Economics & Philosophy (2023), **39**, 308–333
doi:10.1017/S0266267122000098

ARTICLE

When utilitarianism dominates justice as fairness: an economic defence of utilitarianism from the original position

Hun Chung

Faculty of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda, Shinjuku-ku, Tokyo, Japan 169-8050
Emails: hun.chung@waseda.jp, hunchung1980@gmail.com

(Received 24 March 2020; revised 8 February 2022; accepted 19 March 2022; first published online 11 August 2022)

Abstract

The original position together with the veil of ignorance have served as one of the main methodological devices to justify principles of distributive justice. Most approaches to this topic have primarily focused on the single person decision-theoretic aspect of the original position. This paper, in contrast, will directly model the basic structure and the economic agents therein to project the economic consequences and social outcomes generated either by utilitarianism or Rawls's two principles of justice. It will be shown that when the differences in people's productive abilities are sufficiently great, utilitarianism dominates Rawls's two principles of justice by providing a higher level of overall well-being to every member of society. Whenever this is the case, the parties can rely on the Principle of Dominance (which is a direct implication of instrumental rationality) to choose utilitarianism over Rawls's two principles of justice. Furthermore, when this is so, utilitarianism is free from one of its most fundamental criticisms that it 'does not take seriously the distinction between persons' (Rawls 1971 [1999]: 24).

Keywords: Rawls; original position; veil of ignorance; difference principle; utilitarianism

1. Introduction: the original position and normative decision rules

The original position together with the veil of ignorance have served as one of the main methodological devices to justify principles of distributive justice. John Rawls (1971 [1999]) used it to justify (along with principles of 'equal basic liberties' and 'fair equality of opportunity') his 'difference principle', according to which, social and economic inequalities should be arranged so that they are 'to the greatest benefit of the least-advantaged members of society' (Rawls 2001: 42–43). John Harsanyi (1955, 1977) used it to justify average utilitarianism, according to which benefits should be distributed to maximize average social welfare.

The basic thought is that principles of distributive justice are justified to the extent that it could be shown that they would be the outcome of a fair and impartial agreement made by rational, reasonable and mutually disinterested agents (Rawls 1971 [1999]: section 25) in a suitably defined hypothetical original position. The veil of ignorance is what guarantees the fairness and impartiality of the resulting agreement by concealing any morally arbitrary information from the contracting parties that may distract and skew their judgements in a partial way.[1] The underlying presumption is that these conditions and the very setup of the original position render the preferences of its contracting parties indicative of what principles of distributive ethics normatively require.

Since the normative justification of the resulting principle of distributive justice derives from the decisions made by the original contracting parties, most of the debate has centred around which normative decision rule it would be reasonable for the parties in the original position to follow in making their decisions. Rawls had argued that following the maximin rule – which requires the parties to choose the option 'the worst outcome of which is superior to the worst outcomes of the others' (Rawls 1971 [1999]: 133) – would lead to the justification of his difference principle (Rawls 1971 [1999]: section 26); while Harsanyi argued that adopting the principle of expected utility maximization would lead the parties to the justification of the principle of average utility (Harsanyi 1975: 598).[2] Hence, as Buchak rightly notes, 'Whether the rule derived from the original position is utilitarianism or maximin equity [i.e. Rawls's difference principle] depends on whether the setup makes expected utility maximization or maximin appropriate. And, of course, if the setup makes a different rule appropriate, then the result will be something else' (Buchak 2017: 627).

Recent scholarly attempts have explored how varying the initial setup and the assumptions of the original position may lead to the justification of alternate principles of distributive justice that are different from Rawls's difference principle or Harsanyi's utilitarianism. For instance, Buchak argues that if the parties in the original position are assumed to be risk-averse and seek to maximize (not just simple expected utility, but) 'risk-weighted' expected utility, then the parties will arrive at (what she calls) 'relative prioritarianism' (or 'weighted-rank utilitarianism'), an intermediate position between Rawls and Harsanyi, according to which, 'the well-being of the relatively worse off counts for more than that of the relatively better off but that everyone's well-being counts for something' (Buchak 2017: 611). Relatedly, according to Stefansson, if we allow the parties in the original position to display 'ambiguity aversion' [meaning 'that other things being equal, they prefer gambles with known chances of outcomes over games with unknown chances' (Stefansson 2021: section 3)] (which Stefansson understands as a particular form

[1] According to Rawls, the morally arbitrary information that the veil of ignorance is designed to block are one's class position or social status; one's natural assets and talents; one's conception of the good; specific features of one's psychology including one's attitude towards risk; the particular circumstances of one's society; to which generation one belongs, etc. (Rawls 1971 [1999]: 118–119).

[2] Although Harsanyi himself claimed that his 'own model yields a moral theory based on the principle of *average* utility' (Harsanyi 1975: 598, emphasis mine), his argument can also be used to justify *total* utilitarianism given that we assume a fixed population. See Weymark (1991) for an excellent exposition of Harsanyi's 'aggregation theorem' and his 'impartial observer theorem'.

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

of risk-aversion), the result of the original position is what Stefansson calls ‘Distribution-Sensitive Utilitarianism’, which Stefansson characterizes as a form of egalitarianism (Stefansson 2021: sec. 5). One way to interpret this, according to Stefansson, is to accept ‘that since the veil of ignorance argument is so sensitive to subtle modelling choices, the argument does not settle the debates between the main competing views in distributive ethics’ (Stefansson 2021: sec. 1).

If so, then one way to advance the debate might be to examine the specific modelling choices and the conditions under which a particular principle of distributive ethics, say, utilitarianism, outperforms others in all relevant aspects and then check whether these conditions can be reasonably assumed by the parties in the original position given their knowledge of ‘general facts of human society’ (which, according to Rawls, includes knowledge of ‘political affairs’, ‘the principles of economic theory’, ‘the basis of social organization’ and ‘the laws of human psychology’) (Rawls 1971 [1999]: 119). If those conditions under which utilitarianism outperforms Rawls’s two principles of justice in all relevant aspects turn out to be reasonable for the parties in the original position to assume given their knowledge of general facts about human society, then we may bypass the issue concerning which normative decision rule it would be most reasonable for the parties in the original position to adopt as the parties can now simply rely on (what I will later introduce as) the Principle of Dominance, which is directly implied by (instrumental) rationality, to choose utilitarianism. Such will be the basic strategy pursued in this paper. This paper will argue that, under conditions that are both reasonably realistic and consonant with Rawls, utilitarianism (Pareto) *dominates* Rawls’s two principles of justice by allowing *everybody* to enjoy a higher overall well-being than what each would expect to enjoy under a basic structure organized by Rawls’s two principles of justice. And, under conditions in which this holds, utilitarianism will further be free from one of its most fundamental criticisms – namely, that it ‘does not take seriously the distinction between persons’ (Rawls 1971 [1999]: 24).

In pursuing this strategy, I will take a rather different approach than what has been typically employed in recent scholarship dealing with the veil of ignorance and the original position. Most research papers that comment and/or critique on the original position along with veil of ignorance primarily focus on the decision-theoretic aspect of the original position;[3] that is, many papers working on this topic treat the ‘choice behind the veil as the choice of a single individual’ (Stefansson 2021: sec. 2) and reduce the problem faced by the parties in the original position to a choice of a ‘lottery’, whose probability distribution over various social outcomes is either known or unknown. Some scholars have criticized this ‘single-individual-decision-theoretic’ approach to the original position on grounds that principled disagreements (what Ryan Muldoon calls ‘disagreement in perspective’) may still persist even behind the veil of ignorance, and hence ‘the device of the ‘veil of ignorance’ in moral and political philosophy does not guarantee that all agents can be effectively reduced to a single agent selected at random’ (Muldoon *et al.* 2014: 379).

[3] See Harsanyi (1953, 1955, 1977), Rawls (1971 [1999]), Buchak (2017), Moehler (2018) and Stefansson (2021).

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

What I see lacking in these single-person-decision-theoretic approaches is rather the lack of a micro-level mechanism that generates various economic and social outcomes contained in the lotteries that the original contracting parties consider. That is, these approaches typically do not model the underlying economy or the basic structure, and, hence, do not illustrate how the various social-economic outcomes contained in the different lotteries that the original contracting parties consider are the results of the intricate interactions of various economic agents, who respond rationally to the different incentives provided by their basic structures. This is an important limitation of single-person-decision-theoretic approaches as Rawls himself took the issue of economic incentives as fundamental and used it as one of his main defences of the difference principle over complete distributional egalitarianism (Rawls 1971 [1999]: 68, 142, 246). In contrast, this paper will assume that the parties in the original position utilize their knowledge of 'general facts of human society' (which, again, includes knowledge of 'political affairs', 'the principles of economic theory', 'the basis of social organization' and 'the laws of human psychology') (Rawls 1971 [1999]: 119) to directly model the basic structure and the economic agents therein to project the economic consequences and social outcomes of the basic structures prescribed either by Rawls's two principles of justice or utilitarianism to inform their choices of principles of distributive ethics in the original position.[4]

2. The separateness of persons objection

Rawls's criticisms against utilitarianism and the reasons that he thinks his justice as fairness is superior to utilitarianism are spread throughout *A Theory of Justice*: he argues that the issues of strains of commitment, stability, publicity, and establishing the social basis for self-respect all favour justice as fairness over utilitarianism (Rawls 1971 [1999]: sec. 29). All of these criticisms stem from Rawls's belief that utilitarianism 'does not take seriously the distinction between persons' (Rawls 1971 [1999]: 24). Call such an objection to utilitarianism, *the separateness of persons objection.*

The problem of utilitarianism, according to Rawls, is that it invalidly extends 'to society the principle of choice for one man, and then, to make this extension work, conflating all persons into one through the imaginative acts of the impartial sympathetic spectator' (Rawls 1971 [1999]: 24). The problem is that when such a principle of balancing benefits and burdens to maximize the overall net satisfaction is extended to society at large (which is the case for utilitarianism), one person's significant welfare loss may be justified (or even required) whenever such a welfare loss is outweighed by the welfare gains of *other* people. This is the sense in which utilitarianism has been criticized that it may, under some circumstances, justify 'if not slavery or serfdom, at any rate serious infractions of liberty for the sake of greater social benefits' (Rawls 1971 [1999]: 26). Given that there exist other feasible social arrangements in which the 'losers' are better off than in the social arrangement utilitarianism prescribes,

[4] Other papers that have taken a similar economic approach include Roemer (2002), Moreno-Ternero and Roemer (2008) and Chung (2020). See also Roemer (1996).

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

'[w]hat the principle of utility asks is precisely … to accept the greater advantages of others as a sufficient reason for lower expectations over the whole course of [their] li[ves]' (Rawls 1971 [1999]: 155). Not only does this require 'those who must make sacrifices strongly identify with interests broader than their own' (Rawls 1971 [1999]: 155), which, according to Rawls, cannot reasonably be assumed given the general facts of moral psychology (Rawls 1971 [1999]: 153–155), but it also neglects that the 'losers' are separate people leading distinct lives, who each deserve to be treated (following Kant) not merely as means but as ends in themselves (Rawls 1971 [1999]: 156, see also Nozick 1974: 33). This is the essence of the separateness of persons objection.

There are three things to clarify about the separateness of persons objection. The first is that the separateness of persons objection ultimately stems from the *aggregative* nature of utilitarianism, according to which 'the gains for group of individuals can *morally* outweigh the losses for a different group of individuals' (Hirose 2013: 185). Utilitarianism and prioritarianism[5] are both aggregative moral principles; both permit that the gains accruing to some group of individuals can morally outweigh the losses incurred by another group of individuals. Hence, both utilitarianism and prioritarianism are subject to the separateness of persons objection.[6] However, there is a sense in which the separateness of persons objection applies to utilitarianism more forcefully. This is because, unlike prioritarianism, which assumes that '[b]enefiting people matters more the worse off these people are' (Parfit 1997: 213), utilitarianism holds that, in calculating the overall value of a distribution, the benefits to the worse off and the benefits to the better off can be traded-off at a one-to-one ratio. This implies that, while requiring society to maximize (either total or average) aggregate social welfare, utilitarianism is completely unconcerned with how welfare is distributed across different individuals.[7] This would be different from prioritarianism, which would hold that given that the total welfare in the two distributions is the same, the distribution that has a more equal spread of welfare is better. This is why Benbaji has characterized prioritarianism as being 'derivatively (if not, directly) egalitarian' (Benbaji 2005: 312).

[5] See Parfit (1997, 2012).

[6] Buchak's 'relative prioritarianism' is no exception. Although Buchak tries to explain how her relative prioritarianism avoids the separateness of persons objection, the explanation she provides is actually closer to biting the bullet, rather than a genuine escape. Buchack (2017: 640) explains that her relative prioritarianism avoids the objection and respects the separateness of persons as it acknowledges 'a plurality of acceptable risk attitudes but a single correct importance attitude'. What Buchak is basically saying here is that although different people can all rationally disagree about how much utility they are willing to trade-off between different possible states that happen exclusively to themselves, they would all have to agree on a single ratio derived from the 'default risk attitude' which is 'the most risk avoidant of the reasonable risk attitudes' (Buchak 2017: 631) of the parties in the original position with which they may justifiably trade-off utility from the worse-off to utility to the better-off. Even if this ratio is heavily in favour of those who are worse-off, Buchak's relative prioritarianism still permits that the loss incurred by the worse-off can be adequately compensated by sufficiently large gains to the better-off, which is precisely what the separateness of persons objection deems unacceptable.

[7] For instance, in a society consisting of two individuals, utilitarianism is indifferent between a distribution that generates 10 units of welfare to the first person and 0 units of welfare to the second person and a distribution that generates 5 units of welfare to the two persons equally.

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

Second, an important presumption of the separateness of persons objection is that the choice of a society's basic structure (i.e. its basic political and economic institutions) likely generates conflicts of interests and potentially different sets of 'winners' and 'losers'. At the heart of the issue is whether we can justifiably *further* sacrifice the interests of the losers for the sake of achieving the greater good of the winners or society as a whole. When the separateness of persons objection is raised against utilitarianism, the underlying assumption is that the institutional arrangement that utilitarianism prescribes will most likely further lower the expectations of the lesser advantaged (relative to the institutional arrangement prescribed by Rawls's two principles of justice) for the sake of maximizing the greater good of society as a whole. As Hirose explains, 'If one alternative benefits some person and harms no other person, it would be agreed unanimously that this alternative should be chosen' (Hirose 2013: 185). In such cases, the separateness of persons objection does not arise; the separateness of persons objection arises only when utilitarianism requires the relatively worse-off to endure *additional* sacrifices to achieve a greater good for society as a whole.

The third point concerns the *directionality* of the separateness of persons objection. The separateness of persons objection applies *asymmetrically*: it applies only when *the least* or *lesser* advantaged are required to go through further sacrifices to improve the situations of the better-off or society as a whole; it does not apply when the better-off groups are asked to forgo additional benefits to improve the situation of the least or lesser advantaged. Rawls's difference principle requires to design a society's basic structure so that it maximizes the amount of primary goods that go to the least advantaged even if this means decreasing the overall welfare enjoyed by the other groups. The question then is: why does Rawls not raise a similar separation of persons objection to his own difference principle on behalf of the better-off groups, who have to accept a lower expectation of well-being in order to maximize the benefits that go to the least advantaged?[8]

Rawls's answer comes from the *principle of reciprocity*, which he believes grounds his difference principle:

> Thus the more advantaged, when they view the matter from a general perspective, recognize that the well-being of each depends on a scheme of social cooperation without which no one could have a satisfactory life; they recognize also that they can expect the willing cooperation of all only if the terms of the scheme are reasonable. *So they regard themselves as already compensated*, as it were, by the advantages to which no one (including themselves) had a prior claim. (Rawls 1971 [1999]: 88, emphasis added)

In other words, the reason that we need the difference principle is that society, as a cooperative scheme for mutual benefit, needs to induce the willing cooperation of all, and, in particular, the members of the least advantaged group. The reason that it is not unfair to lower the expectations of the more advantaged for this purpose is that the more advantaged (by simply being more advantaged in the cooperative

[8] This is exactly the criticism that Nozick raises against Rawls's difference principle. See Nozick (1974: 192-197).

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

scheme) are already sufficiently compensated by the advantages to which they had no prior claim. This explains why the separateness of persons objection applies only in one direction; it applies only when the worst-off are made even worse for the sake of benefiting other groups or society as a whole.

3. Epistemic and psychological assumptions of the original position

As explained in the Introduction, varying the initial setup and the basic assumptions of the original position may lead to the justification of different principles of distributive justice. Many of these basic assumptions are epistemic or psychological in nature. The following is a list of epistemic and psychological assumptions that crucially affect the outcome of the original position:

1. **Knowledge of Probabilities:** Do the parties in the original position have an objective basis to estimate probabilities? If not, are they still allowed to use subjective probabilities to calculate expectations? Or should the reliance on such probabilistic calculations be strictly disallowed?
2. **Risk Attitudes:** What attitude toward risk or uncertainty should the original contracting parties have? Should they be risk (or ambiguity) averse, risk (or ambiguity) neutral, or risk (or ambiguity) seeking?
3. **Which Normative Decision Rule?:** Which normative decision rule should the original contracting parties ultimately use to base their decisions? The principle of expected utility maximization? (Or the principle of risk-weighted expected utility maximization?) Or the maximin rule? Etc.
4. **Shape of Utility Functions:** How should the original contracting parties conceive the general shape of individual utility functions? Should the original contracting parties assume individual utility functions are concave? Linear? Convex?[9]
5. **Distribution of What?:** What are principles of distributive justice designed to distribute or regulate? People's welfare? Their (index of) primary goods[10] or resources?[11] Their capabilities?[12] Their opportunities?[13] Or some other measure of advantage?

These epistemic and/or psychological assumptions are not fully independent and tend to be closely correlated. For instance, one important reason that made Rawls think that it would be appropriate for the parties in the original position to use maximin as their normative decision rule was that because of 'the veil of ignorance [which] excludes all knowledge of likelihoods', Rawls thought that the

[9] See Chung (Forthcoming) for a discussion of how characterizing individual utility functions in accordance with Kahneman and Tversky (1979)'s prospect theory (viz., as being convex below and concave above a given reference point) affects the overall plausibility of utilitarianism (in comparison to Justice as Fairness) for the parties in the original position. The resulting utilitarianism is what Chung calls 'prospect utilitarianism'. Chung (2017) argues that prospect utilitarianism is better than sufficientarianism by retaining all of sufficientarianism's main attractions while avoiding its drawbacks.

[10] See Rawls (1971 [1999]).

[11] See Dworkin (1981*a*, 1981*b*).

[12] See Sen (1980).

[13] See Roemer (2009).

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

parties 'have no basis for probability calculations' (Rawls 1971 [1999]: 134).[14] Conversely, it is well-known that Harsanyi advocated the principle of expected utility maximization, and one important reason for this [aside from his thinking that the maximin rule is 'highly irrational' (Harsanyi 1975: 595)] was that Harsanyi believed that it would not only be perfectly admissible [on the basis of the principle of indifference/insufficient reason] for the parties to assign the same probability ($1/n$) of taking the place of each of the n individuals in society, but it may even be morally required to do so in order to 'give the same a priori weight to the interests of all members of the society' (Harsanyi 1975: 598 footnote 10). While accepting Harsanyi's 'equiprobability assumption', Buchak argues that there could be a wide range of *reasonable* risk attitudes that the original contracting parties can have, and, hence, proposes that the parties, instead, adopt the principle of *risk-weighted expected utility maximization* (REU-maximization) (for which expected utility maximization is a special case), while applying 'the most risk avoidant of the reasonable risk attitudes' (Buchak 2017: 614–620, 631).

However, one should note that whether the parties in the original position should choose maximin or expected utility maximization or REU-maximization is not directly implied by instrumental rationality alone; each normative decision rule can only be justified in combination with other assumptions concerning the parties' available probabilistic information (or the lack thereof) and their associated risk attitudes. However, there does exist a normative decision principle that is independent of one's assumptions concerning probability distributions or risk attitudes, and, further, is directly implied solely by instrumental rationality: it is what I call *the principle of dominance*.

- **The Principle of Dominance:** Given a choice between any two options X and Y, if option X *dominates* options Y – that is, if X generates outcomes that are no worse than Y in all possible states and in some states generates outcomes that are strictly better than Y, – then rationality requires one to choose X over Y.

Of course, the principle of dominance has a rather limited scope in its application; once we rule out options that are obviously bad and unworthy of consideration, there tend to be few remaining cases in which one option clearly dominates another option. But if there does exist one dominant option even after eliminating obviously bad options that are unworthy of consideration, then it would just be plainly instrumentally irrational not to choose it – and furthermore, this is true independent of one's particular assumptions concerning knowledge of probabilities or risk attitudes.

Hence, one potential way to advance the debate concerning which principle of distributive ethics will be eventually justified from the original position in the direction of supporting utilitarianism would be to show that the economic

[14] This is the first among 'three chief features' (Rawls 1971 [1999]: 134) that Rawls suggests are jointly sufficient to make it suitable to adopt the maximin rule for a given situation. The second condition that makes it suitable to adopt the maximin rule is that the minimum stipend guaranteed by relying on the maximin rule is sufficiently satisfactory (Rawls 1971 [1999]: 133). The third condition is that the other options that are rejected by the maximin rule 'have outcomes that one can hardly accept' (Rawls 1971 [1999]: 133). See also Rawls (1974*a*).

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

consequences produced by a basic structure organized by utilitarianism dominates those produced by a basic structure organized by any other principle of distributive ethics. That will be the basic strategy that will be employed in the remainder of this paper. Here, I will primarily focus on comparing the economic consequences of utilitarianism and Rawls's two principles of justice as they represent the two dominant pillars of the classical Rawls-vs-Harsanyi debate, but the argument can easily be extended to support other aggregative distributive principles such as (relative) prioritarianism. If it turns out that the economic consequences generated by a basic structure organized by utilitarianism *Pareto dominates* those generated by a basic structure organized by Rawls's difference principle (both in terms of people's welfare levels as well as their index of primary goods), not only can it be claimed that the parties in the original position will choose utilitarianism over Rawls's difference principle on the basis of *the principle of dominance* (which requires no probabilistic information for its application), but the force of the criticism that utilitarianism does not take seriously the separateness of distinct persons will also be significantly diffused because there will simply be no conflict of interests or trade-offs across different individuals under utilitarianism as nobody's interests can be said to be sacrificed for the sake of achieving the greater good of society as a whole.

4. The model

4.1. The setup

Recall that the parties in the original position, in spite of being behind the veil of ignorance, are equipped with knowledge of 'general facts about human society' including knowledge of the principles of political science, economics, sociology and human psychology (Rawls 1971 [1999]: 119). Suppose that, based on this knowledge, the parties in the original position construct a model of society for the purpose of projecting the economic consequences that are likely to be generated by the institutional arrangements respectively prescribed by utilitarianism and Rawls's two principles of justice, reference to which the parties plan to base their decisions.

Following Rawls, we assume that the parties model society as consisting largely of two groups of people: MAG (the more advantaged group) and LAG (the less advantaged group) (Rawls 2001: 61). The parties further assume that the members of MAG and LAG both have 'physical needs and psychological capacities within the normal range' (Rawls 1971 [1999]: 83–84, see also Rawls 1974*b*). The parties reflect this by assuming that both MAG and LAG share *the same* utility functions for income/wealth, which, according to Rawls, serves as the 'first approximation' of the index of primary goods one enjoys (Rawls 1971 [1999]: 53). Let $u_M^{Income} : \mathbb{R}_+ \to \mathbb{R}$ denote MAG's utility function for income and let $u_L^{Income} : \mathbb{R}_+ \to \mathbb{R}$ denote LAG's utility function for income. Following Rawls's 'normatively constructed' utility function,[15] we assume that the parties characterize

[15] Rawls's normatively constructed utility function is introduced on p. 108 of *Justice as Fairness – A Restatement* (see Figure 1: Rawls's Normatively Constructed Utility Function).

According to Rawls, '[t]his constructed utility function is based on the needs and requirements of citizens – their fundamental interests – conceived as such persons; it is not based on people's

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

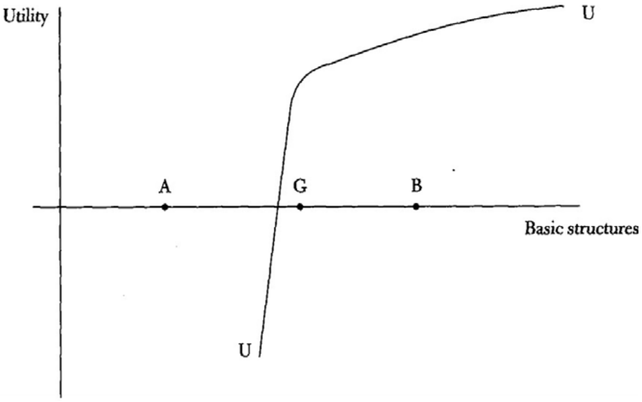

Figure 1. Rawls's normatively constructed utility function (Rawls 2001: 108).

both MAG's and LAG's utility functions for income as being *strictly concave*.[16] In particular, suppose that the parties assume that, for any given level of income $x \in \mathbb{R}_+$, both MAG's and LAG's experienced welfare level (i.e. utility) is the same as the square-root of their income/wealth level x: hence, we have for all $x \in \mathbb{R}_+$, $u_M^{Income}(x) = u_L^{Income}(x) = \sqrt{x}$.

We now add production into the model.[17] As Rawls explains, '[s]ocial cooperation, we assume is always productive, and without cooperation there would be nothing produced and so nothing to distribute' (Rawls 2001: 61). Suppose that MAG and LAG are each endowed with $T > 0$ hours of freedom or leisure time. The social cooperation between MAG and LAG allows members of each group to use a portion of their leisure time to work and earn income. Let $L_M \in [0,\ T]$ denote the working time spent by MAG and let $L_L \in [0,\ T]$ denote the working time spent by LAG. Let us now represent MAG's and LAG's respective productivity by the following income functions, $x_M(L_M)$ and $x_L(L_L)$, which represent the amount of income earned as a function of each group's working time:

- $x_M(L_M) = \beta\, L_M$ and
- $x_L(L_L) = L_L$.

We assume $\beta > 1$. Hence, not only do members of each group earn income proportional to their working time, the members of MAG earn $\beta > 1$ times more income than members of LAG for the same working time spent. Following Rawls, we may think of the members of MAG as the 'entrepreneurs' and the members of LAG as the 'unskilled

actual preferences and interests' but rather 'the parties use a utility function … so constructed as to reflect the ideal normative conceptions used to organize justice as fairness' (Rawls 2001: 107).

[16] A real-valued function $u : \mathbb{R}^n \to \mathbb{R}$ is strictly concave (resp. convex) if for all $x, y \in \mathbb{R}^n$ and for all $\alpha \in (0, 1)$, we have: $u(\alpha x + (1-\alpha)y) > \alpha u(x) + (1-\alpha)u(y)$ (resp. $u(\alpha x + (1-\alpha)y) < \alpha u(x) + (1-\alpha)u(y)$). In words, a real-valued function is strictly concave (resp. convex) if the functional value of any weighted average over the two points is greater (resp. smaller) than the weighted average of the functional values of the same two points.

[17] This is different from Moreno-Ternero and Roemer (2008) or Chung (2020), where there is no production in the model.

workers' (Rawls 1971 [1999]: 69) and think of β as reflecting MAG's relative wage when LAG's wage is normalized to 1. This is the way in which the members of LAG, despite being normal, are modelled as being relatively disadvantaged to members of MAG: members of MAG are more productive in earning income than members of LAG for the same working time. Given MAG's and LAG's respective working times, $L_M, \ L_L \in [0, \ 1]$, we define MAG's and LAG's utility functions for 'leisure' or 'freedom' as follows:

- $u_M^{Leisure}(L_M) = T - L_M$
- $u_L^{Leisure}(L_L) = T - L_L.$

From this, let us define the 'total well-being' of MAG and LAG as the 'sum' of the welfare generated from both 'income' and 'leisure (or freedom)'. Hence, the total welfare functions of MAG and LAG can be written as:

- $TW_M(x_M(L_M), \ L_M) = u_M^{Income}(x_M(L_M)) + u_M^{Leisure}(L_M) = \sqrt{x_M(L_M)} + [T - L_M]$
- $TW_L(x_L(L_L), \ L_M) = u_L^{Income}(x_L(L_L)) + u_L^{Leisure}(L_L) = \sqrt{x_L(L_L)} + [T - L_L].$

Suppose that the parties in the original position are concerned about choosing principles of justice that determine the society's basic structure, which, in turn, determines its redistributive tax rate $t \in [0, \ 1]$. Redistributive taxes are levied solely on MAG's earned income, which is then transferred to LAG. So, if members of MAG spend L_M of their leisure as working time, they earn βL_M amount of income, from which they contribute $\beta t L_M$ for redistributive taxation, the amount of which is then transferred to LAG. So, given that members of MAG use L_M of their leisure as working time and members of LAG use L_L of their leisure as working time, the after-tax income that each group receives is:

- $x_M(L_M) = \beta \, (1 - t) L_M$
- $x_L(L_L) = L_L + \beta t L_M.$

Following Rawls, the parties in the original position assume that members of both MAG and LAG are 'mutually disinterested' in the sense that 'they are not willing to have their interests sacrificed to the others' (Rawls 1971 [1999]: 112). Hence, the members of MAG and LAG will be assumed to choose their working time so that they strike the optimum work-and-life/freedom balance that maximizes their *own* total well-being, consisting of the sum of their income-welfare and their leisure/freedom-welfare, given the redistributive tax rate $t \in [0, \ 1]$ imposed by the basic structure of their society. Assuming that the members of both groups respond and behave optimally to the socially imposed redistributive tax rate, different principles of justice will tend to prescribe different basic structures that implement different redistributive tax rates for the purpose of achieving their particular normative aims. Utilitarianism will prescribe a basic structure and its corresponding redistributive tax rate t_U^* so that the sum total of MAG's and LAG's welfare levels is maximized. Rawls's difference principle will prescribe a basic structure and

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

its corresponding redistributive tax rate t_R^* so that LAG's income/wealth level (which, according to Rawls, serves as the 'first approximation' (Rawls 1971 [1999]: 53) of the index of primary social goods that members of LAG enjoy) is maximized.[18]

4.2. Results and analysis

We first consider how MAG and LAG will optimally allocate their working time given that they face a redistributive tax rate $t \in [0,\ 1)$.

Proposition 1: Suppose $T > \frac{\beta}{4}$. Let $t \in [0,\ 1]$ be the redistributive tax rate. Then, MAG's optimal working time (L_M^*) and LAG's optimal working time (L_L^*) are:

- $L_M^*(t) = \frac{\beta(1-t)}{4}$
- $L_L^*(t) = \frac{1-\beta^2 t(1-t)}{4}$.

We can see from Proposition 1 that MAG's optimal working time is decreasing in the redistributive tax rate t – i.e. a higher redistributive tax rate t will induce the higher-productive group, MAG, to work less, which is what we would normally expect. Note that by differentiating $L_M^*(t)$ with respect to t, we get $L_M^{*\prime}(t) = -\frac{\beta}{4}$. This implies that the members of MAG will react more sensitively and, thereby, decrease their optimal working time more rapidly to a given increase in the redistributive tax rate t as they become more productive relative to the members of LAG. In other words, the disincentivizing effect of the redistributive tax rate of inducing the members of MAG to work less is stronger the more productive they are. Interestingly, the way LAG reacts to the redistributive tax rate t is less straightforward. By differentiating $L_M^*(t)$ with respect to t, we get $L_L^{*\prime}(t) = \frac{\beta^2}{2}t - \frac{\beta^2}{4}$, which is *negative* when $t < \frac{1}{2}$ and *positive* when $t > \frac{1}{2}$. In other words, just like it was the case for MAG, increasing the redistributive tax rate t will induce the members of LAG to work less *given that the redistributive tax rate* t *does not exceed* $\frac{1}{2}$. Of course, since the redistributive tax is imposed only on MAG's and not on LAG's earned income, the particular way in which increasing the redistributive tax rate t induces each group to work less is different: for the members of MAG, a higher redistributive tax rate t disincentivizes to work because a higher redistributive tax rate means that they will earn less after-tax income for the same working time, which incentivizes them to allot more time to leisure instead of work; for the members of LAG, a higher redistributive tax rate t disincentivizes to work because, with a higher redistributive tax rate, a greater portion of MAG's earned income will be redistributed to LAG, which makes it possible for LAG to achieve the same level of disposable income through redistributive subsidy while working less. However, once the redistributive tax rate

[18] Many people conflate well-being/welfare and primary social goods when commenting on the different principle (Moreno-Ternero and Roemer 2008; Buchak 2017; Gustafsson 2018). However, this is a key factor that must be kept distinct as completely different distributional prescriptions will follow depending on which conception of advantage one uses to apply different principles of distributional justice. See Chung (2021) for a criticism of Gustafsson (2018) on this matter.

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

t exceeds $\frac{1}{2}$, increasing it further will actually induce the members of LAG (but not the members of MAG)[19] to *work more*. The intuition is that once the redistributive tax rate t becomes too high, this will disincentivize the members of MAG to earn income so much that the portion of MAG's income that gets collected for the purpose of redistributive taxation to subsidize LAG will no longer be enough to meet LAG's optimal income level, which, in turn, forces LAG to supplement their income shortage through their own labour. From the optimal working times of MAG and LAG that we derived from Proposition 1, we are able to characterize MAG's and LAG's after-tax income as follows:

Corollary of Proposition 1: Suppose $T > \frac{\beta}{4}$. Let $t \in [0,\ 1]$ be the redistributive tax rate. Then, MAG's and LAG's respective after-tax incomes are:

- $x_M^*(t) = \frac{\beta^2(1-t)^2}{4}$
- $x_L^*(t) = \frac{1}{4}$.

An interesting thing to note is that the redistributive tax rate $t \in [0,\ 1]$ only affects MAG's (and not LAG's) after-tax income. Specifically, MAG's after-tax income is $\frac{\beta^2(1-t)^2}{4}$, which we can see is decreasing in the redistributive tax rate $t \in [0,\ 1]$. In other words, with a higher redistributive tax rate, the members of MAG will work less and, as a result, enjoy less tangible after-tax income. By contrast, LAG's after-tax income is fixed at $x_L^*(t) = \frac{1}{4}$. As we have already seen, the redistributive tax rate t affects both MAG's and LAG's optimal working time. Since LAG's after-tax income remains fixed at $x_L^*(t) = \frac{1}{4}$ regardless of the redistributive tax rate t, this means that any portion of income that falls short of $\frac{1}{4}$ will be fully subsidized and financed through redistribution taxation imposed on MAG. Another implication of LAG's after-tax income being fixed at $x_L^*(t) = \frac{1}{4}$ is that increasing the redistributive tax rate t does not necessarily raise the index of primary social goods enjoyed by members of LAG; rather, what raising the redistributive tax rate does is that it reduces the relative gap in the indices of primary social goods enjoyed by members of MAG and members of LAG by lowering MAG's after-tax income.

Although LAG ends up receiving a fixed after-tax income, namely, $x_L^*(t) = \frac{1}{4}$, independent of the redistributive tax rate t, this does not mean that redistributive taxation does not affect LAG's total well-being. As a matter of fact, given that the redistributive tax rate t is below $\frac{1}{2}$, an increase in redistributive taxation improves LAG's total well-being while at the same time reducing MAG's total well-being. This is so because although redistributive taxation does not increase LAG's after-tax income, it provides more leisure time to the members of LAG, which allows them to earn the same after-tax income while working less. Conversely, increasing redistributive taxation decreases the total well-being of the members of MAG by decreasing the amount of tangible after-tax income earned through their labour.[20]

[19] The members of MAG will continue to reduce their working time as the redistributive tax rate t increases, and when $t = 1$, then they will not work at all!

[20] However, once the redistributive tax rate t exceeds $\frac{1}{2}$, further increasing the redistributive tax rate will reduce not only MAG's total well-being, but also LAG's total well-being as well. This is so because, as explained after Proposition 1, a redistributive tax rate that is too high will force the members of LAG to

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

So, when the redistributive tax rate t is below $\frac{1}{2}$, we are faced with a familiar trade-off: by raising the redistributive tax rate t, we can increase the total well-being of the members of LAG at the expense of lowering the total well-being of the members of MAG.[21] Faced with such a trade-off, how would utilitarianism and Rawls's two principles of justice prescribe the society's redistributive tax rate t? And under which principle of justice would the members of both MAG and LAG find it better to live? Again, utilitarianism will prescribe a redistributive tax rate t^*_U that maximizes the sum of total well-being enjoyed by MAG and LAG; while Rawls's justice as fairness will prescribe a redistributive tax rate t^*_R that maximizes the minimum after-tax income earned either by LAG or MAG.

Given our understanding of what each principle of justice requires and assuming that both MAG and LAG choose their working time that strikes the optimum balance of after-tax income and leisure as a response to the specific redistributive tax rate they face, let us now derive the redistributive tax rate t^*_U that utilitarianism prescribes and the Rawlsian redistributive tax rate t^*_U that Rawls's difference principle prescribes:

Proposition 2: The utilitarian redistributive tax rate is: $t^*_U = \frac{\beta-1}{2\beta}$.

Proposition 3: The Rawlsian redistributive tax rate is: $t^*_R \leq \frac{\beta-1}{\beta}$.

Proposition 2 shows that the utilitarian redistributive tax rate is uniquely determined by $t^*_U = \frac{\beta-1}{2\beta}$. By contrast, given that we follow Rawls and measure a person's advantage as the person's index of primary social goods, which in our case is identified with each group's after-tax income, Proposition 3 shows that there exists a range of redistributive tax rates (namely, $t^*_R \leq \frac{\beta-1}{\beta}$) that would be compatible with the prescriptions of the difference principle, which seeks to maximize the after-tax income of the lower income group, LAG. Among the range of redistributive tax rates that are compatible with Rawls's difference principle, let $t^*_{\hat{R}} = \frac{\beta-1}{\beta}$ denote the 'maximum Rawlsian redistributive tax rate'.

The maximum Rawlsian redistributive tax rate $t^*_{\hat{R}} = \frac{\beta-1}{\beta}$ is the redistributive tax rate from which the members of LAG can achieve the highest total well-being under the basic structure prescribed by Rawls's two principles of justice given that the redistributive tax rate is below $\frac{1}{2}$. This is so because the maximum Rawlsian redistributive tax rate allows the members of LAG to enjoy the most amount of leisure time (or freedom) without reducing their disposable after-tax income. Given that leisure is an essential component to exercise the *fair worth* (Rawls 1971 [1999]: 179) of the several basic liberties guaranteed by the first principle of justice, which Rawls deemed fundamentally important, we might think that a basic

put in additional working hours to compensate for the shortage of income that was not adequately provided through redistributive taxation levied on MAG, and as LAG's after-tax income is fixed at $x^*_L(t) = \frac{1}{4}$, putting in more working time means that there will be less time spent for leisure, which results in an overall reduction of LAG's total well-being.

[21] However, when the redistributive tax rate t exceeds $\frac{1}{2}$, any further increase in the redistributive tax rate will only make *everybody* worse-off. See previous footnote.

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

structure organized by Rawls's two principle of justice *taken together* would adopt, among the continuum of redistributive tax rates compatible with Rawls's difference principle taken in isolation, the maximum Rawlsian redistributive tax rate $t^*_{\hat{R}} = \frac{\beta-1}{\beta}$. Furthermore, given that the redistributive tax rate is below $\frac{1}{2}$, any redistributive tax rate that is lower than the maximum Rawlsian redistributive tax rate will increase the total welfare of MAG at the expense of lowering the total welfare of LAG. Hence, from the Rawlsian perspective, any redistributive tax rate that is below the maximum Rawlsian redistributive tax rate $t^*_{\hat{R}} = \frac{\beta-1}{\beta}$ would be *prima facie* objectionable on grounds that it does not take seriously the separateness of persons as discussed previously. So, let us assume that among the range of redistributive tax rates compatible with Rawls's difference principle, the parties assume that Rawls's two principles of justice prescribe the *maximum* Rawlsian redistributive tax rate $t^*_{\hat{R}} = \frac{\beta-1}{\beta}$.

By comparing the utilitarian tax rate $t^*_U = \frac{\beta-1}{2\beta}$ and the maximum Rawlsian redistributive tax rate $t^*_{\hat{R}} = \frac{\beta-1}{\beta}$, we can see that the maximum Rawlsian redistributive tax rate $t^*_{\hat{R}} = \frac{\beta-1}{\beta}$ is higher than that of the utilitarian tax rate $t^*_U = \frac{\beta-1}{2\beta}$; specifically, it is *twice as high* as the utilitarian tax rate inside our model. Bearing in mind such a difference in the redistributive tax rates prescribed by utilitarianism and Rawls's two principles of justice, let us now examine what levels of total well-being the members of MAG and LAG will experience under these two alternative basic structures:

Corollary of Proposition 2: Under the utilitarian redistributive tax rate $t^*_U = \frac{\beta-1}{2\beta}$, the total well-being achieved by MAG and LAG are:

- $TW_M\left(t^*_U = \frac{\beta-1}{2\beta}\right) = \frac{\beta+1}{8} + T$
- $TW_L\left(t^*_U = \frac{\beta-1}{2\beta}\right) = \frac{\beta^2+3}{16} + T.$

Corollary of Proposition 3: Under the maximum Rawlsian redistributive tax rate $t^*_{\hat{R}} = \frac{\beta-1}{\beta}$, the total well-being achieved by MAG and LAG are:

- $TW_M\left(t^*_{\hat{R}} = \frac{\beta-1}{\beta}\right) = \frac{1}{4} + T$
- $TW_L\left(t^*_{\hat{R}} = \frac{\beta-1}{\beta}\right) = \frac{\beta}{4} + T$

Remember that the parameter $\beta > 1$ represents the productive advantage that MAG has relative to LAG: specifically, it represents MAG's relative wage when LAG's wage is normalized to 1. Under the basic structure organized by utilitarianism, we can see that the total well-being of *both* MAG and LAG is increasing in the parameter β; that is, both MAG's and LAG's total well-being levels increase as the productive advantage of MAG relative to LAG becomes greater. This is in alignment with Rawls's view of regarding 'society as a cooperative venture for mutual advantage' (Rawls 1971 [1999]: 73–74) as the productive advantage of MAG makes everybody – i.e. not just members of MAG, but also members of

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

LAG – better off. Furthermore, between MAG and LAG, it is LAG's (and not MAG's) total well-being that increases more rapidly as MAG becomes more relatively productive. This is consonant with Rawls's emphasis on 'reciprocity' that we have discussed previously.

By contrast, under the basic structure organized by Rawls's two principles of justice, it is only LAG's (and not MAG's) total well-being that improves when MAG's relative productive advantage increases; this is so because the results show that the members of MAG experience a *fixed* level of total well-being [namely, $TW_M\left(t^*_{\hat{R}} = \frac{\beta-1}{\beta}\right) = \frac{1}{4} + T$] regardless of how productive they are relative to the members of LAG. This means that there is a sense in which the members of MAG are being sacrificed under a basic structure that is organized by Rawls's two principles of justice, as their productive talents are being socially utilized entirely for the purpose of improving the total well-being of another group, LAG. Of course, this does not necessarily make Rawls's two principles of justice subject to the separateness of persons objection because, as explained previously, the separateness of persons objection applies asymmetrically only in one direction – that is, only when the members of LAG are sacrificed to achieve the greater overall good of society as a whole and in particular the members of MAG.

One thing to note is that since $\beta > 1$ by assumption (i.e. since we are assuming that MAG is more productive than LAG), we will always have: $TW_M\left(t^*_U = \frac{\beta-1}{2\beta}\right) = \frac{\beta+1}{8} + T > \frac{1}{4} + T = TW_M\left(t^*_{\hat{R}} = \frac{\beta-1}{\beta}\right)$. That is, the members of MAG will always experience a higher level of total well-being under a basic structure organized by utilitarianism than what they would experience under a basic structure organized by Rawls's two principles of justice (which is not too surprising). Another thing to note is that although it is true that LAG starts out better-off under a basic structure organized by Rawls's two principles of justice than what they would be under a basic structure organized by utilitarianism, as the members of MAG become more and more productive, the total well-being of LAG increases *quadratically* (i.e. it increases at a faster and faster rate) under utilitarianism, while it increases *linearly* (i.e. it increases at a constant rate) under Rawls's two principles of justice. From this, we can conjecture that once the relative gap between MAG's and LAG's productive abilities becomes sufficiently large – that is, as the members of MAG (the entrepreneurs) become more and more productive relative to the members of LAG (the unskilled workers) – there will exist a point after which utilitarianism will start to outperform Rawls's two principles of justice in improving the total well-being of members of LAG. This conjecture is confirmed in the following result:

Proposition 4: Suppose $1 < \beta < 3$. Then,

(1) $t^*_U < t^*_{\hat{R}}$
(2) $TW_M\left(t^*_U\right) > TW_M\left(t^*_{\hat{R}}\right)$
(3) $TW_L\left(t^*_U\right) < TW_L\left(t^*_{\hat{R}}\right)$.

Now, suppose $\beta \geq 3$. Then,

(4) $t_U^* < t_{\hat{R}}^*$
(5) $TW_M(t_U^*) > TW_M\left(t_{\hat{R}}^*\right)$
(6) $TW_L(t_U^*) \geq TW_L\left(t_{\hat{R}}^*\right)$ [and $TW_L(t_U^*) > TW_L\left(t_{\hat{R}}^*\right)$ when $\beta > 3$.]

The first set of results [from (1) to (3)] summarize what happens under the two alternate basic structures when the relative productive gap between MAG and LAG is not so great (viz. when $1 < \beta < 3$). In such situations, the results show that through implementing a higher redistributive tax rate, a basic structure that is organized by Rawls's two principles of justice will provide a higher level of total well-being for members of LAG than what they would expect under a basic structure organized by utilitarianism. By contrast, as already explained, the total well-being of MAG is greater under a basic structure organized by utilitarianism than it is under a basic structure organized by Rawls's two principles of justice. So far, the results simply confirm our familiar intuition – namely, that a basic structure organized by utilitarianism maximizes aggregate social welfare at the expense of lowering the total well-being of the lesser advantaged group than what they could achieve under alternate social arrangements, in particular, under a basic structure organized by Rawls's two principles of justice. This makes utilitarianism subject to the separateness of persons objection that we have discussed previously.

However, the next set of results [from (4) to (6)] shows that such a familiar intuition no longer applies when the relative productive gap between MAG and LAG becomes sufficiently large (viz. when $\beta \geq 3$). Here, it is not only the members of MAG, but also the members of LAG who are better-off under a basic structure organized by utilitarianism than a basic structure organized by Rawls's two principles of justice. In other words, when MAG is sufficiently more productive relative to LAG, not only does a basic structure organized by utilitarianism maximize aggregate social welfare, it also allows *everybody* to experience a higher level of overall well-being than what each would expect to experience under a basic structure organized by Rawls's two principles of justice. This means that the higher redistributive tax rate that is imposed on the members of MAG under a basic structure organized by Rawls's two principles of justice does not really protect the interests of LAG as it was initially designed to do, but, instead, simply levels everybody down, when the relative productive gap between MAG and LAG is sufficiently great. Recall the Principle of Dominance:

- **The Principle of Dominance:** Given a choice between any two options X and Y, if option X *dominates* options Y – that is, if X generates outcomes that are no worse than Y in all possible states and in some states generates outcomes that are strictly better than Y, – then rationality requires one to choose X over Y.

As already explained, one practical limitation of the Principle of Dominance is its scope – i.e. the principle of dominance is seldom practically applicable because in most decision situations with a choice of two considered options, there are few cases

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

in which one option clearly dominates the other. However, in those rare cases in which one option does dominate the other, the application of the Principle of Dominance is uncontroversial as it is directly implied by instrumental rationality.

We can see that when the relative productive gap between MAG and LAG is sufficiently large (viz. when $\beta \geq 3$), the choice between a basic structure organized by utilitarianism and one organized by Rawls's two principles of justice turns into one of those rare decision problems to which the Principle of Dominance *can* be applied. Specifically, from the perspective of the original contracting parties behind the veil of ignorance, the choice is between utilitarianism and Rawls's two principles of justice, and, for each choice option, the two possible outcomes that they must consider are either being born as a member of MAG or as a member of LAG. When the relative productive gap between MAG and LAG is sufficiently large (in particular, when $\beta > 3$), utilitarianism strictly Pareto dominates Rawls's two principles of justice as a basic structure organized by utilitarianism will generate total well-being levels that are strictly greater than those generated by Rawls's two principles of justice for *both* MAG and LAG. In such situations, unlike what Rawls had initially presupposed, choosing utilitarianism does *not* 'lead to institutions that the parties would find intolerable' (Rawls 1971 [1999]: 135). On the contrary, utilitarianism leads to institutions that make everybody better off, which, in turn, liberates it from the separateness of persons objection as it no longer requires the members of LAG 'to accept the greater advantages of others as a sufficient reason for lower expectations over the whole course of [their] li[ves]' (Rawls 1971 [1999]: 155).

I would like to emphasize that this result does not substantially change even if we assume that the parties in the original position primarily care about securing the highest index of primary social goods (i.e. income) instead of securing the highest level of total well-being. Recall from Corollary of Proposition 1 that given the redistributive tax rate t, MAG's after-tax income is: $x_M^*(t) = \frac{\beta^2(1-t)^2}{4}$, while LAG's after-tax income is fixed at: $x_L^*(t) = \frac{1}{4}$. Hence, under the utilitarian redistributive tax rate $t_U^* = \frac{\beta-1}{2\beta}$, MAG's after-tax income becomes $\frac{(\beta+1)^2}{16}$, while LAG's after-tax income becomes $\frac{1}{4}$. Similarly, under the maximum Rawlsian redistributive tax rate $t_{\hat{R}}^* = \frac{\beta-1}{\beta}$, both MAG's and LAG's after-tax income becomes $\frac{1}{4}$. Since $\beta > 1$ (i.e. since we are assuming that MAG is more productive than LAG), we have $x_M^*\left(t_U^* = \frac{\beta-1}{2\beta}\right) = \frac{(\beta+1)^2}{16} > \frac{1}{4} = x_M^*\left(t_{\hat{R}}^* = \frac{\beta-1}{\beta}\right)$ for any $\beta > 1$. (In particular, we do not need $\beta \geq 3$, that is, we do not need to assume that the relative productive gap between MAG and LAG is sufficiently large). In other words, the members of MAG earn strictly more after-tax income under utilitarianism than what they would expect to earn under Rawls's two principles of justice; while the amount of after-tax income earned by the members of LAG is the same in both alternative basic structures. Hence, we can see that utilitarianism still dominates Rawls's two principles of justice even if we use the index of primary goods as a measure of people's advantage instead of people's total welfare levels. Moreover, this is so regardless of how large MAG's and LAG's relative productive gap is.

Based on the discussion so far, we can understand that given that the relative productivity of MAG is sufficiently greater than that of LAG (specifically, whenever $\beta > 3$) utilitarianism simply dominates Rawls's two principles of justice both in terms of total well-being as well as in terms of the index of primary goods in our model. In such cases, the issue of which specific normative decision rule (e.g. expected utility maximization, risk-weighted expected utility maximization, maximin, etc.) it would be reasonable for the parties in the original position to adopt becomes peripheral as the parties can simply rely on the Principle of Dominance (which is directly implied by instrumental rationality) to choose utilitarianism. Furthermore, in doing so, we are able to avoid the unnecessary controversies regarding probability assignments in the original position as (just like the maximin rule) the application of the Principle of Dominance does not require any probabilistic information. Of course, if the relative productive gap between MAG and LAG is not too great (i.e. if $1 < \beta < 3$), then the members of LAG would be better-off (not necessarily in terms of the index of primary goods, but in terms of total well-being) under a basic structure organized by Rawls's two principles of justice than a basic structure organized by utilitarianism. So, then, an important part of the debate hinges on whether or not it would be reasonable for the parties in the original position to assume, on the basis of their knowledge of 'general facts about human society' (Rawls 1971 [1999]: 119), that the relative productive gap between MAG and LAG is sufficiently large (in particular, $\beta > 3$) in their model society.

Of course, figuring out the relative productive gap between MAG and LAG is primarily an empirical issue that goes outside of the purview of this paper. But I would like to briefly point out that according to the Human Development Reports published by the United Nations Development Programme, the ratio of the average income of the richest 20% of the population to the average income of the poorest 20% of the population (which roughly corresponds to the parameter β in our model) in the USA during the years 2010–2017 was 9.4.[22] Among the roughly 150 countries whose data were displayed in the Human Development Reports, no country had a lower ratio than 3 (with Ukraine having the lowest ratio of 3.5). This suggests that it would not be too unreasonable for the parties in the original position to assume that the relative productive gap between MAG and LAG in their model society would be sufficiently large (i.e. $\beta > 3$) and that such an assumption would be largely consistent with the existing empirical data to which the original contracting parties would have access given their knowledge of 'general facts about human society' (Rawls 1971 [1999]: 119).

5. Publicity and stability

To this, somebody might argue that since the utilitarian redistributive tax rate t_U^* generates a higher total well-being for members of LAG than the maximum Rawlsian redistributive tax rate $t_{\hat{R}}^*$ when the relative productive gap between MAG and LAG is sufficiently great (i.e. $\beta > 3$), it is actually the utilitarian redistributive tax rate t_U^* and not the maximum Rawlsian redistributive tax rate $t_{\hat{R}}^*$ that Rawls's two principles of justice will prescribe. This type of response is not available for Rawls due to his firm commitment to the 'publicity condition'. The publicity condition requires that the endorsement of any conception of justice must be 'public' in

[22] See http://hdr.undp.org/en/indicators/135106.

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

the sense that it is 'widely known' and 'explicitly recognized' (Rawls 1971 [1999]: 115) that such a conception of justice is 'publicly accepted and followed as the fundamental charter of society' (Rawls 1971 [1999]: 158). This means that a society's conception of justice must not be 'esoteric' in the sense that its fundamental aims are pursued indirectly in public disguise in a way that only a few political elites understand. For Rawls, if a basic structure is widely publicized to be utilitarian [resp. Rawlsian], then it is publicly affirmed to be a utilitarian [resp. Rawlsian] society.

Remember that the utilitarian redistributive tax rate t_U^* was designed with an explicit aim to maximize aggregate social welfare. This means that a society that implements the utilitarian tax rate t_U^* will be defined, by the publicity condition, as a 'utilitarian' society even if such a society happens to better meet the demands of justice as fairness, and, in particular, the difference principle by providing a higher overall well-being to members of LAG than a society that implements the maximum Rawlsian redistributive tax rate $t_{\hat{R}}^*$. At this point, it is instructive to recall how Rawls rejected utilitarianism's potential resort to esotericism (or what some have called 'indirect utilitarianism') on grounds of publicity in *A Theory of Justice*:

> suppose that the average utility is actually greater should the two principles of justice be publicly affirmed and realized in the basic structure. For the reasons mentioned, this may conceivably be the case. ... The utilitarian cannot reply that one is now really maximizing the average utility. In fact, the parties would have chosen the two principles of justice. (Rawls 1971 [1999]: 158)

Given that the relative productive gap between MAG and LAG is sufficiently great (i.e. $\beta > 3$), we can repeat exactly the same argument with the roles of the difference principle and the principle of average utility reversed:

> suppose that LAG's overall well-being is actually greater should utilitarianism be publicly affirmed and realized in the basic structure. For the reasons mentioned, this may conceivably be the case. ... The Rawlsian cannot reply that one is now really implementing the difference principle. In fact, the parties would have chosen the principle of utility.

A closely related issue of (psychological) stability can also be easily taken care of in a similar manner. For instance, one might argue that even if the parties in the original position can choose utilitarianism on the basis of their knowledge of general facts about human society combined with the Principle of Dominance, utilitarianism, once practically implemented, may be more unstable than Rawls's two principles of justice. According to Rawls, 'A conception of justice is stable when the public recognition of its realization' 'generates its own support' (Rawls 1971 [1999]: 155). The reason that Rawls thought that utilitarianism will be unstable is that

> When the principle of utility is satisfied ... there is no such assurance that everyone benefits. Allegiance to the social system may demand that some, particularly the less favored, should forgo advantages for the sake of the greater good of the whole. Thus the scheme will not be stable unless those who must make sacrifices strongly identify with interests broader than their

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

> own. … Even when we are less fortunate, we are to accept the greater advantages of others as a sufficient reason for lower expectations over the whole course of our life. This is surely an extreme demand. (Rawls 1971 [1999]: 155)

Here, we can reaffirm Rawls's presumption that utilitarianism requires, for the sake of maximizing average utility, further sacrifices from members of LAG than what he expects his two principles of justice require. The point is that, because of this, utilitarianism will be unstable as it will fail to elicit ongoing allegiance to its political system from all of its members, especially, from the members of LAG. As argued in the previous section, the presumption that utilitarianism imposes further sacrifices on members of LAG to maximize average utility is false whenever $\beta > 3$; on the contrary, we have seen that whenever $\beta > 3$, we do have assurance that everyone benefits and will enjoy a higher total well-being than what each could expect to enjoy under Rawls's two principles of justice. When this is so, unlike what Rawls thinks, utilitarianism does not require members of LAG to have 'a greater identification with the interests of others than the two principles of justice' (Rawls 1971 [1999]: 154). Hence, as long as it is well-publicized that utilitarianism, by aiming to maximize society's aggregate social welfare, will actually better improve the overall well-being of the members of LAG than Rawls's two principles of justice, there is no reason why we should think that the members of LAG will withdraw their support for their utilitarian basic structure. This means that, when $\beta > 3$, utilitarianism will be stable in exactly the same sense Rawls thought his two principles of justice would be stable: 'Since everyone's good is affirmed, all acquire inclinations to uphold the scheme' (Rawls 1971 [1999]: 155).

5. Concluding remarks: the value of the economic approach

In this paper, I have tried to propose a potential way to resolve the debate concerning which principle of distributive ethics will eventually be justified from the original position. Unlike many research papers that have primarily taken an individual decision-theoretic approach to the original position, this paper has attempted to model the economy and basic structure that allow the original contracting parties to project the economic consequences and the overall well-being of the two groups, MAG and LAG, as a result of living under different basic structures respectively designed by utilitarianism and Rawls's two principles of justice. Of course, the basic model employed in this paper is not the only way to model the economy and the basic structure of society. What is important is to realize that Rawls's two principles of justice do not always better improve the overall well-being of LAG than utilitarianism, and one value of taking the economic approach is that it helps us understand the specific empirical conditions under which utilitarianism not only maximizes aggregate social welfare, but also better improves the overall well-being of LAG than Rawls's two principles of justice. Once this is done, we can then resort to the academic division of labour to allow the various empirical sciences to help us understand whether our actual societies meet those empirical conditions.

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

Acknowledgements. This research has been generously funded by the 40th Sakurada-kai Political Science Research Grant (No. 16) 櫻田會「政治研究助成 (助第40-16号)」.

References

Benbaji Y. 2005. The doctrine of sufficiency: a defense. *Utilitas* **17**, 310–332.

Buchak L. 2017. Taking risks behind the veil of ignorance. *Ethics* **127**, 610–644.

Chung H. 2017. Prospect utilitarianism: a better alternative to sufficientarianism. *Philosophical Studies* **174**, 1911–1933.

Chung H. 2020. Rawls's self-defeat: a formal analysis. *Erkenntnis* **85**, 1169–1197.

Chung H. 2021. On choosing the difference principle behind the veil of ignorance. *Journal of Philosophy* **118**, 450–463.

Chung H. Forthcoming. Prospect utilitarianism and the original position. *Journal of the American Philosophical Association*.

Dworkin R. 1981*a*. What is equality – Part 1: equality of welfare. *Philosophy & Public Affairs* **10**, 185–246.

Dworkin R. 1981*b*. What is equality – Part 2: equality of resources. *Philosophy & Public Affairs* **10**, 283–345.

Gustafsson J.E. 2018. The difference principle would not be chosen behind the veil of ignorance. *Journal of Philosophy* **115**, 588–604.

Harsanyi J. 1953. Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy* **61**, 434–435.

Harsanyi J. 1955. Cardinal welfare, individualistic ethics, and the interpersonal comparisons of utility. *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. *American Political Science Review* **69**, 594–606.

Harsanyi J. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.

Hirose I. 2013. Aggregation and the separateness of persons. *Utilitas* **25**, 182–205.

Kahneman D. and A. Tversky 1979. Prospect theory: an analysis of decision under risk. *Econometrica* **47**, 263–291.

Moehler M. 2018. The Rawls–Harsanyi dispute: a moral point of view. *Pacific Philosophical Quarterly* **99**, 82–99.

Moreno-Ternero J. and J. Roemer 2008. The veil of ignorance violates priority. *Economics and Philosophy* **24**, 233–257.

Muldoon R., C. Lisciandra, M. Colyvan, G. Sillari and J. Sprenger 2014. Disagreement behind the veil of ignorance. *Philosophical Studies* **170**, 377–394.

Nozick R. 1974. *Anarchy, State, and Utopia*. New York, NY: Basic Books.

Parfit D. 1997. Equality and priority. *Ratio* **10**, 202–221.

Parfit D. 2012. Another defense of the priority view. *Utilitas* **24**, 399–440.

Rawls J. 1971 [1999]. *A Theory of Justice (Revised Edition)*. Cambridge, MA: Harvard University Press.

Rawls J. 1974a. Some reasons for the maximin criterion. *American Economic Review* **64**, 141–146.

Rawls J. 1974b. Reply to Alexander and Musgrave. *Quarterly Journal of Economics* **88**, 633–655.

Rawls J. 2001. *Justice as Fairness – A Restatement*. Cambridge, MA: Harvard University Press.

Roemer J. 1996. *Theories of Distributive Justice*. Cambridge, MA: Harvard University Press.

Roemer J. 2002. Egalitarianisms against the veil of ignorance. *Journal of Philosophy* **99**, 164–184.

Roemer J. 2009. *Equality of Opportunity*. Cambridge, MA: Harvard University Press.

Sen A. 1980. Equality of what? In *Tanner Lectures on Human Values, Vol.* **1**, ed. S. McMurrin. Cambridge: Cambridge University Press.

Stefansson H.O. 2021. Ambiguity aversion behind the veil of ignorance. *Synthese* **198**, 6159–6182.

Weymark J. 1991. A reconsideration of the Harsanyi–Sen debate on utilitarianism. In *Interpersonal Comparisons of Well-Being*, ed. J. Elster and J. Roemer, 255–320. Cambridge: Cambridge University Press.

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

Appendix: Statement and proofs of main results

Proposition 1: Suppose $T > \frac{\beta}{4}$. Let $t \in [0,\ 1]$ be the redistributive tax rate. Then, MAG's optimal working time (L_M^*) and LAG's optimal working time (L_L^*) are:

- $L_M^* = \frac{\beta(1-t)}{4}$
- $L_L^* = \frac{1-\beta^2 t(1-t)}{4}$.

Proof:
MAG solves the following maximization problem:

$$\max_{L_M \in [0,\ T]} \sqrt{\beta\ (1-t)L_M} + [T - L_M]$$

Note that the objective function (which is MAG's total welfare function) is strictly concave. Hence, the first-order condition will be sufficient to give us the value of L_M that uniquely maximizes MAG's total welfare. Taking the first-order condition of the objective function, we have:

$$\frac{1}{2} \cdot \frac{\beta(1-t)}{\sqrt{\beta(1-t)L_M}} - 1 = 0 \Rightarrow L_M^* = \frac{\beta(1-t)}{4}.$$

Since $T > \frac{\beta}{4}$, $L_M^* = \frac{\beta(1-t)}{4}$ is the unique interior solution to MAG's maximization problem. Similarly, LAG solves the following problem:

$$\max_{L_L \in [0,\ T]} \sqrt{L_L + \beta t L_M} + [T - L_M].$$

Given $L_M^* = \frac{\beta(1-t)}{4}$, LAG's maximization problem can be re-written as:

$$\max_{L_L \in [0,\ T]} \sqrt{L_L + \frac{\beta^2 t(1-t)}{4}} + [T - L_M].$$

Again, since the objective function (which is LAG's total welfare function) is strictly concave, the first-order condition will be sufficient to give us the value of L_L that uniquely maximizes LAG's total welfare. Taking the first-order condition of the objective function, we have:

$$\text{FOC}: \frac{1}{2\sqrt{\frac{1}{4}\beta^2 t(1-t) + L_L}} - 1 = 0 \Rightarrow L_L^* = \frac{1 - \beta^2 t(1-t)}{4.}$$

∎

Corollary of Proposition 1: Suppose $T > \frac{\beta}{4}$. Let $t \in [0,\ 1]$ be the redistributive tax rate. Then, MAG's and LAG's respective after-tax incomes are:

- $x_M^*(t) = \frac{\beta^2(1-t)^2}{4}$
- $x_L^*(t) = \frac{1}{4}$.

Proof:
Note that MAG's and LAG's after-tax incomes are:

- $x_M(L_M) = \beta\ (1-t)L_M$
- $x_L(L_L) = L_L + \beta t L_M$.

Plugging in $L_M^* = \frac{\beta(1-t)}{4}$ and $L_L^* = \frac{1-\beta^2 t(1-t)}{4}$ gives us the result. ∎

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

Proposition 2: The utilitarian redistributive tax rate is: $t_U^* = \frac{\beta-1}{2\beta}$.

Proof: By Proposition 1, given a redistributive tax rate $t \in [0,\ 1]$, we know that the optimum working times for MAG and LAG are:

- $L_M^* = \frac{\beta(1-t)}{4}$
- $L_L^* = \frac{1-\beta^2 t(1-t)}{4}$.

By Corollary of Proposition 1, this results in the following after-tax income for MAG and LAG:

- $x_M^*(t) = \frac{\beta^2(1-t)^2}{4}$
- $x_L^*(t) = \frac{1}{4}$.

This in turn generates the following indirect total welfare functions for MAG and LAG as a function of the redistributive tax rate $t \in [0,\ 1]$:

- $TW_M(t) = \sqrt{\frac{\beta^2(1-t)^2}{4}} + T - \frac{\beta(1-t)}{4}$
- $TW_L(t) = \sqrt{\frac{1}{4}} + T - \frac{1-\beta^2 t(1-t)}{4}$.

Utilitarianism chooses the redistributive tax rate $t \in [0,\ 1]$ that maximizes the sum of the total welfare levels of MAG and LAG. Hence, utilitarianism solves:

$$\max_{t\in[0,1]}[TW_M(t) + TW_L(t)]$$

$$= \max_{t\in[0,1]}\left[\sqrt{\frac{\beta^2(1-t)^2}{4}} + T - \frac{\beta(1-t)}{4}\right] + \left[\sqrt{\frac{1}{4}} + T - \frac{1-\beta^2 t(1-t)}{4}\right].$$

The objective function (which is the sum of two strictly concave functions) is again strictly concave. Hence, the first-order condition will be sufficient to give us the value of t^* that uniquely maximizes the sum of MAG's and LAG's total welfare levels given $t^* \in [0,\ 1]$. Taking the first-order condition, we have:

$$\text{FOC}: \frac{\beta^2-\beta}{4} - \frac{\beta^2}{2}t = 0 \ \Rightarrow \ t = \frac{\beta-1}{2\beta}.$$

Since $\beta > 1$, we have $0 < t = \frac{\beta-1}{2\beta} < 1$. Hence, we conclude that the utilitarianian tax rate is $t_U^* = \frac{\beta-1}{2\beta.}$ ∎

Proposition 3: The Rawlsian redistributive tax rate is: $t_R^* \leq \frac{\beta-1}{\beta}$.

Proof:

Rawls's difference principle prescribes a redistributive tax rate t_R^* so that the index of primary social goods enjoyed by either MAG or LAG, whose index of primary social goods is lower than the other, is maximized. In our case, the index of primary social goods is measured [as a 'first approximation' (Rawls 1971 [1999]: 53)] by each group's after-tax income. Hence, Rawls's difference principle will choose a redistributive tax rate t_R^* that maximizes the minimum after-tax income received by either MAG or LAG. By Corollary of Proposition 1, the after-tax income for MAG and LAG are:

- $x_M^*(t) = \frac{\beta^2(1-t)^2}{4}$
- $x_L^*(t) = \frac{1}{4}$.

Claim: It is not the case that the Rawlsian redistributive tax rate t_R^* is $t_R^* > \frac{\beta-1}{\beta}$. To see this, suppose $t_R^* > \frac{\beta-1}{\beta}$. Note that if $\frac{\beta-1}{\beta} < t < 1$, then $x_M^*(t) = \frac{\beta^2(1-t)^2}{4} < \frac{1}{4} = x_L^*$. Hence, if $t > \frac{\beta-1}{\beta}$, then, between MAG and LAG, the members of MAG have a lower after-tax income than the members of LAG.

https://doi.org/10.1017/S0266267122000098 Published online by Cambridge University Press

So, $t_R^* > \frac{\beta-1}{\beta}$ must be the (Rawlsian) redistributive tax rate that maximizes MAG's after-tax income. Pick a $t' \in [0,\ 1]$ such that $\frac{\beta-1}{\beta} < t' < t_R^*$. Then, we have:

$$x_M^*(t_R^*) = \frac{\beta^2(1-t_R^*)^2}{4} < x_M(t') = \frac{\beta^2(1-t')^2}{4} < \frac{1}{4} = x_L^*$$

So, MAG receive a higher after-tax income under the redistribute tax rate t' than what they receive under the redistributive tax rate t_R^*, while their after-tax income is still lower than that of LAG. This contradicts that t_R^* is the Rawlsian redistributive tax rate that maximizes the lower after-tax income of either MAG or LAG. Hence, we must have $t_R^* \leq \frac{\beta-1}{\beta}$. Note that if $t \leq \frac{\beta-1}{\beta}$, then $x_M^* = \frac{\beta^2(1-t)^2}{4} \geq \frac{1}{4} = x_L^*$. So, if $t \leq \frac{\beta-1}{\beta}$, then the members of LAG will have a lower after-tax income than members of MAG. By the Corollary of Proposition 1, LAG's after-tax income is fixed at $x_L^* = \frac{1}{4}$ independent of the redistributive tax rate. Hence, any redistributive tax rate $t_R^* \leq \frac{\beta-1}{\beta}$ will generate an after-tax income of $x_L^* = \frac{1}{4}$ for LAG and will maximize the after-tax income of LAG, whose after-tax income is lower than that of MAG, establishing the claim. ∎

Corollary of Proposition 2: Under the utilitarian redistributive tax rate $t_U^* = \frac{\beta-1}{2\beta}$, the total well-being achieved by MAG and LAG are:

- $TW_M\left(t_U^* = \frac{\beta-1}{2\beta}\right) = \frac{\beta+1}{8} + T$
- $TW_L\left(t_U^* = \frac{\beta-1}{2\beta}\right) = \frac{\beta^2+3}{16} + T.$

Proof:
Plugging in $t = \frac{\beta-1}{2\beta}$ to

- $TW_M(t) = \sqrt{\frac{\beta^2(1-t)^2}{4}} + T - \frac{\beta(1-t)}{4}$
- $TW_L(t) = \sqrt{\frac{1}{4}} + T - \frac{1-\beta^2 t(1-t)}{4}$

gives us the result. ∎

Corollary of Proposition 3: Under the maximum Rawlsian redistributive tax rate $t_R^* = \frac{\beta-1}{\beta}$, the total well-being achieved by MAG and LAG are:

- $TW_M\left(\frac{t_R^* = \beta-1}{\beta}\right) = \frac{1}{4} + T$
- $TW_L\left(\frac{t_R^* = \beta-1}{\beta}\right) = \frac{\beta}{4} + T.$

Proof:
Plugging in $t = \frac{\beta-1}{\beta}$ to

- $TW_M(t) = \sqrt{\frac{\beta^2(1-t)^2}{4}} + T - \frac{\beta(1-t)}{4}$
- $TW_L(t) = \sqrt{\frac{1}{4}} + T - \frac{1-\beta^2 t(1-t)}{4}$

gives us the result. ∎

Proposition 4: Suppose $1 < \beta < 3$. Then,

(1) $t_U^* < t_R^*$
(2) $TW_M(t_U^*) > TW_M(t_R^*)$
(3) $TW_L(t_U^*) < TW_L(t_R^*)$.

Now, suppose $\beta \geq 3$. Then,

(4) $t_U^* < t_R^*$
(5) $TW_M(t_U^*) > TW_M(t_R^*)$
(6) $TW_L(t_U^*) \geq TW_L(t_R^*)$ [and $TW_L(t_U^*) > TW_L(t_R^*)$ when $\beta > 3$.]

Proof:
First, note:

$$t_U^* = \frac{\beta - 1}{2\beta} < \frac{\beta - 1}{\beta} = t_R^*$$

for all $\beta > 1$. This establishes (1) and (4).
By Corollaries of Propositions 2 and 3, we have:

$$TW_M\left(t_U^* = \frac{\beta - 1}{2\beta}\right) - TW_M\left(t_R^* = \frac{\beta - 1}{\beta}\right) = \left[\frac{\beta + 1}{8} + T\right] - \left[\frac{1}{4} + T\right] = \frac{\beta - 1}{8} > 0$$

for all $\beta > 1$. This establishes (2) and (5).
Similarly,

$$TW_L\left(t_U^* = \frac{\beta - 1}{2\beta}\right) - TW_L\left(t_R^* = \frac{\beta - 1}{\beta}\right)$$
$$= \left[\frac{\beta^2 + 3}{16} + T\right] - \left[\frac{\beta}{4} + T\right] = \frac{\beta^2 - 4\beta + 3}{16} = \frac{(\beta - 2)^2 - 1}{16} < 0$$
$$\Leftrightarrow (\beta - 2)^2 < 1 \Leftrightarrow \ -1 < \beta - 2 < 1 \ \Leftrightarrow 1 < \beta < 3.$$

Hence, $TW_L\left(t_U^* = \frac{\beta-1}{2\beta}\right) < TW_L\left(t_R^* = \frac{\beta-1}{\beta}\right)$ if and only if $1 < \beta < 3$, which establishes (3), and $TW_L\left(t_U^* = \frac{\beta-1}{2\beta}\right) \geq TW_L\left(t_R^* = \frac{\beta-1}{\beta}\right)$ [resp. $TW_L\left(t_U^* = \frac{\beta-1}{2\beta}\right) > TW_L\left(t_R^* = \frac{\beta-1}{\beta}\right)$] if and only if $\beta \leq\ 1$ or $\beta \geq 3$ [resp. $\beta <\ 1$ or $\beta > 3$], which establishes (6) as desired. ∎

Hun Chung is an Associate Professor (with Tenure) at the Faculty of Political Science and Economics at Waseda University, Tokyo, Japan. He received his first PhD in Philosophy at Cornell University and his second PhD in Political Science at the University of Rochester. His main areas of research interests lie in PPE (Philosophy, Politics & Economics) and the intersection of Political Philosophy and Formal Theory (Game/Social Choice Theory.) URL: http://hunchung.com.

Cite this article: Chung H (2023). When utilitarianism dominates justice as fairness: an economic defence of utilitarianism from the original position. *Economics & Philosophy* **39**, 308–333. https://doi.org/10.1017/S0266267122000098