W} - (r_M + r_D) - x_1$ for x_2, the problem is now further simplified to maximizing $u_M(r_M + x_1) + u_M(\bar{W} - r_L - x_1)$. Since u_M is strictly concave in x_1, $u_M(r_M + x_1) + u_M(\bar{W} - r_L - x_1)$ is also strictly concave in x_1, and hence, the first order condition is sufficient for it to obtain its maximum. Taking derivatives with respect to x_1 and setting it equal to zero we have:

$$u'_M(r_M + x_1) - u'_M(\bar{W} - r_L - x_1) = 0$$

$$\Rightarrow u'_M(r_M + x_1) = u'_M(\bar{W} - r_L - x_1)$$

$$\Rightarrow r_M + x_1 = \bar{W} - r_L - x_1$$

$$\Rightarrow x_1 = \frac{\bar{W} - r_M - r_L}{2} \ \ and \ \ x_2 = \frac{\bar{W} - r_M - r_L}{2} \ .$$

Hence, $(x_M, \ x_L) = \left(r_M + \frac{\bar{W} - r_M - r_L}{2}, \ r_L + \frac{\bar{W} - r_M - r_L}{2}\right) = \left(\frac{\bar{W} + r_M - r_L}{2}, \ \frac{\bar{W} - r_M + r_L}{2}\right)$ is the utilitarian solution.

To derive the distribution prescribed by Rawls's difference principle, for the same reason provided in the proof of Proposition 3, we would need to find a distribution (x_1, x_2) such that $x_1 + x_2 = \bar{W}$ and $u_M(x_1) = u_L(x_2)$; that is, we would need to find a distribution that uses up all the available resources and equalizes the utilities of MAG and LAG. Let $(x_1, \ x_2) = \left(\frac{\bar{W} + r_M - r_L}{2}, \ \frac{\bar{W} - r_M + r_L}{2}\right)$. Then, we can see that $x_1 + x_2 = \bar{W}$ and $u_L\left(\frac{\bar{W} - r_M + r_L}{2}\right) = u_M\left(\frac{\bar{W} - r_M + r_L}{2} - (r_L - r_M)\right) = u_M\left(\frac{\bar{W} + r_M - r_L}{2}\right)$ as desired. ■

https://doi.org/10.1017/apa.2022.31 Published online by Cambridge University Press