John Rawls's Difference Principle and The Strains of Commitment: A Diagrammatic Exposition

Greg Hill

Available at: https://works.bepress.com/greg_hill/3/
The Difference Principle and the Strains of Commitment: A Diagrammatic Exposition

I. Introduction

II. The Difference Principle and the Strains of Commitment

III. Utilitarianism

IV. Libertarianism

V. The Psychology of Reciprocity

VI. Conclusion
I. INTRODUCTION
Teaching philosophy to students majoring in physics, chemistry, computer science, or even economics poses unique challenges. Difficulties can arise when trying to convey philosophical ideas that bear little relation to the work these students are doing in their other courses. There are, however, some concepts in our discipline that can be conveyed in ways that draw on the skills acquired by students who are comfortable employing the methods of math and science, and, with luck, these young scientists may find themselves in a position to make a distinctive contribution to their fellow students (and perhaps to their teachers as well). One such concept is John Rawls’s notion of the “strains of commitment,” which plays an important, if little analyzed, role in Rawls’s critique of Utilitarianism and in his argument for the Difference Principle. In this essay, I present Rawls’s argument through a series of diagrams displaying the relationship between the distribution of income, on the one hand, and the strains of commitment among various groups of citizens under alternative conceptions of justice, on the other. In addition, I show how “strains-of-commitment-type” considerations can be brought to bear in evaluating Utilitarianism and Libertarianism using this same diagrammatic device, an exercise that brings into clearer view some important aspects of “justice as fairness” and its principal rivals.

II. THE DIFFERENCE PRINCIPLE AND THE STRAINS OF COMMITMENT
Rawls’s theory of justice is built around a thought experiment. We are asked to imagine that parties in an “original position” are to choose the principles that will govern the fundamental terms of social cooperation.¹ In light of the finality of the decision they are to make, Rawls stipulates that the parties must choose principles they will be able to willingly affirm no matter what their place in the distribution of social and natural
endowments. Put the other way around, the parties must reject principles that could give rise to obligations they could not discharge without rancor or resentment, that is to say, principles that would create excessive “strains of commitment.” One of Rawls’s main arguments against Utilitarianism is that those citizens whose interests are sacrificed in order to maximize the net balance of benefits would find it difficult to willingly embrace policies that served this objective. For example, while a majority of citizens might prefer that certain religions be suppressed, the parties in the original position would not adopt a principle of selective suppression because any one of them might turn out to be a member of an unpopular religious sect and could not willingly acquiesce in the suppression of their deeply-held beliefs. In a similar fashion, the parties would insist upon equality of opportunity in employment because, even if discrimination in employment turned out to be economically efficient, citizens who were discriminated against could not willingly accept their unequal treatment even if it meant greater incomes for the majority of citizens.

In light of these and other considerations, Rawls argues that an equal right to the basic liberties and fair equality of opportunity must be (in this order) the fundamental principles of a just state. In thinking through the principles that should guide the economic organization of a just state, Rawls again rejects Utilitarianism because the application of its “greatest good for the greatest number” criterion could leave some citizens impoverished if their adverse condition were the by-product of economic arrangements that produced greater benefits for the vast majority of citizens. Instead, Rawls contends that the parties would select the Difference Principle, which sanctions inequalities of wealth and income only insofar as they work to “the long-run advantage of the least well off.” According to Rawls, the Difference Principle would produce much
weaker strains of commitment than Utilitarianism because the terms of social cooperation sanctioned by this principle are consistent with the “psychology of reciprocity,” which Rawls regards as a powerful motive deeply rooted in human nature.³

More specifically, the main argument Rawls develops in favor of the Difference Principle runs as follows. The parties in the original position, who do not know their place in the distribution of social and natural advantages, and who represent citizens “as free and equal,” begin their deliberations about the basic terms of social cooperation by assuming that the benefits of social cooperation are to be equally distributed. Starting from this baseline, the parties then consider the gains that could be achieved for everyone if unequal shares were offered as an incentive for especially productive effort, innovation, risk taking, and the like. The parties continue to reason in this vein, examining the consequences of marginal increases in inequality until they reach a point where further increases in inequality no longer work to everyone’s advantage. Inequality in excess of this threshold will only raise the income of more productive citizens at the expense of less productive citizens. Since the parties must be able to meet their public responsibilities no matter what their place in the distribution of social and natural advantages, they focus their attention on the income of “the least well off,” and therefore select the degree of inequality that maximizes the income of this group.

This line of reasoning is depicted in figure one, which displays the income of a representative citizen among the least advantaged group (on the horizontal axis) in relation to the degree of income inequality in the political community (on the vertical axis). (In all that follows, I shall always be referring to either the representative citizen, or to the average income, of some group). According to Rawls’s account, we are to envision the parties beginning at the origin, where income shares are equal, and then
gradually moving up the “income-inequality curve” in the direction of the arrows until they reach the point at which the income of the least well off reaches its maximum, where we find the Difference Principle.

Figure 1
Income Inequality and the Income of the Least Well Off

In figure one, the points, $\text{DP}^{\text{ineq}}$ and $\text{DP}^{\text{inc}}$, designate, respectively, the degree of inequality and the income of the least well off, under the Difference Principle. Any further increase in inequality above $\text{DP}^{\text{ineq}}$ reduces the income of the least well off. Since Rawls assumes that everyone gains in the move from equal shares to the degree of inequality marked by the Difference Principle, and that the income of the least advantaged declines with any further increase in inequality, this version of Rawls’s argument for the Difference Principle has been called the Pareto interpretation.⁴
III. Utilitarianism and the Strains of Commitment

Although the income of the least advantaged declines as inequality increases beyond the threshold marked by the Difference Principle, Rawls assumes that the income of more favored citizens (not shown in figure one) continues to rise as the degree of inequality increases. To illustrate this relationship, figure two displays the average income of the least well off group (LWO) and the average income of all other citizens, the non-least well off (NLWO), under varying degrees of inequality. For purposes of comparison, I’ve shown two principles of distribution, the Difference Principle and Utilitarianism.

Figure 2

Income of the Least Well Off (LWO) and All Others (NLWO) under Alternative Principles

In figure two, the income of the least well off (LWO) and the average income of all other citizens (NLWO) are both represented on the horizontal axis. The points, $L^U$ and $L^D$, designate the income of the least well off (LWO) under Utilitarianism and the
Difference Principle, respectively. Similarly, the points $A^U$ and $A^D$ designate the average income of all other citizens (NLWO) under Utilitarianism and the Difference Principle, respectively. Note that in moving up the income inequality curve from the Difference Principle to Utilitarianism, the income of the least well off declines from $L^D$ to $L^U$, whereas the average income of the remaining population rises from $A^D$ to $A^U$, an amount that exceeds the lost income of the LWO by a considerable amount. There are many reasons why this is likely to be the case, but the one most relevant to our concerns is that the gains made possible by incentives exceeding those sanctioned by the Difference Principle would increase the income of many citizens while decreasing the income of only a few, a “trade off” sanctioned by the Utilitarian objective of maximizing the greatest good for the greatest number.\(^5\)

Looking at figure two, some students may ask why the parties in the original position would choose the Difference Principle rather than Utilitarianism in light of the fact that, given certain additional, but plausible, premises, Utilitarianism promises greater expected income (and even greater expected utility) than the Difference Principle.\(^6\) One of Rawls’s arguments against this conclusion is that Utilitarianism and, more specifically, the aim of maximizing expected (i.e., probability-weighted) utility, is not an appropriate decision rule in the circumstances of the original position. In the first place, Rawls emphasizes that the parties do not have sufficient information to attach probabilities to possible outcomes; hence they cannot estimate expected values.\(^7\) Looking back at figure two, and bearing in mind Rawls’s assumption that the parties in the original position don’t know their place in the distribution of natural and social endowments, we can interpret this requirement by stipulating that the parties don’t know whether they are on the LWO or the NLWO curve, i.e., whether they belong to the least well off group or not.
Secondly, Rawls insists that the parties, who are conceived as moral persons, have very little interest in income above some minimal standard, so that the marginal value of income diminishes rapidly once this threshold is reached. 

In addition to the foregoing arguments, Rawls also makes a case for the superiority of the Difference Principle over Utilitarianism by introducing considerations that revolve around the strains of commitment. Since the parties choosing principles of justice know they “cannot enter into agreements that may have consequences they cannot accept,” they must “consider the strains of commitment,” weighing “with care whether they will be able to stick by their commitment in all circumstances.” Although the parties “have only a general knowledge of human psychology to go on,” Rawls insists that “this information is enough to tell which conception of justice involves the greater stress.” To begin with, Rawls argues that this knowledge is sufficient to reject Utilitarianism because the principle of utility may require that the interests of some citizens be sacrificed in order to advance the interests of others. Consequently, a social system organized around the principle of utility “will not be stable unless those who must make sacrifices strongly identify with interests broader than their own.” Although Rawls allows that sympathy and benevolence can be called upon to prompt sacrifices from time to time, he contends that these motives are not strong enough to prompt willing sacrifices from the same groups on an ongoing basis.

By contrast, Rawls argues that the Difference Principle would not give rise to such intense strains of commitment. His argument rests on the following three claims: 1) any degree of inequality greater than that sanctioned by the Difference Principle would allow the more favored to gain at the expense of the least favored, thereby coming into conflict with the “psychology of reciprocity,” which Rawls regards as a deep and, most
likely, permanent, part of human nature; 2) since the least favored do as well as possible under the Difference Principle, they have no reason for complaint; and 3) since those more favored by nature do even better than the least favored, they, too, have no reason for complaint. These ideas are illustrated in figure three, which displays the strains of commitment among the least well off (LWO) and among all other citizens (NLWO). Although figure three resembles figure two, it’s important to recognize that the intensity of the strains of commitment (SOC) replaces income on the horizontal axis, so the SOC become weaker as we move rightward along this axis (whereas, in figure two, income increases as we move rightward along the horizontal axis.)

Figure 3

Strains of Commitment under Alternative Principles

![Figure 3](image-url)
Two aspects of figure three deserve emphasis. First, notice that the strains of commitment are greater for the least well off (LWO) than for the more favored (NLWO) at every point north of the origin, that is, under all distributions except the baseline of equal shares. This is because the strains of commitment decrease as a citizen’s absolute income increases, and the least well off (LWO), by definition, always have less income than others. Second, it’s important to stress that the strains of commitment among the least well off (LWO) increase at any point above that marked by the Difference Principle for two reasons, first because their income decreases when inequality exceeds this threshold, and second because the more favored gain at their expense, which, by coming into conflict with the psychology of reciprocity, gives rise to resentment and perhaps even a diminution of self-respect.

Since Rawls’s strains of commitment argument is directed mainly against Utilitarianism, it may be fruitful to ask how a Utilitarian might respond to this argument. One natural reply would be to deny the relevance of the strains of commitment as an independent consideration in the design of society’s most fundamental institutions. For an unabashed Utilitarian, the strains of commitment command attention only insofar as this stress involves a cost, or disutility, which must be included in calculating the net benefits of the alternatives under consideration. Having insisted on this point, the Utilitarian may go on to argue that Rawls has underestimated the strength of our capacity for sympathy, that is, our willingness to incur costs so that others may enjoy greater benefits, a capacity which, it may be added, could be further strengthened by education and other means. According to this understanding, the difference between the strains of commitment among the least well off under the Difference Principle, on the one hand, and under Utilitarianism, on the other, could be less than Rawls imagines. Specifically,
the strains of commitment curve for the least well off (LWO) in figure three would be less backward-bending where it exceeds the threshold level of inequality marked by the Difference Principle.

Although this response, assuming its premises are granted, weakens the force of Rawls’s strains of commitment argument in favor of the Difference Principle, it’s worth mentioning that the income distribution consistent with Utilitarianism, insofar as the Utilitarian takes Rawls’s notion of the strains of commitment seriously (at least as a cost that must be taken into account), would perhaps be more egalitarian than would be the case in the absence of this consideration. We can take a further step along this path if we abandon the strict Utilitarian doctrine and accept Rawls’s notion that the strains of commitment must not be excessive for any group. In this case, the parties in the original position may find the notion of “Utilitarianism with a Social Minimum” (as proposed by Jeremy Waldron) an appealing alternative to the Difference Principle.13 For the addition of a social minimum to an otherwise straightforward Utilitarianism would go a considerable distance toward addressing Rawls’s claim that Utilitarianism would create excessive strains of commitment among the least well off. In figure three above, “Utilitarianism with a Social Minimum” could be represented by a horizontal line lying between the lines representing Utilitarianism and the Difference Principle, allowing greater inequality than the Difference Principle, but less than unrestricted Utilitarianism, while generating greater average income than the Difference Principle and a higher minimum than Utilitarianism.

Nevertheless, the Rawlsian may insist, any degree of inequality greater than that sanctioned by the Difference Principle would come into conflict with the psychology of reciprocity and on this ground alone would produce unacceptable strains of commitment
among the least well off. We shall consider this argument more closely below, but first let’s turn to an entirely different view about the strains of commitment and their bearing upon the hypothetical choice to be made in the original position.

IV. Libertarianism and the Strains of Commitment

Rawls’s most famous libertarian critic, Robert Nozick, argues that under the Difference Principle the strains of commitment experienced by the most well off would, in fact, be greater than the strains experienced by the least well off because the income of the most well off under the Difference Principle is much less than it would be under a variety of alternative principles, but most especially Laissez Faire. According to Nozick, the proper baseline for comparison is not equal shares, but rather whatever distribution would result from the free exchange of justly acquired goods. Looking at figure four below for reference, the proponents of Laissez Faire would argue that (within Rawls’s strains-of-commitment framework), the correct starting point for measuring gains and losses is whatever degree of inequality would result from just acquisition and free exchange under the policy of Laissez Faire. And beginning from this starting point, it is obvious that the most favored lose at the expense of the least favored as we move down the “inequality-income curve” from Laissez Faire to the Difference Principle. In the libertarian view, therefore, the Difference Principle, by imposing heavy losses upon the most favored, would create more powerful strains of commitment among this group than among the least well off. We will examine this claim in more detail later, but first let’s give diagrammatic expression to the moral psychology at work in the libertarian argument. Figure four displays the strains of commitment for the least well off (LWO) and the most well off (MWO) groups under Laissez Faire and under the Difference Principle.
In the libertarian version of the strains of commitment depicted in figure four, the
group that experiences the greatest stress depends on which principle of distribution is
adopted. Under *Laissez Faire*, the least well off (LWO) experience the greatest strains,
marked by \( L^L \), because their income is significantly less than it would be under the
Difference Principle, that is, at \( L^D \). Contrariwise, under the Difference Principle, the
most well off (MWO) face the greatest strains because their income, at \( M^D \), is
significantly less than it would be under *Laissez Faire*, that is, at \( M^L \). It’s worth
underscoring the crucial significance of the baseline against which “gains” and “losses”
are measured. For Rawls, the appropriate baseline for free and equal citizens is equal
income shares, whereas for Nozick, it is whatever distribution arises from the
combination of just appropriation and free exchange.\(^{16}\)

V. The Psychology of Reciprocity
Let’s take a closer look at Rawls’s conception of reciprocity, as it plays an important role in his argument for the Difference Principle. In figures three and four, the strains of commitment among the least well off gradually decrease as their income rises, and then gradually increase as the degree of inequality exceeds the threshold sanctioned by the Difference Principle, and their income falls. But Rawls’s characterization of the psychology of reciprocity, in combination with his view that the more favored gain *at the expense of* the least favored when inequality exceeds the threshold prescribed by the Difference Principle, suggests that there is a *discontinuous* break in the strains of commitment among the least well off at the point where inequality exceeds this threshold. This notion is illustrated in figure five, which displays the strains of commitment among the least well off (LWO) under Laissez Faire and the Difference Principle.

![Figure 5](Image)

**Figure 5**
Strains of Commitment among the Least Well Off
Under Rawls’s Moral Psychology

---

[Figure Image Description]

- **High** and **Low** levels of Strains of Commitment
- **Degree of Inequality** ranges from Low to High
- **Laissez Faire** and **Difference Principle** stages
- **Exploitation** and **Cooperation** regions
- **Equal Shares** and **L^L** to **L^D** markers
In figure five, the strains of commitment among the least well off gradually diminish as the degree of inequality increases up to the threshold marked by the Difference Principle, that is, \( L^D \). Beyond this threshold, which marks the endpoint of mutually advantageous cooperation, there is a sharp, discontinuous increase in the strains of commitment among the least well off because any increase in inequality beyond this point means that, in Rawls’s conception, one segment of the community gains at the expense of another. And this state of affairs, in turn, comes into conflict with the psychology of reciprocity, thereby creating powerful strains of commitment among a group that not only has less income than any other group, but, from Rawls’s perspective, makes sacrifices that benefit those who are already more favored. Notice that the curve representing the strains of commitment among the least well off (LWO) is convex above the Difference Principle, which means that the strains of commitment rise even faster in this region, as if additional inequality beyond the Difference Principle “rubs salt into the wounds” of the least well off. Under the Difference Principle, by contrast, everyone benefits from the resulting scheme of cooperation, (so that the discontinuous segment of the psychology of reciprocity-driven strains of commitment does not come into play). “Since,” in Rawls’s words, “everyone’s good is affirmed, all acquire inclinations to uphold this scheme.”

This argument is not entirely satisfying. We may agree with Rawls that the psychology of reciprocity is a deep part of our nature, and that we tend to develop loyalties to people and institutions that advance our good. But Rawls gives us no reason to think that in judging whether a particular relationship or social arrangement advances our good, we naturally assume a baseline of equal shares against which we measure our gains and losses. One can easily envision relationships of reciprocity in which the
benefits accruing to each participant are not measured in relation to any recognizably egalitarian baseline. For example, consider a voluntary association where, in Rawls’s words, the members “view one another as equals, as friends and associates, joined together in a system of cooperation known to be for the advantage of all and governed by a common conception of justice.”

Although association members may fulfill their responsibilities because they enjoy advantages as members of the association, there is no reason to believe these benefits are measured, even informally, against some kind of “equal shares” baseline, rather than against, say, the advantages a person might enjoy as a member of a different association, or simply by spending more time with one’s family.

More generally, we can say that among competing conceptions of justice that specify different baselines for the purpose of assessing gains and losses, appeals to reciprocity are question-begging because they presuppose precisely what is at issue, that is, the baseline against which costs and benefits are to be measured. In Rawls’s argument, everyone can be said to benefit from the adoption of the Difference Principle, provided 1) we start from the equal shares baseline, 2) move up the income-inequality curve by small increments, and 3) never cross the inequality threshold marked by the Difference Principle. For Nozick, on the other hand, the Difference Principle is not a principle of reciprocity because it transfers income from the more productive to the less productive, thereby leaving the first group worse off when measured against a baseline in which privately owned assets have been justly acquired. *Laissez faire*, on this view, gives reign to reciprocity because exchanges are voluntary and therefore typically work to the advantage of contracting parties. This debate is not really about which institutional arrangement is consistent with our propensity to reciprocate, but about the appropriate baseline from which to assess gains and losses. The real work in Rawls’s argument for
the Difference Principle is being done by the equal shares baseline, the assumption of rationality, and the considerations Rawls brings to bear in their defense, and not by the “psychology of reciprocity,” which, in Rawls’s account, is part of our natural constitution, but which, we have argued, does not necessarily embody an egalitarian baseline. This point helps us to appreciate Rawls’s remark that “it is hard to imagine” any conception other than the Difference Principle that would satisfy the “reciprocity condition.” Yes, of course, provided we start from a baseline of equal shares.²¹

VI. CONCLUSION

A diagrammatic presentation of Rawls’s arguments concerning the strains of commitment serves two purposes. In the first place, it is an alternative way of illustrating some of the basic premises and arguments of Rawls’s conception of justice as fairness. In fact, some interesting nuances in Rawls’s theory can be brought out simply by focusing upon the shape of the curves displayed above. Secondly, these diagrams may be especially helpful to students who are unfamiliar with philosophical discourse, but comfortable with the graphical representation of concepts. Although the deepest questions about the nature of justice do not revolve around diagrams, these figures may, hopefully, facilitate discussion about the first principles of political philosophy.

2 See *A Theory of Justice*, pp. 176-183

3 See ibid., pp. 102 ff.

4 The claim that everyone benefits from inequality up to the point marked by the Difference Principle requires a further premise, which Rawls calls “the chain connection.” Rawls also refers to it as a Kantian justification of the Difference Principle because no person uses another as a mean, that is to say, no one gains at the expense of someone else. *A Theory of Justice*, pp. 81-83, 251-257.

5 Strictly speaking, income must be translated into utility before balancing gains and losses. In addition, the diminishing utility of income imposes an egalitarian constraint on distributional alternatives.

6 Such assumptions would include the premise that the probability of being among the least favored is low in comparison with the probability of not belonging to this group.

7 Moreover, Rawls argues, there are strong reasons against even the “default” position wherein all outcomes are assumed to be equally likely.

8 To be sure, this premise does not fit comfortably with the theory of incentives that supports the operation of the Difference Principle.


10 Ibid.

11 Ibid., pp. 177-178.


15 It may be noted that the appropriate starting point, *laissez faire*, is also the correct ending point from the libertarian vantage point.

16 It’s harder to say what Utilitarianism’s comparable baseline would be, for while Utilitarians claim to treat everyone’s interests equally, I cannot find a line of reasoning that leads from here to an equal shares baseline, nor to an “equal utility” baseline, nor, finally, to anything resembling the libertarian baseline.
To avoid misunderstanding, I should emphasize that the issue in this instance is not whether equal shares is the morally appropriate baseline, but rather whether this baseline, and the gains and losses measured in relation to it, constitute the reference point in terms of which the psychology of reciprocity is activated or not.

To summarize, appeals to the psychology of reciprocity only carry weight if the competing principles under consideration specify the same baseline for measuring gains and losses. What is left of Rawls’s strains-of-commitment argument in favor the Difference Principle if it turns out that the natural psychology of reciprocity cannot be plausibly harnessed to Rawls’s equal shares baseline? Rawls could still argue that the strains of commitment decrease as a citizen’s absolute income increases and therefore the Difference Principle, which maximizes the income of the least well off, necessarily minimizes the strains of commitment in comparison with any alternative principle. Although this fallback position is straightforward, it does not carry the same force as Rawls’s original argument, which includes the additional strain that would arise if, under greater inequality than that sanctioned by the DP, the more well off gain at the expense of the least well off.


References


