Erkenntnis (2020) 85:1169–1197
https://doi.org/10.1007/s10670-018-0079-4

ORIGINAL RESEARCH

# Rawls's Self-Defeat: A Formal Analysis

**Hun Chung**[1]

Published online: 7 December 2018


## Abstract

One of John Rawls's major aims, when he wrote *A Theory of Justice*, was to present a superior alternative to utilitarianism. Rawls's worry was that utilitarianism may fail to protect the fundamental rights and liberties of persons in its attempt to maximize total social welfare. Rawls's main argument against utilitarianism was that, for such reasons, the representative parties in the original position will not choose utilitarianism, but will rather choose his justice as fairness, which he believed would securely protect the worth of everybody's basic rights and liberties. In this paper, I will argue that, under close formal examination, Rawls's argument against utilitarianism is self-defeating. That is, I will argue that Rawls's own reasons, assumptions, and the many theoretical devices he employs demonstrably imply that the representative parties in the original position will choose utilitarianism instead of justice as fairness.

I thank Ill-Soo Cho, John Duggan, Jerry Gaus, Adam Gjesdal, Jinhee Jo, Brian Kogelmann, Philippe Lemoine, Julia Markovits, Kian Mintz-Woo, Ryan Muldoon, Alex Schaefer, Abe Singer, John Thrasher, Kevin Vallier, David Wiens, participants at the Philosophy Department Workshop at Cornell University, participants at the Institute of Peace and Democracy Workshop at Korea University, participants at the Department of Political Science and International Relations Graduate Seminar at Kyunghee University, participants at the Korean Association of Ethics 2017 Winter Conference, participants at the "Topics in Philosophy, Politics, and Economics" Session at the 86th Southern Economic Association, the faculty of the Philosophy and Economics Departments at the University of Bayreuth, and two anonymous reviewers for their helpful comments on earlier drafts of this paper.

✉ Hun Chung
hunchung1980@gmail.com

[1] Faculty of Political Science and Economics, Waseda University, Tokyo, Japan

## 1 John Rawls's *Justice as Fairness,* A Triumph Against Utilitarianism?

In *A Theory of Justice*, John Rawls wished to present a theory of distributive justice that was superior to utilitarianism. The resulting theory is what Rawls called "justice as fairness" which is composed of the following three principles[1] stated in the order of strict priority,

1. *The Principle of Maximum Equal Basic Liberties*: Each person is to have an equal right to the most extensive scheme of equal basic liberties compatible with a similar scheme of liberties for others.
2. *The Principle of Fair Equal Opportunity*: Social economic inequalities should be attached to positions and offices opened to all under conditions of fair equal opportunity.
3. *The Difference Principle*: Social and economic inequalities should be arranged in a way that is the greatest benefit to the least-advantaged members of society.

The thought was that these three principles would be chosen over utilitarianism by the representative parties of "the original position" behind "the veil of ignorance." The original position is an initial situation where the representative parties of society decide the fundamental guiding principles regulating the basic structure of their society by their own voluntary agreement. The veil of ignorance is a theoretical device that guarantees the fairness of the resulting agreement by depriving the original contracting parties of morally irrelevant information.[2]

Rawls's argument for his three principles of justice—i.e. justice as fairness—and his argument against the principle of average utility—i.e. utilitarianism—are like two sides of the same coin. Rawls's justification for justice as fairness derives from the purported fact that the parties in the original position will choose it over utilitarianism; Rawls's refutation of utilitarianism derives from the purported fact that the original contracting parties will not choose it over justice as fairness. For Rawls, it is this very choice for justice as fairness from the original position that lends its very justification as well as a conclusive refutation of utilitarianism.[3]

[1] Actually, Rawls described himself as presenting two principles of justice (see Rawls 1999, 53), in which the principle of fair equal opportunity and the difference principle comprise two parts of the second principle. The main reason why I restate Rawls's theory of justice as being composed of three (instead of two) principles is to isolate the difference principle, which will be the focus of our discussion later on.

[2] According to Rawls, the contracting parties "do not know their place in society, their class position or social status, their place in the distribution of natural assets and abilities, their deeper aims and interests, or their particular psychological makeup." (Rawls 1974a, 141) "And to insure fairness between generations, we must add that they do not know to which generation they belong and thus information about natural resources, the level of productive techniques, and the like, is also forbidden to them" (Rawls 1974b, 637). Rawls's 'thick' veil of ignorance is in contrast with Ronald Dworkin's 'thin' veil of ignorance behind which the parties are cognizant of their own preferences. (see Dworkin 1981a, b).

[3] For an excellent survey of how different rational choice models of the original position of both John Rawls and John Harsanyi have developed in the literature, see Gaus and Thrasher (2015); See also D'agostino et al. (2014) for a survey of contemporary approaches to the social contract.

In a similar period, Harsanyi (1955, 1977) proved two mathematical results which he believed to formally justify utilitarianism.[4] There remain some controversies on whether Harsanyi's formal results really support utilitarianism.[5] Even so, Harsanyi's contributions ignited one of the most hotly debated topics in contemporary political philosophy—namely, whether the parties in Rawls's original position would choose justice as fairness or utilitarianism. Call this the "Rawls-vs-Harsanyi Debate." The debate has now reached a complete deadlock—each side simply denying the crucial assumption made by the other.

For instance, supporters of utilitarianism claim that the "maximin rule" (from which Rawls seems to derive support for the original parties' choice for the difference principle) is either straight out irrational (Harsanyi 1975) or depends on assuming that the original contracting parties are extremely risk-averse (Roemer 2002; Moreno-Ternero and Roemer 2008).[6] To this, Rawls argues that the selection for his justice as fairness, and, in particular, the difference principle does not depend on any assumption regarding the original contracting parties' attitude toward risk (Rawls 1971/1999: 149; 2001: 106–107).

On the other hand, Rawlsians argue that in order for the parties of the original position to choose utilitarianism, they would have to be able to calculate expectations, but, the specific characteristics of the original position provide absolutely no basis for the parties to make any kind of probability judgments which renders the task of calculating expectations unfeasible (Rawls 1971/1999: 146–149). To this, utilitarians claim that, in such cases, we should rely on "the principle of insufficient reason" and assign equal probabilities (Harsanyi 1977: 50) for being born in each starting place in society, and once this is done, it would be rational for the original contracting parties to perform an expected utility calculation which would result in them adopting the principle of average utility (Harsanyi 1977: 50).

[4] In his 1955 paper, Harsanyi (1955) showed that if we assume that individual and social preferences conform to the axioms of von Neumann–Morgenstern's expected utility theory, and we further assume Pareto Indifference, social preferences can be represented as a weighted sum of individual utilities. This is known as "Harsanyi's Aggregation Theorem." In his 1977 book, Harsanyi (1977) showed that an impartial observer who imagines him/herself to have an equal chance of being any person in society will choose a social arrangement that maximizes society's average utility level. This is known as "Harsanyi's Impartial Observer Theorem".

[5] Sen (1976) argues that Harsanyi's Aggregation Theorem is merely a representation theorem, and not a defense of utilitarianism. Weymark (1991) formalizes Sen's critique of Harsanyi, and Roemer (1996, 2008) raises a similar point against Harsanyi's Impartial Observer Theorem. Broome (1987) argues that Harsanyi's theorem can be reinterpreted as a defense of utilitarianism once we interpret utility to represent goodness rather preferences. Risse (2002) follows Broome and argues that Harsanyi's Aggregation Theorem does support utilitarianism.

[6] Just to be clear. I am not claiming that Roemer is a defender of utilitarianism or that he thinks that Harsanyi's representation theorems give support for utilitarianism; he actually denies both. These are simply Roemer's criticisms against adopting the maximin rule in the original position.

After observing that the Rawls-vs-Harsanyi debate has reached a deadlock, Moehler (2018) concludes that there is no clear winner of the Rawls-Harsanyi dispute as each author attempts to model different moral ideals.

My approach in this paper is different. This paper is not intended to be another paper that comments on the Rawls-vs-Harsanyi debate[7] or a paper that tries to defend Harsanyi's version of utilitarianism. Rather, the purpose of this paper is to provide a very close formal examination of Rawls's own arguments in support for his justice as fairness as well as his arguments against utilitarianism, and see whether each micro-component of Rawls's arguments, once analyzed formally, hold together well. Hence, my main focus will be on the *validity* rather than the *soundness* of Rawls's argument. For this purpose, I will grant pretty much all of Rawls's major assumptions, and then try to show that Rawls's own weapons (viz. his assumptions, reasons and theoretical devices), on which he relies to justify justice as fairness and refute utilitarianism, actually work against him—that is, they actually support utilitarianism and undermine his own justice as fairness. With due respect, I will try to show that Rawls's arguments are self-defeating.

## 2 Rawls's Criticism of Utilitarianism

As explained, simply put, Rawls's refutation of utilitarianism was that it would not be chosen over justice as fairness from the original position. Then, why wouldn't the parties in the original position choose utilitarianism instead of justice as fairness? Rawls presents several reasons; he talks about strains of commitment, the distinction between persons, the publicity condition, and issues related to stability and self-respect.[8] To understand the central idea that underlies these criticisms, it might be helpful to remind ourselves of a standard criticism against utilitarianism that is frequently made—namely, that utilitarianism may, in principle, justify the institution of slavery. Basically, we might interpret Rawls's criticisms as inviting us to imagine a situation in which we chose utilitarianism behind the veil of ignorance, but, then, discovered ourselves to be slaves after the veil of ignorance has been lifted. Would we be able to honor our original agreement? No. (This is the argument from strains of commitment.) If our society officially affirms that it will follow utilitarianism as its fundamental guiding principle (which is required by the publicity condition) and tries to publicly justify that the sacrifices of slaves are required to maximize social welfare, would it be possible for us, as slaves, to retain our self-respect? No. (This is the argument from the publicity condition and self-respect.) Wouldn't this be a case

[7] See Fleurbaey et al. (2008) for the most up-to-date collection of articles that concerns this debate.

[8] Most of these criticisms are contained in Section 29 of Rawls (1971/1999).

of failing to take the distinctness and separateness of different persons seriously? (this is the argument from the distinctness of persons.) And, if utilitarianism will be unable to generate wide universal support, would a political society regulated by utilitarianism be stable? No. (This is the argument from stability.)

All of these are importantly distinct considerations that may explain why the original contracting parties in the original position would favor justice as fairness over utilitarianism. However, we can see that there is a central theme that penetrates all of these considerations; namely, there is a real danger that utilitarianism, once affirmed, might require one to sacrifice one's most fundamental interests and basic rights/liberties for the sake of maximizing total or average social welfare. The reasons stemming from strains of commitment, stability, and self-respect are all mere implications of this possible consequence of utilitarianism. And, the purported fact that justice as fairness, with its three principles, is able to securely protect these fundamental interests and basic rights/liberties (by its first principle) as well as their social worth (by its second and third principle; Rawls 1971/1999, 179) is the decisive reason why Rawls believes that the representative parties of the original position will choose his justice as fairness over any form of utilitarianism (see Rawls 2001, 102).

## 3 The Difference Principle and Primary Social Goods

Now, if it is true that each individual's fundamental interests and basic rights/liberties are firmly secured by the very first principle (i.e. the principle of maximum equal basic liberties) of justice as fairness, then what role do the other two principles (i.e. the principle of fair equal opportunity and the difference principle) really play in the representative parties' decision to choose justice as fairness over utilitarianism? Why, for instance, should they not choose a conception of justice that combines the principle of maximum equal basic liberties with, say, the principle of average utility[9] instead of the difference principle? One of Rawls's reasons is that the difference principle (compared to the principle of average utility) better secures *the worth* of the basic liberties that the principle of equal basic liberties formally guarantees. According to Rawls,

> Freedom as equal liberty is the same for all… 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. … Taking [the principle of maximum equal basic liberties and the difference principle] 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 1971/1999, 179 emphasis added)

[9] Rawls calls such conception of justice, the "principle of restricted utility" (Rawls 2001, section 38), and ultimately rejects it.

Again, the main point of the difference principle is that, by maximizing the expectation of the least advantaged group in society, it best guarantees that every member of society, especially the least advantaged group, enjoys *the best worth* of their basic right/liberties that the principle of equal basic liberties formally guarantees as much as possible. That is why we need the difference principle along with the principle of equal basic liberties.

Now, Rawls claims that the difference principle is designed to apply to the distribution of what he calls "primary social goods." According to Rawls,

> … primary goods … are things which it is supposed a rational man wants whatever else he wants. … The primary social goods, to give them in broad categories, are rights, liberties, and opportunities, and income and wealth. (Rawls 1971/1999, 79)

The main reason for introducing the idea of the primary social goods is as follows. In order for the difference principle to apply, one needs to identify which group is the least advantaged group in society. This requires interpersonal comparisons. However, Rawls wanted to avoid making interpersonal comparisons in terms people's welfare levels. A search for a more objective basis for interpersonal comparison is what led Rawls to rely on primary social goods.

> …the difference principle introduces a simplification for the basis of interpersonal comparisons. These comparisons are made in terms of expectations of primary social goods. In fact, I define these expectations simply as the *index* of these goods which a representative individual can look forward to. *One man's expectations are greater than another's if this index for someone in his position is greater.* (Rawls 1971/1999, 79 emphasis added)

The basic thought is that we can assign numbers to different bundles of primary social goods in a way that bundles that are assigned higher numbers are more valuable to *everybody*, regardless of his/her particular aims and goals, than bundles that are assigned lower numbers. From this, the problem of interpersonal comparison becomes greatly simplified; person *A* is better off than person *B* if and only if *A* possesses a bundle of primary social goods that is assigned a greater number than what is assigned to *B's* bundle of primary social goods.

An immediate question is whether such indexing of primary social goods is actually possible. This is called the "indexing problem."[10] I will simply note that Rawls quite frequently sidestep this issue by using income and wealth as "first approximation(s)" for the purpose of applying the difference principle[11] and later relies on such simplifying assumptions when he discusses several illustrative

[10] The problem is not easy. Let $X$ be the set of all possible bundles of primary social goods. A generic element $x \in X$ will be a bundle that contains varying degrees of rights, liberties, opportunities, income, and wealth. In order for it to be possible to index (i.e. assign numbers to) different bundles of primary social goods in such a way that allows us to make interpersonal comparisons regarding who is better off than whom on the basis of such indices, two things must be satisfied:

(a) There must be an underlying order (which is complete and transitive) on $X$.
(b) Every individual must be able to uniformly order different bundles of primary social goods in $X$ according to this underlying order.

However, these two conditions will very likely conflict with the idea that different individuals will very likely pursue very different conceptions of a good life and will, thereby, order the different bundles of primary social goods differently.

This might be better illustrated with a specific example. Let $x, y \in X$ be two bundles of primary social goods. Let $x$ be a bundle that contains more liberty but less income than $y$. Suppose that Alice wishes to be an academic, while Bob wishes to be a corporate lawyer. Based on their different plans of life, it is very likely that Alice and Bob will order the two bundles of primary social goods differently; suppose that, based on their different plans of life, Alice prefers $x$ to $y$, while Bob prefers $y$ to $x$. Note that this is a perfectly reasonable situation; a bundle of primary social goods that is best for the purpose of a certain plan of life might not be best for the purpose of another plan of life. In this scenario, there is no way to assign numbers to the two bundles, $x$ and $y$, that would make it possible for both individuals to uniformly judge that a given bundle of primary social goods is more valuable than another bundle of primary social goods if and only if the former is assigned a higher number than the latter. If we assign a higher number to bundle $x$, then this will contradict Bob's valuations of the two bundles; if we assign a higher number to bundle $y$, then this will contradict Alice's valuations of the two bundles; if we assign the same number to both $x$ and $y$, then this will respect neither individual's valuations. Roemer summarizes the problem as follows:

"But if we wish to consider [Rawls's] full theory, with a variety of primary goods, then we must conclude that the following three claims are together inconsistent:

(1) There is a single index of primary goods,
(2) All plans of life go better (or weakly better) with more primary goods, as measured by the index, and
(3) Plans of life vary sufficiently that individuals will not order bundles of primary goods in the same way." (Roemer, forthcoming)

The three claims are jointly inconsistent; there is no way to make them all true given that we assume that there are both multiple primary goods and sufficiently different life plans. However, rejecting any of the three claims will go against Rawls's entire project. By rejecting either (1) or (2), primary social goods would no longer be able to perform the kind of interpersonal comparisons that Rawls hoped it to perform. This, in turn, would render the difference principle practically inapplicable, as, without being able to make interpersonal comparisons in a meaningful way, there is no way for us to identify which group in society would be the worst-off group, whose expectations the difference principle aims to maximize. Rejecting (3) would be tantamount to rejecting Rawls's *political liberalism* – one of the key assumptions of which is that "it is a permanent feature of the public culture of democracy" that "under the political and social conditions secured by the basic rights and liberties of free institutions, a diversity of conflicting and irreconcilable– and what's more, reasonable – comprehensive doctrines will come about and persist if such diversity does not already obtain." (Rawls 1993/2005: 36)

[11] "The second principle applies, in the first approximation, to the distribution of income and wealth and to the design of organizations that make use of differences in authority and responsibility." (Rawls 1971/1999, 53).

examples.[12] A similar move has been made by other scholars when they discuss about Rawls.[13] Hence, from now on, I will follow Rawls, and simply assume that the indexing problem of primary social goods is adequately solved by regarding each individual's wealth levels as proxies for the amount of primary social goods he/she enjoys.

## 4 Formal Characterization of Utilitarianism and the Difference Principle

Now, we wish to compare Rawls's difference principle and utilitarianism in terms of how each conception of justice better secures the equal worth of the basic rights and liberties of each member of society. Let us try to do this a little more precisely. Let $N=\{1,\ldots,n\}$ be the set of $n$ individuals who are members of a given society, and let $X \subseteq \mathbb{R}^n$ be the set of all feasible distributions of monetary wealth, which, for our purpose, is the single primary social good that serves as a "first approximation" for the purpose of comparing different bundles of primary social goods. So, a typical element $\boldsymbol{x} \in X$ is going to be a vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$, where each component $x_i \in \mathbb{R}$ represents the amount of wealth distributed to individual $i$. Let $u_i\colon X \to \mathbb{R}$ be individual $i$*'s* utility function which transforms wealth into welfare that is *unit* comparable.[14] So, given a distribution of monetary wealth $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, the notation, $u_i(\boldsymbol{x}) = u_i(x_1, \ldots, x_n)$, represents individual $i$'s welfare level at that specific distribution. We follow Rawls and assume that individuals are *mutually disinterested* (Rawls 1971/1999, 12). That is, we assume that each individual cares only about his/her wealth levels and that the wealth levels of other individuals do not 'feed into' each individual's utility function. Hence, we may simply write $u_i(x_1, \ldots, x_n)$ as $u_i(x_i)$.

A social welfare function $W\colon X \to \mathbb{R}$ is designed to represent a society's ranking of different distributions in $X$ such that distribution $\boldsymbol{x} \in X$ is socially preferred to distribution $\boldsymbol{y} \in X$ if and only if $W(\boldsymbol{x}) > W(\boldsymbol{y})$.

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

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

[12] See, for example, the examples he uses in section 12 of Rawls (1971/1999) and explanation of figure 1 in Rawls (2001).

[13] For instance, when examining Rawls's arguments, Brian Barry writes: "For the present purpose I shall take for granted that "better off" means the same as "having more income."" (Barry 1989, 229).

[14] This is the minimum informational requirement that makes utilitarianism technical sense. More formally, let $U^1 = (u_1^1, \ldots, u_n^n)$ and $U^2 = (u_1^2, \ldots, u_n^2)$ be any two profiles of utility functions. We say that utility (or welfare) is unit comparable if and only if the social orderings induced by $U^1$ and $U^2$ are the same if there exist $\alpha > 0$ and $\beta_1, \ldots, \beta_n \in \mathbb{R}$ such that for all $i \in N$ and all $x \in X$, $u_i^2(x) = \alpha u_i^1(x) + \beta_i$. See Gaertner (2009, *A Primer in Social Choice Theory*, 124); Mongin and Claude (1998) Roemer (1996, section 1.1) Bossert and Weymark (2004) for different informational requirements of individual utility functions for different theoretical purposes.

That is, according to the utilitarian social welfare function, a distribution $\boldsymbol{x} \in X$ is socially preferred to another distribution $\boldsymbol{y} \in X$ if and only if the total sum of individual welfare generated by $\boldsymbol{x} \in X$ is greater than the total sum of individual welfare generated by $\boldsymbol{y} \in X$—i.e. if and only if $U(\boldsymbol{x}) > U(\boldsymbol{y}) \Leftrightarrow \sum_{i=1}^{n} u_i(x_i) > \sum_{i=1}^{n} u_i(y_i)$.

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

$$R(\boldsymbol{x}) = \min\{x_1, \ldots, x_n\}.$$

So, according to the Rawlsian social welfare function, a distribution $\boldsymbol{x} \in X$ is socially preferred to another distribution $\boldsymbol{y} \in X$ if and only if the person who receives the lowest bundle of primary social goods (viz. wealth) under distribution $\boldsymbol{x} = (x_1, \ldots, x_n)$ has more wealth than the person who receives the lowest bundle of primary social goods under distribution $\boldsymbol{y} = (y_1, \ldots, y_n)$. In other words, according to the Rawlsian social welfare function, one distribution is better than another distribution if and only if the expectations measured in terms of the index of primary social goods one enjoys—which, in our present case, is simply the amount of wealth one enjoys—of the least advantaged person under the former distribution is greater than that of the least advantaged person under the latter distribution.

With these two social welfare functions, we can now formally define the type of distributions both utilitarianism as well as Rawls's difference principle would respectively choose for any given distributional problem. Given the set of all feasible social distributions, $X$, the specific distributions that would be chosen by utilitarianism would be the solutions to the following maximization problem:

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

Similarly, Given the set of all feasible social distributions, $X$, the specific distributions that would be chosen by Rawls's difference principle would be the solutions to the following maximization problem:

$$\max_{\boldsymbol{x} \in X} \min\{x_1, \ldots, x_n\}.$$

Note how utilitarianism and the difference principle use different information when deciding the specific distribution for a given distributional problem. Utilitarianism is concerned with the *welfare levels* different distributions of wealth generate for each individual and tries to choose the distribution that maximizes the total sum of individual welfare in society. In contrast, the difference principle is concerned with the *amount of primary social goods*—i.e. *wealth levels*—of each individual and tries to choose the distribution under which the individual with the lowest wealth level is maximized.[15]

[15] Note how this is importantly different from the usual way normative welfare economists generally characterize either the *maximin* or the *leximin* principle. See Moulin (2003, Chapter 3); D'Aspremont (2010, section 2.2.3.); Arrow et al. (2002, 2010); Hammond (1976, Section 6).

Since utilitarianism is concerned with each individual's welfare levels, and not simply his/her bundle of primary social goods, and, since each individual is assumed to transform his/her bundle of primary social goods into his/her welfare from his/her specific utility function, for any given distributional problem, we will not be able to compare utilitarianism and the difference principle in terms of their respective distributional consequences before we know the specific way in which each individual's utility function transforms wealth into welfare. To know this, we would need to find a non-arbitrary way to characterize each individual's utility function.

## 5 Rawls's Characterization of Individual Utility Functions

When discussing the type of society to which the difference principle applies, Rawls explains that he

> … shall assume that everyone has physical needs and psychological capacities *within the normal range*, so that questions of health care and mental capacity do not arise. … The first problem of justice concerns the relations among those who in the everyday course of things are full and active participants in society and directly or indirectly associated together over the whole span of their life. Thus the difference principle is to apply to citizens engaged in social cooperation; if the principle fails for this case, it would seem to fail in general. (Rawls 1971/1999, 83–84, emphasis added)

Let us call this the "normality assumption." The normality assumption dictates that the physical and mental capabilities of each individual in our target society (to which Rawls intends his principles of justice to apply) are all *within the normal range* which allows each individual to be a full and active participant of social cooperation. The most important implication of the normality assumption for our current discussion is that we need not consider issues of disabled or handicapped people who are generally *poor translators* of wealth-to-welfare. An example of such a poor translator would be Sen's "cripple" who "gets half the utility that the pleasure-wizard person … does from any given level of income." (Sen 1979, 203) Rawls had deliberately tried to avoid this issue by restricting his theoretical focus to situations in which "questions of health care and mental capacity do not arise."

Based on such normality assumption, Rawls presents what he conceives to be a typical utility function of a "normal" person. (See Fig. 1 in the next page.)

Figure 1 is taken from page 108 of Rawls's *Justice as Fairness—A Restatement.* In the picture, the horizontal axis measure the amount of primary social goods, and the vertical axis measure the amount of utility/welfare the individual enjoys. Points "A", "B", and "G" represent the different amounts of primary social goods guaranteed by alternate social arrangements that have different "basic structures." Point "G" is what Rawls calls the "guaranteeable level." G is the amount of primary social goods one expects to receive when one applies the maximin rule in the original position; it is supposed to denote the amount of primary social goods that best secures the equal worth of the basic rights/liberties one can enjoy.

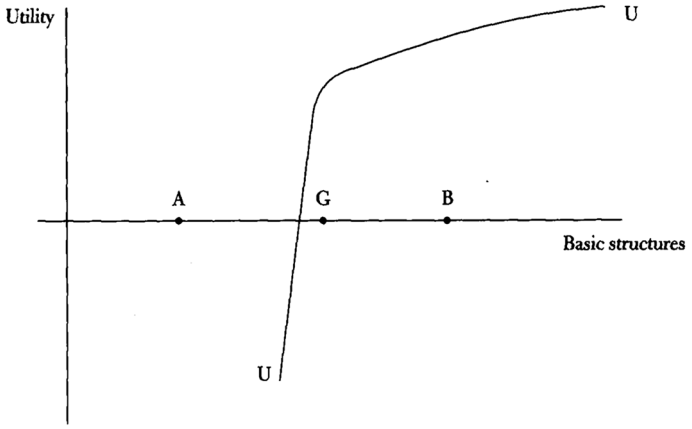


**Fig. 1** Rawls's utility function

The following summarizes the general characteristics of Rawls's utility function,

(C1) The utility function is *strictly increasing* in primary social goods.
(C2) There exists a *reference point* (G in the above figure) below which and above which the slope and curvature of the utility function abruptly changes.
(C3) The slope *below* the reference point is *linear* and *steeper* than the slope above the reference point.
(C4) The slope *above* the reference point is *strictly concave* and *flatter* than the slope below the reference point.

To be clear, the main purpose of introducing such a utility function at that point of the book was to show how utilitarianism, once individual utility functions are suitably characterized, can support justice as fairness in an overlapping consensus (Rawls 2001, 108–109). In 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)

Rawls further uses the particular shape of individual utility function depicted in Fig. 1 to explain why it would be rational for the representative parties in the original position to adopt *the maximin rule*[16] to guide their decision process, which results in their choosing justice as fairness. According to Rawls, there are three basic conditions under which it would be rational to adopt the maximin rule,

(a) … the first condition is that the parties have no reliable basis for estimating the probabilities of the possible social circumstances that affect the fundamental interests of the persons they represent. …

[16] The maximin rule is different from what welfare economists call "the maximin principle." The welfare economist's maximin principle a principle used to order different social distributions; it says that a distribution $x$ is socially preferred to another distribution $y$ if and only if the lowest welfare generated in the former is greater than that of the latter. The maximin rule is a decision procedure that one could follow when one faces risk or uncertainty; it tells the individual to choose the option that will provide the best outcome in the worst possible circumstances.

(b) … it must be rational for the parties as trustees not to be much concerned for what might be gained above what can be guaranteed [by the maximin rule.] Let's call this best–worst outcome the "guaranteeable level." The second condition obtains, then, when the guaranteeable level is itself quite satisfactory. …

(c) … the third condition is that the worst outcomes of all the other alternatives are significantly below the guaranteeable level. … (Rawls 2001, 98)

According to Rawls, the original position meets condition (a) by its very definition and the utility function depicted in Fig. 1 explains why the other two conditions [i.e. (b) and (c)] obtain for the representative parties in the original position,

> …to the right of the bend, at point G in the figure, everyone's utility curves become suddenly quite flat. This explains why the parties as citizens' representatives are not much concerned with outcomes superior to the guaranteeable level, and hence the second condition of the maximin rule holds. To the left of the bend everyone's utility curve falls precipitously, and hence the third condition of the maximin rule also holds. This explains why the parties must reject alternatives that fail to guarantee the basic equal liberties. (Rawls 2001, 108)

What Rawls wanted to argue was that once individual utility functions are suitably characterized (as depicted in Fig. 1), not only would the representative parties in the original position find it rational to adopt the maximin rule (which leads them to choose justice as fairness over utilitarianism), but even utilitarian reasoning will lead them to endorse justice as fairness. For the remainder of the paper, I will try to demonstrate why this argument is flawed.

## 6 Distributional Consequences of Utilitarianism and the Difference Principle

### 6.1 The Model

Consider a liberal democratic society in which everyone is formally/constitutionally guaranteed a full set of basic rights and liberties—i.e. the type of basic rights and liberties guaranteed by the principle of maximum equal basic liberties. We consider two representative groups living in our society. In line with Rawls, we assume as a first approximation that there is a single primary social good called 'wealth' that individuals could transform into utility or welfare. Let $\bar{W} \in \mathbb{R}_+$ (i.e. $\bar{W}$ is a non-negative real number) be the amount of social wealth (which may vary under different social circumstances) that can be distributed to the two representative groups in our society.

Some people may criticize that this does not accurately represent the type of situation that Rawls envisions, as it simply assumes that a fixed amount of social wealth is given exogenously. The reason why this may be problematic is because assuming a fixed amount of social wealth may seem to get rid of incentive issues, which Rawls deemed important in justifying his difference principle. However, one must clearly understand the specific context in which Rawls relied on incentive issues to justify his

difference principle. Rawls relied on incentive issues to argue for the superiority of the difference principle *over strict egalitarianism* (under which social wealth is distributed in a perfectly equal manner) *not utilitarianism.* The point was that although the difference principle allows inequalities, such inequalities will actually work in favor of the least advantaged group, as the better prospects that are achievable by allowing such inequalities "act as incentives [for the entrepreneurs] so that the economic process is more efficient, innovation proceeds at a faster pace" (Rawls 1971/1999, 68) that would improve the situations of the least advantaged group as well.

The main point is that incentives considerations would favor the difference principle when the difference principle is compared to strict egalitarianism; however, incentives considerations will *not* favor the difference principle when it is compared to utilitarianism. A utilitarian society will generally provide an even better prospect for the entrepreneurs than a society regulated by the difference principle, as the upper bound of economic benefits that the entrepreneurs is allowed to accumulate would not be constrained by any considerations to benefit the least advantaged group in their society. By Rawls's same logic, this would serve as an incentive for the entrepreneurs to work even harder to achieve more than what they would have done in a society regulated by the difference principle. So, incentive considerations would give reasons to support utilitarianism rather than the difference principle.[17]

Rawls's worry was not that utilitarianism may fail to give proper incentives for people; his worry was, rather, that the better incentives utilitarianism provides may come with a significant cost—namely, that it may generate extreme inequalities such

[17] Let me illustrate this with a simple example. Suppose there are two representative individuals: 1 and 2. Each individual is endowed with 1 unit of labor, which they can feely use to earn income. For $i = 1, 2$, let $L_i$ denote the amount of labor spent by individual $i$, and let $Y_i$ denote individual $i$'s earned income. Let us assume that each individual's welfare is determined both by his/her income and spent labor; specifically, suppose for $i = 1, 2$, $U_i(L_i, Y_i) = Y_i - L_i^2$. For illustrative purposes, we will assume that individual 1 is *twice as more productive* than individual 2; that is, for every amount of spent labor, individual 1 earns twice as much income (individual 1 is the entrepreneur) as what individual 2 would earn by spending the same amount of labor. Specifically, suppose $Y_1 = 2L_1$ while $Y_2 = L_2$. Here, we will impose an income tax $t \in [0, 1]$ to individual 1 (the entrepreneur)'s earned income, and, then, transfer this tax to individual 2 to compensate for individual 2's relative productive disadvantage. Hence, after expending $L_1$ and $L_2$ amount of labor, individual 1 earns $Y_1 = 2L_1(1 - t)$ amount of income and individual 2 earns $Y_2 = L_2 + 2L_1 t$ amount of income. Individual 1 then solves:

$$\max_{L_1 \in [0,1]} U_1(L_1, Y_1) = \max_{L_1 \in [0,1]} Y_1 - L_1^2 = \max_{L_1 \in [0,1]} 2L_1(1 - t) - L_1^2,$$

while individual 2 solves:

$$\max_{L_2 \in [0,1]} U_2(L_2, Y_2) = \max_{L_2 \in [0,1]} Y_2 - L_2^2 = \max_{L_2 \in [0,1]} (L_2 + 2L_1 t) - L_2^2.$$

By taking the first-order conditions, we can compute each individual's optimal labor amounts, which turns out to be: $L_1^* = 1 - t$ and $L_2^* = \frac{1}{2}$. From this, we can derive each individual's *indirect utility function* as a function of the tax rate $t$: $U_1^* = (1 - t)^2$ and $U_2^*(t) = -2t^2 + 2t + \frac{1}{4}$. With these indirect utility functions, utilitarianism solves:

$$\max_{t \in [0,1]} U_1^*(t) + U_2^*(t) = \max_{t \in [0,1]} (1 - t)^2 - 2t^2 + 2t + \frac{1}{4} = \max_{t \in [0,1]} -t^2 + \frac{3}{4}.$$

That is, utilitarianism chooses an income tax rate that maximizes total social welfare between the two individuals. It can be easily seen that the utilitarian social welfare function is maximized when $t_U^* = 0$,

that the least advantaged group in society may not properly secure their basic rights/liberties and their proper worth thereof. So, focusing on societies that have a fixed amount of wealth to distribute actually takes away one advantage that utilitarianism has over the difference principle; namely, incentive considerations. In this sense, our model is actually handicapping utilitarianism. What would be surprising is if utilitarianism turned out to be superior over the difference principle even when incentive issues disfavoring the difference principle is nullified.

I will now follow Rawls's convention[18] and call the two representative groups MAG (more advantaged group) and LAG (less advantaged group.) I assume that members of both MAG and LAG possess physical and mental capabilities *within the normal range*. Let $u_M\colon \mathbb{R}_+ \to \mathbb{R}$ be the utility function of MAG, and let $u_L\colon \mathbb{R}_+ \to \mathbb{R}$ be the utility function of LAG. We assume that both utility functions conform to all of the general characteristics [i.e. (C1)–(C4)] of Rawls's utility function that we have seen in the previous section, with one wrinkle; we assume that the reference points of MAG's and LAG's utility functions are different. Specifically, let $r_M \in \mathbb{R}$ be the reference point of MAG, and let $r_L \in \mathbb{R}$ be the reference point of LAG. We will assume that $0 < r_M < r_L$.

Remember that the reference point of each group's utility function is supposed to denote the amount of primary social goods (in this case, wealth) each member of the group needs to fully enjoy *the equal worth* of the full set of basic rights and liberties that is formally guaranteed by our model society meeting Rawls's first principle of justice. Hence, the assumption $0 < r_M < r_L$ simply means that LAG, given its members' specific ends as well as their relative disadvantage in their overall natural talents and social circumstances, needs a greater bundle of social primary goods, $r_L$, to equally enjoy the worth of their basic liberties and freedoms that members of MAG

Footnote 17 (continued)
(call this the utilitarian tax rate) and, hence, utilitarianism chooses a *zero* tax rate on earned income. The reason for this is obvious; utilitarianism chooses a zero income tax rate because this incentivizes individual 1 (the entrepreneur) to work harder, which leads society to maximize total social welfare. Compare this to Rawls's difference principle, which solves:

$$\max_{t\in[0,1]} \min\{Y_1, Y_2\} = \max_{t\in[0,1]} \min\{2L_1(1-t), L_2 + 2L_1 t\} = \max_{t\in[0,1]} \min\left\{2(1-t)^2, \frac{1}{2} + 2t(1-t)\right\},$$

where we use the fact that the optimal labor amounts are $L_1^* = 1 - t$ and $L_2^* = \frac{1}{2}$. This problem is reduced to finding a tax rate $t_R^*$ that equalizes $2(1-t)^2$ and $\frac{1}{2} + 2t(1-t)$. (Here, we are using the result of Proposition 1, which we will later see.) As a result, the Rawlsian tax rate is $t_R^* = \frac{1}{4}\left(3 - \sqrt{3}\right) \approx 0.32$. So, Rawls's difference principle chooses an income tax rate of approximately 32%. From this, we are now able to compute how hard individual 1 (the entrepreneur) would work under the two alternate systems, and how much total social wealth, as a result, would be produced. Under the utilitarian tax rate (i.e. $t_U^* = 0$), individual 1 spends his/her entire endowed labor to earn income, while under the Rawlsian tax rate (i.e. $t_R^* \approx 0.32$), individual 1 spends only about $\frac{2}{3}$ of his/her endowed labor. As a result, total social wealth produced under the utilitarian tax rate is 2.5, while total social wealth produced under the Rawlsian tax rate is 1.86. In short, if one wishes to compare utilitarianism and Rawls's difference principle in terms of which distributive principle better incentivizes entrepreneurs (i.e. those with high productive ability) to work harder to increase the total social pie, utilitarianism will generally be the winner. □

[18] See Rawls (2001, 62–63).

fully enjoy at wealth level $r_M$. This formally models that, between MAG and LAG, the lesser advantaged group is LAG.[19]

The fact that people's utility functions within the normal range can have different reference points is not only consonant with Rawls's overall project, but it is actually what underlies the very design of the veil of ignorance. Remember that the main purpose for introducing the veil of ignorance is to conceal the original contracting parties from knowing their relative social, economic, and natural advantages and/or disadvantages, which Rawls thought to be morally arbitrary. This presupposes that, even within the normal range, there exist people who are advantaged or disadvantaged relative to other people. We need a way to represent this relative advantage and/or disadvantage within our model. And, the most sensible way to represent this while being faithful to Rawls's own utility function is to assume that the utility functions of MAG and LAG have different reference points. Note that Rawls explains that "[i]t is not implied that those with the same index [of primary social goods] have equal well-being, all things considered; for their ends are generally different and many other factors are relevant." (Rawls 1974b, 643) If we agree to use Rawls's utility function and further assume that everybody's utility function has exactly the same reference point, the same bundle of primary social goods would necessarily generate equal well-being, contradicting what Rawls had just explained. Hence, there is a sense in which assigning *different* reference points to MAG's and LAG's utility functions is not simply a matter of theoretical discretion, but a logical implication of Rawls's own assumptions taken together.

The following summarizes the main assumptions of both MAG and LAG's utility functions:

(A1) $u_M(x), u_L(x)$ are unit comparable.
(A2) $u_M(x), u_L(x)$ are differentiable in the left and right regions of their respective reference points, i.e. $u_M(x)$ is differentiable for all $x \in \mathbb{R}_+ \backslash \{r_M\}$, and $u_L(x)$ is differentiable for all $x \in \mathbb{R}_+ \backslash \{r_L\}$.
(A3) For all $x \in \mathbb{R}_+ \backslash \{r_M\}, u'_M(x) > 0$, and for all $x \in \mathbb{R}_+ \backslash \{r_L\}, u'_L(x) > 0$, i.e. $u_M, u_L$ are strictly increasing in wealth.
(A4) Both $u_M$ and $u_L$ are linear below, strictly concave above their respective reference points, and the slopes of $u_M$ and $u_L$ are steeper below the reference points than above.

[19] The utility functions of different people within the normal range can have different reference points for variety of reasons. Some people might need additional resources to fully realize their rational life plans (and, hence, enjoy the equal worth of the basic rights and liberties that are formally guaranteed to them by the constitution) simply because, despite having human capacities within the normal range, they are not gifted with splendid natural talents or born from prestigious social classes. Others might need additional resources because they happen to affirm a particular religion that requires them to travel to the holy land multiple times a year to worship its gods; to them, enjoying the worth of religious freedom requires them to have additional resources for travel. These are just a couple of examples. Whatever the particular reasons that make people need greater bundles of primary social goods, it is these sorts of contingent facts that the veil of ignorance was initially designed to conceal from the original contracting parties.

(A5) To simplify our analysis, let $u_M(r_M + \Delta) = u_L(r_L + \Delta)$ and $u_M(r_M - \Delta) = u_L(r_L - \Delta)$ for $\Delta > 0$, i.e. both MAG and LAG receive the same utility for wealth levels that are at the same distances from their respective reference points. Or to put it another way, $u_L$ is obtained by moving $u_M$ an increment of $(r_L - r_M)$ to the right, i.e. $u_L(x) = u_M(x - (r_L - r_M))$.

These assumptions are meant to ensure that $u_M$ and $u_L$ conform to the general characterizations of individual utility functions described by Rawls that we have previously seen.[20] The differentiability assumptions are added to allow us to use calculus techniques to get specific answers to distributional problems. The last assumption is, strictly speaking, not completely needed and can be significantly weakened[21]; however, it will greatly simplify our analysis, and, can also be interpreted as an implication of Rawls's normality assumption. This completes the setup of our model.

Before moving on, I would like to point out that focusing on a simple model of a society consisting of two representative groups is an exercise that Rawls himself invokes quite frequently throughout his works.[22] So, the type of exercise we are trying to conduct is a method that Rawls himself frequently employs.

### 6.2 Results of the Model

Let us now derive the specific distributional consequences of utilitarianism and justice as fairness (more specifically, the difference principle) of our model. Before doing this, I would like to emphasize that the only reason why we are considering a liberal democratic society that meets the principle of equal basic liberties (i.e. Rawls's first principle) is to meet the preconditions that allow us to apply Rawls' difference principle. Such an assumption is not meant to restrict the distributional consequences of utilitarianism. All the distributional results of utilitarianism that we will soon derive will remain intact even if we dropped this assumption. So, what we are really comparing is the distributional consequences of utilitarianism-full-stop (not utilitarianism restricted by Rawls's first principle) and Rawls's difference principle. In doing so, we will vary the social wealth level $\bar{W} \in \mathbb{R}_+$. 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 wealth. Here is our first result that will be used frequently throughout our analysis.

[20] I would like to note that, technically speaking, none of the formal results proved in the next subsection depends on the fact that there exists a point in each group's utility function that is continuous but not differentiable. All of the formal results proved in the next subsection will follow even if we generally assumed that each group's utility function is strictly concave throughout. Hence, most of the specific assumptions assumed here are stated mainly to be faithful to Rawls's characterization of individual utility functions; they are not assumed to make it easier to derive my desired results.

[21] Specifically, it can be weakened to, $\forall x > 0, \min\{u'_M(r_M - x), u'_L(r_L - x)\} > max\{u'_M(r_M + x), u'_L(r_L + x)\}$.

[22] See Rawls (1971/1999, sections 12–13); Rawls (2001, 62–63).

**Proposition 1** *For any* $\bar{W} \geq 0$*, the difference principle prescribes* $(x_M, x_L) = \left(\frac{\bar{W}}{2}, \frac{\bar{W}}{2}\right)$.[23]

Proposition 1 shows that the difference principle will always divide social wealth into half and distribute it equally to each individual. With this in mind, let us consider the distributional consequences of utilitarianism under different levels of social wealth.

One thing to remember is that Rawls has stated that his principles of justice, which include the difference principle, only apply to situations of *moderate scarcity* (Rawls 1971/1999, section 22). Of course, Rawls remained vague on what he exactly meant by the condition of moderate scarcity. He explains that it refers to a situation in which "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." He also notes that such a condition is "understood to cover a wide range of situations." (Rawls 1971/1999, 110) However, a society whose social wealth level allows everybody to just barely enjoy the equal worth of his/her basic rights and liberties would seem to fit into this wide range of situations that exemplify conditions of moderate scarcity. In our model, this situation can be represented by the social wealth level $\bar{W} = r_M + r_L$.

**Proposition 2** *Suppose* $\bar{W} = r_M + r_L$*. Then, utilitarianism prescribes* $(x_M, x_L) = (r_M, r_L)$*, while the difference principle prescribes* $(x_M, x_L) = \left(\frac{r_M + r_L}{2}, \frac{r_M + r_L}{2}\right)$.[24]

Proposition 2 is important. The situation that is assumed is one in which there is enough resources (albeit barely) to secure the worth of *everybody's* basic rights and liberties. Society's resource level is scarce; but not extremely so. The situation exemplifies that of moderate scarcity—one to which Rawls's difference principle should obviously apply.

Yet, compare the distributional consequences of utilitarianism and the difference principle. Utilitarianism prescribes a distribution that secures the equal worth of basic rights and liberties for *everybody*. By contrast, the difference principle prescribes to distribute $\frac{r_M + r_L}{2}$ amount of wealth to each group equally. Note that $\frac{r_M + r_L}{2}$ is an amount that is greater than $r_M$, but less than $r_L$. This means that not only does the difference principle only secure the worth of basic rights and liberties of one representative group even when it was practically feasible to secure the equal worth of basic rights and liberties for every group, it secures the worth of basic rights and liberties of MAG, who, by assumption, is the more advantaged group, rather than LAG, who, by assumption, is the lesser advantaged group, and, in doing so, the difference principle gives MAG more resources than what is actually necessary to secure the fair worth of its members' basic rights and liberties at the very expense

[23] See "Appendix" for Proof.
[24] See "Appendix" for Proof.

of the members of LAG. In other words, the distribution that the difference principle prescribes in our two group society under conditions of moderate scarcity goes against the very purpose of why the difference principle was initially proposed and designed in the first place; namely, to protect the least advantaged group in society.

For a devoted Rawlisan, the result seems to be something that would be hard to swallow. Hence, one might wonder what, in intuitive terms, has gone wrong with Rawls's difference principle at this point. The problem stems from Rawls's complete reliance on bundles of resources (i.e. index of primary social goods) as a measure a person's advantage. By focusing solely on maximizing the size of the bundle of primary social goods (i.e. wealth) that the person who receives the lowest bundle gets, the difference principle is completely blind to the issue of what different bundles of primary social goods *can actually do for different people*; based on an individual's specific needs, the specific bundle distributed by the difference principle might still be insufficient.

In contrast, utilitarianism *does* care about what different bundles of primary social goods do for different people; it cares about how these different bundles of resources translate to people's welfare and attempts to maximize their total sum. Given Rawls's own characterizations of individual utility functions, this results in utilitarianism's attempt to secure each group's reference point whenever society's resource situation allows it. In this way, Proposition 2 can be simply seen as a logical corollary of what Amartya Sen has earlier noted as Rawls's resource "fetishism." 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 1979: 215–216)

This kind of perversion in which the difference principle goes against its very own *raison d'être* is not limited to situations of moderate scarcity as the next proposition demonstrates.

**Proposition 3** *Suppose* $r_M + r_L < \bar{W} < 2r_L$. *Then, utilitarianism prescribes* $(x_M, x_L) = \left(\frac{\bar{W}+r_M-r_L}{2}, \frac{\bar{W}-r_M+r_L}{2}\right)$, *while the difference principle prescribes* $(x_M, x_L) = \left(\frac{\bar{W}}{2}, \frac{\bar{W}}{2}\right)$.[25]

[25] See "Appendix" for Proof.

Proposition 3 concerns a situation in which there is more than enough social wealth to secure everybody's basic rights and liberties, but not so much social wealth to secure the worth of everybody's basic rights and liberties *if everybody needed as much resources as LAG*. Society's resource level is abundant; but, not extremely so. So, the situation may be thought of as exemplifying conditions of *moderate abundance*.

The way utilitarianism distributes the available resources accords very well with our basic moral intuitions. It prescribes, first, to secure the equal worth of everybody's basic rights and liberties by giving everybody the amount of resources needed to satisfy his/her specific reference point. Afterwards, utilitarianism prescribes to divide and distribute the remaining social wealth—i.e. what is left after everybody secures the equal worth of his/her basic rights and liberties—equally to everybody. The claim that utilitarianism will, for the sake of maximizing aggregate social welfare, result in vast inequalities, which would likely put the least advantaged group below what Rawls calls 'the guranteeable level' is demonstrably false in our model.

Now, compare this with the distributional consequences of the difference principle. The specific distribution the difference principle prescribes is $\left(\frac{\bar{W}}{2}, \frac{\bar{W}}{2}\right)$, again, an equal distribution of social wealth. As $r_M + r_L < \bar{W} < 2r_L$ by assumption, we have $r_M < \frac{r_M + r_L}{2} < \frac{\bar{W}}{2} < r_L$, which implies that only MAG, and not LAG, will be able to receive enough resources to secure the worth of his/her basic rights and liberties. Again, this contradicts the difference principle's very own *raison d'être*.

The only situation in which the difference principle will be able to secure the equal worth of *everybody's* basic rights and liberties would be when the available social wealth $\bar{W}$ *exceeds* $2r_L$. In other words, the only way the difference principle will be able to secure the equal worth of *everybody's* basic rights and liberties is for there to be *more than* a moderate abundance of social wealth, which goes far beyond the moderate scarcity condition that Rawls himself assumes.

This is already a major failure on part of the difference principle. However, I would like to throw a final blow. Suppose social wealth levels are extremely abundant—that is, suppose $\bar{W} > 2r_L$. Under these conditions, both utilitarianism and the difference principle will still respectively prescribe $\left(\frac{\bar{W} + r_M - r_L}{2}, \frac{\bar{W} + r_M - r_L}{2}\right)$ and $\left(\frac{\bar{W}}{2}, \frac{\bar{W}}{2}\right)$. Since $\bar{W} > 2r_L$, we have $\frac{\bar{W}}{2} > r_L > r_M$, and, hence, both MAG and LAG will have more than enough resources to secure the equal worth of their basic rights and liberties under both distributive principles. So, wouldn't this suggest a tie, at least, under the circumstances that exceed moderate abundance?

Not quite. This is because, within our model, not only would the total amount of aggregate social welfare generated by utilitarianism be greater than what would be generated by the difference principle, the level of welfare enjoyed by LAG, in particular, will always be greater under utilitarianism than under the difference principle whenever the levels of social wealth is greater than or equal to moderate scarcity.

**Proposition 4** *Suppose* $\bar{W} \geq r_M + r_L$. *Then, the following two claims are true,*

(a) *The total social welfare generated by utilitarianism is strictly greater than the total social welfare generated by the difference principle; and*
(b) *The welfare level that LAG enjoys is always greater under utilitarianism than under the difference principle.*[26]

We know that utilitarianism strives to maximize total social welfare. Rawls's worry was not that utilitarianism might achieve a smaller total social welfare than it otherwise could have, but that the maximization of social welfare might be achieved at the expense of sacrificing the basic rights, liberties, and welfare of the least advantaged group in society. This is the essence of Rawls's criticism that utilitarianism does not take the separateness of persons seriously (as well as all of his other substantive criticisms based on concerns for stains of commitment, stability, and self-respect.) What Proposition 4 shows is that, whenever conditions are more favorable than or equal to moderate scarcity, the increase in total social welfare that is achieved under utilitarianism is not achieved by any sacrifice of the least advantaged group in society. Since we are only concerned with social conditions that are more favorable than or equal to moderate scarcity, from the perspective of LAG, who, in our model, is supposed to represent the lesser advantaged group in society, Proposition 4 shows us that utilitarianism *dominates* the difference principle, not merely in a single way, but in two different ways; (i) not only does utilitarianism dominate the difference principle in terms of securing the worth of LAG's basic rights and liberties, but (ii) it also dominates utilitarianism in terms of allowing LAG, the lesser advantaged group between the two, to enjoy a greater welfare level.

Before ending this section, I would like to clarify two things to avoid unnecessary confusion. First, I will state without proof that none of the formal results depend on the size of the two groups (viz. MAG and LAG,) being equal; all of the formal results will go through regardless of how we vary the size of the two groups. Second, as I have already noted, the specific distributional results of utilitarianism in our formal model do not depend on our assumption that our model society is a liberal democratic society that formally satisfies the first principle of justice as fairness (i.e. the principle of equal basic liberties.) All of the distributional results of utilitarianism would be the same without this formal restriction. In other words, I am not considering a restricted form or utilitarianism—that is, something that Rawls called "mixed conceptions" (Rawls 1971/1999, section 49) or "the principle of restricted utility" (Rawls 2001, Section 38). Again, the only purpose for introducing the formal restriction that our model society satisfies the principle of equal basic liberties was to apply the difference principle, as the difference principle would not be applicable unless the principle of equal basic liberties is already satisfied.

[26] See "Appendix" for Proof.

## 7 Defending Utilitarianism from the Original Position

Now, let us go back to Rawls' original position, and consider the decision problem faced by the representative parties.[27] Although the two groups, LAG and MAG, are not themselves parties of the original position, the representative parties in the original position use the distributional consequences generated by utilitarianism and the difference principle with respect to these two groups (which we have derived from our model in the previous section) as a part of their general background knowledge during their reasoning process. Let us remind ourselves of the three basic conditions that, according to Rawls, would make it rational to follow the maximin rule when choosing principles of justice for the basic structure of society:

(a) … the first condition is that the parties have no reliable basis for estimating the probabilities of the possible social circumstances that affect the fundamental interests of the persons they represent. …
(b) … it must be rational for the parties as trustees not to be much concerned for what might be gained above what can be guaranteed [by the maximin rule.] Let's call this best–worst outcome the "guaranteeable level." The second condition obtains, then, when the guaranteeable level is itself quite satisfactory. …
(c) … the third condition is that the worst outcomes of all the other alternatives are significantly below the guaranteeable level. … (Rawls 2001, 98)

Let us follow Rawls and grant that not only do these three conditions obtain in the original position, but they also make it rational for the representative parties to adopt the maximin rule to guide their choices on distributional principles. The maximin rule implies that the representative parties, who are behind the veil of ignorance, which renders them unaware of their real identities, would have to make their choices from the perspective of LAG, who is the lesser advantaged group in our society. The "guaranteeable level" would simply denote the amount of wealth required to secure the equal worth of LAG's basic rights and liberties; it would denote LAG's reference point $r_L$.

Since being guaranteed a wealth level of $r_L$ would allow LAG to enjoy the fair worth of his/her basic rights and liberties, such 'guranteeable level' would be quite satisfactory as condition (b) claims. Condition (b) along with the fundamental importance Rawls puts on securing the equal worth of basic rights and liberties jointly imply that the representative parties would give utmost priority to securing a wealth level of $r_L$ for LAG.

Under conditions of moderate scarcity or above, utilitarianism does guarantee LAG a wealth level of $r_L$ or more. By contrast, the difference principle can only

[27] Since the representative parties in the original position are situated symmetrically, many have claimed that the decision problem that the representative parties face in the original position is essentially a rational choice of a single individual. For this reason, people like Hampton (1980) and Gauthier (1985) have objected that the choice made in the original position cannot really be seen as a genuine mutual agreement or a (social) contract.

guarantee LAG a wealth level of $r_L$ only when society is more than moderately abundant. Whenever the level of social wealth is lower than moderate abundance—particularly, when the level of social wealth is, just as Rawls assumes, moderately scarce—the amount of social wealth distributed to LAG, as we have seen, is *guaranteed to be lower than* $r_L$. One thing to note is that Rawls assumes that, by the veil of ignorance, "information about natural resources, the level of productive techniques, and the like, is also forbidden to [the original contracting parties.]" (Rawls 1974b, 637) So, the representative parties cannot choose the difference principle on the basis of knowing that their society would be more than moderately abundant; they would need to consider the distributional consequences of each principle at lower resource levels (particularly, when the resource level is moderately scarce.) This implies that the worst outcome of the difference principle is, as Rawls fears, "significantly below the guaranteeable level" which condition (c) strongly urges to avoid.

Hence, under all major assumptions that Rawls himself suggests, utilitarianism simply *dominates* the difference principle along with justice as fairness. By invoking such dominance-based reasoning, the representative parties will simply choose utilitarianism over justice as fairness. And, by doing so, they are not relying on any kinds of judgments concerning probability, which is simply what condition (a) requires.

Now, some critics might think that the problem is not with the difference principle itself, but rather with how the difference principle was derived; that is, many critics have thought that the difference principle, despite its intuitive plausibility, cannot be derived from the original position. So, the culprit here, according to these critics, is the original position, not the difference principle, as the difference principle may be justified in alternate ways. For instance, Barry (1989, chapter 6) explains how we can arrive at the difference principle without relying on the original position. The basic thought is to start with a default distribution of strict equality of primary social goods, and, then successively move to more unequal distributions by successively applying Pareto improvements until we reach a point at which no further Pareto improvements are possible. The resulting distribution is the one that would accord with the one prescribed by the difference principle. The moral is that if we use a different theoretical device than the original position, we may very well arrive at the difference principle after all.

This criticism confuses the *explanandum* and the *explanans* of my argument. My point is *not* that the difference principle is implausible *because* it cannot be derived from the original position; rather, my point is that we cannot derive the difference principle from the original position *because* the difference principle, when applied to the index of primary social goods, is implausible *in itself*. The problem stems from the difference principle's failure to recognize each individual's reference point. The difference principle will, hence, distribute primary social goods mechanically without taking into consideration the specific needs of each individual. The result is that people born under fortunate circumstances (i.e. people who have low reference points) will likely receive a bundle of primary social goods that is greater than what they would minimally need to secure the equal worth of their basic rights and liberties, while people born under unfortunate circumstances (i.e. people who have high reference points) will likely receive a bundle of primary social goods that is smaller than what they would minimally need to secure the equal worth of their basic rights

and liberties. If the parties in the original position wish to protect themselves from the likely scenario of being born with a high reference point, their purpose can only be served by rejecting the difference principle and opting for utilitarianism.

## 8 Concluding Remarks

In section 14 of A Theory of Justice, Rawls distinguishes between *pure procedural justice* and *perfect procedural justice*. Perfect procedural justice has both an "independent standard for deciding which outcome is just and a procedure guaranteed to lead to it." (Rawls 1971/1999, 74) "By contrast, pure procedural justice obtains when there is no independent criterion for the right result, instead there is a correct or fair procedure such that the outcome is likewise correct or fair, whatever it is, provided that the procedure has been properly followed." (Rawls 1971/1999, 75) Rawls made it clear that the original position was designed to instantiate pure (as opposed to perfect) procedural justice (Rawls 1971/1999, 118). This means that Rawls would have to accept its results whatever they turn out to be. Our previous discussion shows Rawls's primary considerations—namely, the importance of protecting each individual's fundamental interests by securing the equal worth of his/her basic rights and liberties—provide very strong reasons for the original contracting parties to choose utilitarianism over justice as fairness under Rawls's own assumptions. If this is correct, Rawls has no choice but to accept utilitarianism.

Some critics might think that Rawls could, at this point, resort to his method of "reflective equilibrium" to counter this unwanted conclusion for utilitarianism. I am afraid that this is not a viable move that Rawls could plausibly make. Remember that in order to figure out the best principles of justice, the method of reflective equilibrium requires us to "work from both ends." (Rawls 1971/1999, 18) That is, we start with what Rawls calls our "considered judgments" (such as our judgments that slavery, racial discrimination, and religious intolerance are unjust) and take them as our "provisional fixed points which we presume any conception of justice must fit" (Rawls 1971/1999, ibid.) We then describe an initial contractual situation that will generate a set of principles of justice from plausibly chosen initial conditions that would hopefully accommodate most (if not all) of our considered judgments. Whenever we find discrepancies, Rawls suggests that we go "back and forth, sometimes altering the conditions of the contractual circumstances, at others withdrawing our judgments and conforming them to principle." (Rawls 1971/1999, ibid.) Through this process, Rawls believes that we will eventually arrive at "a description of the initial situation that both expresses reasonable conditions and yields principles which match our considered judgments duly pruned and adjusted." (Rawls 1971/1999, ibid.) This is the state which Rawls calls a "reflective equilibrium."

Now, in order to use the method of reflective equilibrium to revert our conclusion for utilitarianism, utilitarianism would have to contradict at least some of our firmest considered judgments. The problem is: it does not (at least in our current model.) We have seen that in our model, utilitarianism firmly secures the equal worth of everybody's basic rights and liberties whenever society's resource situation is equal to or greater than moderate scarcity. If we take the belief, "justice requires society to

secure the equal worth of everybody's basic rights and liberties whenever the society's resource situation allows it", to be one of our firmly held considered judgments, we can see that utilitarianism accommodates it while the difference principle fails to do so.

Similar remarks can be said to the issue of *relative stability*. Rawls's considerations for strains of commitment, distinction between persons, publicity, stability, and self-respect can all be seen as subsumed under the general considerations for relative stability. The basic thought is that people will likely renege on their agreement for utilitarianism once they find themselves in a disadvantaged position after the veil of ignorance is lifted. This, again, as we have seen, is untrue in our model. Unlike the difference principle, utilitarianism will always secure the equal worth of everybody's basic rights and liberties whenever society's resource situation is equal or greater than moderate scarcity. So, if there is anybody who would be inclined to renege on his/her original agreement, it would be those who agreed to the difference principle, but who have found themselves denied of receiving an adequate amount of primary social goods that would secure the equal worth of their basic rights and liberties. This would inevitably render the difference principle less stable than utilitarianism. In either case, it seems that the method of reflective equilibrium would have to favor utilitarianism over the difference principle in our model.

In order to reject utilitarianism and defend the difference principle via the method of reflective equilibrium, we would have to start with the assumption that utilitarianism is in itself implausible while the difference principle is plausible, and, then adjust all of our other assumptions, accordingly, in such a way that would make it so that we arrive at the difference principle and not utilitarianism. The problem with this approach is that not only does this *beg the question*, but neither the belief that the difference principle is plausible nor the belief that utilitarianism is implausible can be taken to be our *considered* moral judgments.

However, let us consider, for illustrative purposes, what kinds of modifications need to be made to make either (1) the selection of the difference principle more plausible or (2) the selection of utilitarianism less plausible in our model.

Let us consider the first option: making the difference principle more plausible. We have seen that part of what drives the difference principle to generate implausible distributional consequences stems from Rawls's *resource fetishism*. Hence, we may correct this problem by reinterpreting the difference principle as applying to people's welfare levels (just as welfare economists do) and not to people's resource levels (i.e. the index of primary social goods.) Given Rawls's characterization of individual utility functions, doing so would guarantee that each social group meets its reference point whenever society's resource levels are equal or greater than moderate scarcity. However, this would only make the difference principle tie with utilitarianism in its distributional prescriptions. So, the representative parties of the original position will lack any decisive reason to favor one conception of justice over the other.

Now, let us consider the second option: making utilitarianism less plausible. This can be done if we discard Rawls's characterizations of individual utility functions and assume that each individual's utility function is strictly convex (i.e. it has an increasing slope.) This will make utilitarianism generate extreme inequalities: utilitarianism will now distribute all the available social resources to one social group

and give nothing to the other in all possible resource situations. This indeed goes against our basic moral intuitions of fairness. However, once individual utility functions are characterized in this particular way, this has the implication of violating the second necessary condition the fulfillment of which is required to make it rational to apply the 'maximin rule' in the original position. This is so because the representative parties will no longer consider what can be guaranteed by applying the 'maximin rule' (i.e. what Rawls calls "the guaranteeable level") to be satisfactory. So, given strictly convex individual utility functions, it will be unlikely that the representative parties of the original position will choose the difference principle over utilitarianism as a result of their deliberation processes.

I conclude that, under close formal examination, Rawls's argument for his justice as fairness is self-defeating.[28] The utilitarian dog has bit the Rawlsian hand that fed it!



## Appendix: Proofs of Main Results

**Proposition 1** *For any* $\bar{W} \geq 0$*, the difference principle prescribes* $(x_M, x_L) = \left(\frac{\bar{W}}{2}, \frac{\bar{W}}{2}\right)$.

***Proof of Proposition 1*:** Fix an arbitrary $\bar{W} \geq 0$. The specific distribution that the difference principle will prescribe will be the solution to the following problem,

$$\max_{x,y \in \mathbb{R}} \min\{x_M, x_L\}$$
$$\text{subject to } x_M, x_L \geq 0 \text{ and } x_M + x_L \leq \bar{W}$$

Let $(x_M^*, x_L^*)$ denote the solution to the problem. I claim that $x_M^* = x_L^*$. For suppose $x_M^* \neq x_L^*$. Without loss of generality, suppose $x_M^* > x_L^*$. Let $\Delta = \frac{x_M^* - x_L^*}{2}$. Then, $(x_M^* - \Delta, x_L^* + \Delta)$ would be another feasible distribution. But, note $x_M^* > x_M^* - \Delta = x_L^* + \Delta > x_L^*$, which contradicts that $x_L^*$ is part of the solution to the above problem. Hence, $x_M^* = x_L^*$. Now, we need to show that $x_M^* = x_L^* = \frac{\bar{W}}{2}$. Suppose not. Then, $x_M^* = x_L^* < \frac{\bar{W}}{2}$ which implies $x_M^* + x_L^* < \bar{W}$. Define $\Delta' = \frac{\bar{W} - (x_M^* + x_L^*)}{2}$. Then, $(x_M^* + \Delta', x_L^* + \Delta')$ would be another feasible distribution such that $x_M^* + \Delta' = x_L^* + \Delta' > x_M^* = x_L^*$ contradicting that $(x_M^*, x_L^*)$ is a solution to the above problem. □

[28] Although not being utilitarians themselves, Brennan (2007) and Tomasi (2012) have employed a similar strategy of showing the self-defeating nature of Rawls's theory by arguing that the main intent behind Rawls's difference principle can actually be better satisfied, not by the direct implementation of the difference principle itself, but rather, by allowing robust free market economic freedoms.

**Proposition 2** *Suppose* $\bar{W} = r_M + r_L$. *Then, utilitarianism prescribes* $(x_M, x_L) = (r_M, r_L)$, *while the difference principle prescribes* $(x_M, x_L) = \left(\frac{r_M + r_L}{2}, \frac{r_M + r_L}{2}\right)$.

***Proof of Proposition 2*:** Let $(x_M, x_L)$ be the distribution that utilitarianism prescribes. We need to show that $(x_M, x_L) = (r_M, r_L)$. First, note that we must have $x_M + x_L = r_M + r_L = \bar{W}$ as if $x_M + x_L < r_M + r_L = \bar{W}$, then we can always increase total social welfare (i.e. the sum of MAG and LAG's welfare) by distributing $\bar{W} - (x_M + x_L)$ to any of the individuals. Hence, $x_M + x_L = r_M + r_L = \bar{W}$. To complete the proof, we need to show that $x_M = r_M$ and $x_L = r_L$. For a proof by contradiction, suppose not. So, we must have $x_M \neq r_M$ or $x_L \neq r_L$, but since $x_M + x_L = r_M + r_L$, we must have $x_M \neq r_M$ and $x_L \neq r_L$.

*Case 1*: Suppose $x_M < r_M$. Then, since $x_M + x_L = r_M + r_L$, we must have $x_L > r_L$. Let $\epsilon \in (0, r_M - x)$. Then,

$$
\begin{aligned}
&u_M(x_M+) + u_L(x_L - \epsilon) \\
&\quad = u_M(x_M) + \{u_M(x_M+) - u_M(x_M)\} + u_L(x_L) - \{u_L(x_L) - u_L(x_L - \epsilon)\} \\
&\quad = u_M(x_M) + u_L(x_L) + \{u_M(x_M + \epsilon) - u_M(x_M)\} - \{u_L(x_L) - u_L(x_L - \epsilon)\} \\
&\quad = u_M(x_M) + u_L(x_L) + \int_{x_M}^{x_M+\epsilon} u'_M(z)dz - \int_{x_L-\epsilon}^{x_L} u'_L(z)dz \\
&\quad = u_M(x_M) + u_L(x_L) + \int_{x_M}^{x_M+\epsilon} u'_M(z)dz - \int_{x_M}^{x_M+\epsilon} u'_M(z + \{(x_L - \epsilon) - (r_L - r_M) - x_M\})dz \\
&\quad = u_M(x_M) + u_L(x_L) + \int_{x_M}^{x_M+\epsilon} [u'_M(z) - u'_M(z + \{(x_L-) - (r_L - r_M) - x_M\})]dz \\
&\quad > u_M(x_M) + u_L(x_L) [\text{as } u'_M(z) > u'_M(z + \{(y - \epsilon) - (r_L - r_M) - x\}), \forall z \in (x_M, x_M + \epsilon)]
\end{aligned}
$$

So, $(x_M + \epsilon, x_L - \epsilon)$ is another feasible distribution that generates a greater total sum of individual utilities than $(x_M, x_L)$, which contradicts that $(x_M, x_L)$ is an distribution that utilitarianism prescribes.

*Case 2*: Suppose $x_M > r_M$. Then, since $x_M + x_L = r_M + r_L$, we must have $x_L < r_L$. Let $\epsilon \in (0, x_m - r_M)$. Then,

$$
\begin{aligned}
&u_M(x_M-\epsilon)+u_L(x_L+\epsilon)\\
&\quad=u_M(x_M)-\{u_M(x_M)-u_M(x_M-\epsilon)\}+u_L(x_L)+\{u_L(x_L+\epsilon)-u_L(x_L)\}\\
&\quad=u_M(x_M)+u_L(x_L)+\{u_L(x_L+\epsilon)-u_L(x_L)\}-\{u_M(x_M)-u_M(x_M-\epsilon)\}\\
&\quad=u_M(x_M)+u_L(x_L)+\int_{x_L}^{x_L+}u_L'(z)dz-\int_{x_M-}^{x_M}u_M'(z)dz\\
&\quad=u_M(x_M)+u_L(x_L)+\int_{x_L}^{x_L+\epsilon}u_L'(z)dz-\int_{x_L}^{x_L+\epsilon}u_L'\big(z+\{(x_M-\epsilon)+(r_L-r_M)-x_L\}\big)dz\\
&\quad=u_M(x_M)+u_L(x_L)+\int_{x_L}^{x_L+\epsilon}\big[u_L'(z)-u_L'\big(z+\{(x_M-\epsilon)+(r_L-r_M)-x_L\}\big)\big]dz\\
&\quad>u_M(x_M)+u_L(x_L)[\text{as } u_L'(z)>u_L'\big(z+\epsilon\{(x_M-\epsilon)+(r_L-r_M)-x_L\}\big),\forall z\in(x_L,x_L+\epsilon)]
\end{aligned}
$$

So, $(x_M-\epsilon,x_L+\epsilon)$ is another feasible distribution that generates a greater total sum of individual utilities than $(x_M,x_L)$, which contradicts that $(x_M,x_L)$ is an distribution that utilitarianism prescribes. That the difference principle prescribes $(x_M,x_L)=\left(\frac{r_M+r_L}{2},\frac{r_M+r_L}{2}\right)$ follows from Proposition 1. □

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

***Proof of Proposition 3*:** By Proposition 2, when total wealth is $r_M+r_L$, $(x_M,x_L)=(r_M,r_L)$ is the distribution that maximizes the sum of individual utilities of MAG and LAG. So, 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 wealth left for further distribution. Now, the problem reduces to,

$$
\begin{aligned}
&\max_{(x_1,x_2)\in\mathbb{R}^2}u_M(r_M+x_1)+u_L(r_L+x_2)\\
&\text{subject to } x_1+x_2=\bar{W}-(r_M+r_L).
\end{aligned}
$$

Note that for all $x\in[0,\bar{W}-(r_M+r_L)]$, we have $u_A(r_M+x)=u_L(r_L+x)$. So, $u_M(r_M+x_1)+u_L(r_L+x_2)=u_M(r_M+x_1)+u_M(r_M+x_2)$. By substituting $\bar{W}-(r_M+r_L)-x_1$ 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 $u_A$ 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,

$$
\begin{aligned}
& u'_M(r_M + x_1) - u'_M(\bar{W} - r_L - x_1) = 0 \\
& \Rightarrow u'_M(r_M + x_1) = u'_M(\bar{W} - r_L - x_1) \\
& \Rightarrow r_M + x_1 = \bar{W} - r_L - x_1 \\
& \Rightarrow x_1 = \frac{\bar{W} - r_M - r_L}{2} \text{ and } x_2 = \frac{\bar{W} - r_M - r_L}{2}.
\end{aligned}
$$

Hence, $(x_M, x_L) = \left(r_M + \frac{\bar{W}-r_M-r_L}{2}, r_D + \frac{\bar{W}-r_M-r_L}{2}\right) = \left(\frac{\bar{W}+r_M-r_L}{2}, \frac{\bar{W}-r_M+r_L}{2}\right)$ is the utilitarian solution. That the difference principle prescribes $(x_M, x_L) = \left(\frac{\bar{W}}{2}, \frac{\bar{W}}{2}\right)$ follows from Proposition 1. □

**Proposition 4** *Suppose $\bar{W} \geq r_M + r_L$. Then, the following two claims are true,*

(a) *The total social welfare generated by utilitarianism is strictly greater than the total social welfare generated by the difference principle; and*
(b) *The welfare level that LAG enjoys is always greater under utilitarianism than under the difference principle.*

***Proof of Claim (a) of Proposition 4*:** That total social welfare is maximized under utilitarianism simply derives from the very definition of utilitarianism. So, in order to prove the claim, all we need to do is to show that total social welfare is *not* maximized under the difference principle. This can be done by showing that the marginal utilities of MAG and LAG are different under the distribution $(x_M, x_L) = \left(\frac{\bar{W}}{2}, \frac{\bar{W}}{2}\right)$, which is the distribution prescribed by utilitarianism. Note $u'_M\left(\frac{\bar{W}}{2}\right) = u'_L\left(\frac{\bar{W}}{2} + (r_L - r_M)\right) \neq u'_L\left(\frac{\bar{W}}{2}\right)$. □

***Proof of Claim (b) of Proposition 4*:** Since $u_L$ is strictly increasing, in order to prove the claim, all we need to do is to show that the amount of wealth distributed to LAG by utilitarianism is greater than the amount of wealth allocated to LAG by the difference principle. Suppose $\bar{W} = r_M + r_L$. Then, by Proposition 3, LAG is distributed $r_L$ amount of wealth under utilitarianism, and $\frac{\bar{W}}{2} = \frac{r_M+r_L}{2}$ amount of wealth under the difference principle. Note $r_L - \frac{\bar{W}}{2} = r_L - \frac{r_M+r_L}{2} = \frac{r_L-r_M}{2} > 0$ as desired. Now, suppose $\bar{W} > r_M + r_L$. Then, by proposition 4, LAG is distributed $\frac{\bar{W}-r_M+r_L}{2}$ amount of wealth under utilitarianism, and $\frac{\bar{W}}{2}$ amount of wealth under the difference principle. Note $\frac{\bar{W}-r_M+r_L}{2} - \frac{\bar{W}}{2} = \frac{r_L-r_M}{2} > 0$ as desired. □

## References

Arrow, K., Sen, A., & Suzumura, K. (2002). *Handbook of social choice and welfare* (Vol. 1). Amsterdam: Elsevier.

Arrow, K., Sen, A., & Suzumura, K. (2010). *Handbook of social choice and welfare* (Vol. 2). Amsterdam: Elsevier.
Barry, B. (1989). *Theories of justices*. Oakland: University of California Press.
Bossert, W., & Weymark, J. (2004). Utility in social choice. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory* (Vol. 2). Dordrecht: Kluwer Academic Publishers.
Brennan, J. (2007). Rawls' paradox. *Constitutional Political Economy, 18,* 287–299.
Broome, J. (1987). Utilitarianism and expected utility. *The Journal of Philosophy, 84*(8), 405–422.
D'Agostino, F., Gaus, G., & Thrasher, J. (2014). Contemporary approaches to the social contract. In *The Stanford encyclopedia of philosophy*. http://plato.stanford.edu/archives/spr2014/entries/contractarianism-contemporary/.
D'Aspremont, C. (2010). Social welfare functionals and interpersonal comparability. In K. Arrow, A. Sen, & K. Suzumura (Eds.), *Handbook of social choice and welfare* (Vol. 1). Amsterdam: Elsevier.
Dworkin, R. (1981a). What is equality: Part 1. Equality of welfare. *Philosophy and Public Affairs (Summer, 1981), 10*(3), 185–246.
Dworkin, R. (1981b). What is equality: Part 2. Equality of resources. *Philosophy and Public Affairs (Autumn, 1981), 10*(4), 283–345.
Fleurbaey, M., Salles, M., & Weymark, J. (Eds.). (2008). *Justice, political liberalism, and utilitarianism: Themes from Harsanyi and Rawls*. Cambridge: Cambridge University Press.
Frankfurt, H. (1987). Equality as a moral ideal. *Ethics, 98*(1), 21–43.
Gaertner, W. (2009). *A primer in social choice theory*. Oxford: Oxford University Press.
Gaus, G., & Thrasher, J. (2015). Rational choice and the original position, the (many) models of Rawls and Harsanyi. In T. Hinton (Ed.), *the original position* (pp. 39–58). Cambridge: Cambridge University Press.
Gauthier, D. (1985). Bargaining and justice. *Social Philosophy and Policy, 2*(2), 29–47.
Hammond, P. (1976). Equity, arrow's conditions, and Rawls' difference principle. *Econometrica, 44*(4), 793–804.
Hampton, J. (1980). Contracts and choices: Does Rawls have a social contract theory? *Journal of Philosophy, 77,* 315–338.
Harsanyi, J. (1955). Cardinal welfare, individualistic ethics, and the interpersonal comparisons of utility. *The Journal of Political Economy, 63,* 309–321.
Harsanyi, J. (1975). Review, can the maximin principle serve as a basis for morality? A critique of John Rawls's theory. *The American Political Science Review, 69*(2), 594–606.
Harsanyi, J. (1977). *Rational behavior and bargaining equilibrium in games and social situations*. Cambridge: Cambridge University Press.
Moehler, M. (2018). The Rawls–Harsanyi dispute: A moral point of view. *Pacific Philosophical Quarterly*, *99*(1), 82–99.
Mongin, P., & Claude, D. (1998). Utility theory and ethics. In S. Barbera, P. Hammond, & C. Seidl (Eds.), *Handbook of utility theory: Volume 1 principles*. Dordrecht: Kluwer Academic Publishers.
Moreno-Ternero, J., & Roemer, J. (2008). The veil of ignorance violates priority. *Economics and Philosophy, 24*(2), 233–257.
Moulin, H. (2003). *Fair division and collective welfare*. Cambridge: MIT Press.
Rawls, J. (1971/1999). *A theory of justice (revised edition).* Harvard University Press.
Rawls, J. (1974a). Some reasons for the maximin criterion. *The American Economic Review, 64*(2), 141–146.
Rawls, J. (1974b). Reply to Alexander and Musgrave. *The Quarterly Journal of Economics, 88*(4), 633–655.
Rawls, J. (1993, 2005). *Political liberalism*. Columbia University Press.
Rawls, J. (2001). *Justice as fairness*. Cambridge: Belknap Harvard.
Risse, M. (2002). Harsanyi's 'utilitarian theorem' and utilitarianism. *Nous, 36*(4), 550–577.
Roemer, J. (1996). *Theories of distributive justice*. Cambridge: Harvard University Press.
Roemer, J. (2002). Egalitarianism against the Veil of Ignorance. *Journal of Philosophy, 99,* 167–184.
Roemer, J. (2008). Harsanyi's impartial observer is *not* a utilitarian. In F. Marc, M. Salles, & J. Weymark (Eds.), *Justice, political liberalism, and utilitarianism: Themes from Harsanyi and Rawls*. Cambridge: Cambridge University Press.
Sen, A. (1976). Welfare inequalities and Rawlsian axiomatics. *Theory and Decision, 7,* 243–262.
Sen, A. (1979). Equality of what? *The Tanner Lecture on Human Value.*
Tomasi, J. (2012). *Free market fairness*. Princeton: Princeton University Press.
Weymark, J. (1991). A reconsideration of the Harsanyi-Sen debate on utilitarianism. In J. Elster & J. Roemer (Eds.), *Interpersonal comparisons of well-being*. Cambridge: Cambridge University Press.