**Justice and the Structure of Reciprocity: How Empirical Results Can Inform Normative Theory**

**1. Introduction**

Over the past two decades a rich empirical literature has developed, reflecting work in economics, psychology, neurobiology, evolutionary biology and other disciplines, concerning human cooperation, and the distribution of the benefits and burdens that it generates. These issues are also a focus of a great deal of normative theorizing, much of it falling within the subject matter of theories of justice. It is a natural thought that the empirical literature must have some bearing on the normative theories, but it is no easy matter to spell out what these connections might be. This paper explores this question, by focusing on empirical literature relevant to one particular notion that plays a role in many normative theories of justice —the notion of reciprocity. I intend the paper as a single case study; there are many other possible ways that empirical results might be relevant to normative theory. I hope that my discussion will encourage investigation of these other connections.

The organization of this paper is as follows. In Sections II -IV, I describe some general ways in which empirical results might be relevant to normative theorizing, focusing mainly on what they tell us about the motivations underlying cooperative behavior and how these constrain which cooperative schemes are likely or not to be stable or implementable. I also describe some general contours of the notion of reciprocity suggested by the empirical studies discussed in subsequent sections. Section V describes some results from studies of prisoner’s dilemmas and public goods games that are relevant to an understanding of reciprocal behavior and explores their significance for moral theory. These include the existence of considerable heterogeneity in human motivation and behavior and the implications of this (explored in Section VI) for the project of constructing ideal theories of distributive justice that assume “full compliance”. Section VII then explores, again from an empirical point of view, the issue of the scope of reciprocity as a moral consideration—whether, so to speak, reciprocity is all of justice or just a part of it. Section VIII discusses the important role played by people’s motives and intentions in reciprocal interactions and the implications of this for the understanding of reciprocity. Section IX discusses the claim that reciprocal interactions require mutual gain. I argue that this requirement is too strong and that this in turn has important consequences for the role played by reciprocity in the justification of social welfare programs, which are discussed in Section X. Finally Section XI draws some more general conclusions about the normative significance of reciprocity and the need for rules and institutions if reciprocators are to successfully cooperate.

**II. Barry and Gibbard on Theories of Justice**

It is a common thought that normative claims cannot be deduced from descriptive claims alone, but also a common thought that the latter can sometimes constrain the former. Assuming this to be so, an obvious strategy is to look for plausible principles connecting the descriptive and normative—for example, connecting principles that tell us that satisfaction of some descriptive condition is necessary or sufficient for the holding of a normative conclusion or supports it in some other way. Following this strategy, I begin by recalling a disagreement between Brian Barry and Alan Gibbard (among others) about the relationship between different theories of justice[[1]](#footnote--1).

In setting up this issue, Barry claimed that

… a theory of justice may be characterised by its answers to three questions. First, what is the motive (are the motives) for behaving justly? Secondly, what is the criterion (are the criteria) for a just set of rules? And thirdly, how are the answers to the first two questions connected? We want to know exactly how somebody with the stipulated motive(s) for behaving justly would be led to comply with rules that are just according to the stipulated criterion or criteria. A theory of justice that cannot answer the third question satisfactorily fails on the ground of internal inconsistency. (1989, p. 46)

Although Barry does not explicitly say this, it presumably matters, within this framework, whether, as a empirical matter, a significant number of people actually have the motive for behaving justly that is postulated by the theory of interest. And of course it also matters whether their possession of that motive actually leads them to behave in ways that the theory recommends. A theory that relies on a motive that no one or very few people actually have or which most people are unlikely to act on would presumably not do a good job of answering Barry’s third question. Formulated very roughly, the connecting principle implicit in Barry’s remarks seems to be something like this: A necessary condition on the normative acceptability of a theory of justice is that a significant number of people either have motivations that will lead them to act in accord with the recommendations of the theory or (more weakly) that a significant number of people could come to posses, under the right circumstances, motivations that would lead them to so act. Since , to a substantial extent, it is an empirical matter what motivations people possess or might come to possess and behavior these might produce, this is a principle that connects the empirical and the normative. My focus in what follows will not be on defending this (or similar) principles, but rather in exploring how, if we are willing to assume something like them, we can bring empirical considerations to bear on normative theorizing.

Barry focuses on two different theories of justice, which he calls Justice as Mutual Advantage (JMA) and Justice as Impartiality (JI). I will understand JMA as the claim that the principles or rules of justice are those that would be adopted by people who are motivated only to advance their own self-interest and who regard it as “rational” or normatively prescriptive to always act in accord this motive[[2]](#footnote-0). According to JI as Barry understands it, the motive for behaving justly is the desire “ to act in accordance with principles that could not reasonably be rejected by people seeking an agreement with others under conditions free from morally irrelevant bargaining advantages and disadvantages”. As Barry interprets this conception, it generates, in some circumstances, a requirement to aid that is not conditional on the choices and behavior of those aided—in other words, it draws on something like a motive of unconditional altruism. (By “unconditional altruism” I mean simply that the motive prompts action directed at the good of others and is not conditional on their behavior. Of course Barry has a more specific conception of what JI requires.) According to Barry, this motive, like the desire to advance one’s own self- interes is a basic, primitive human motive. Barry thinks of JMA and JI as the only two coherent alternative ways of thinking about justice and sees Rawls’ theory as a not entirely successful attempt to combine elements from both approaches in a kind of hybrid theory that “hover[s] uneasily between impartiality and mutual advantage”.

Commenting on this suggestion in his review of Barry’s book, Alan Gibbard (1991) writes:

Perhaps [Barry] is right in this, but Rawls long ago seemed to have his eye on a third perch: one he called Justice as Reciprocity.

Gibbard characterizes justice as reciprocity as follows:

If I return favor for favor, I may be doing so in pursuit of my own advantage, as a means to keep the favors rolling. My motivation might, though, be more intrinsically reciprocal: I might be decent to him because he has been decent to me. I might prefer treating another well who has treated me well, even if he has no further power to affect me. We tip for good service in strange restaurants.

He adds:

Is Justice as Fair Reciprocity a distinct alternative to Barry's two? Is it different in any way from Justice as Mutual Advantage and Justice as Impartiality? The case that it is, in a nutshell, is this: On the one hand, it is distinct from Justice as Mutual Advantage because it draws on non- egoistic motives. On the other, it is distinct from Justice as Impartiality because it says that a person cannot reasonably be asked to support a social order unless he gains from it.

Rawls has repeatedly described his own theory of justice as based on reciprocity. Moreover, he explicitly endorses Gibbard’s claim that his theory is distinct both from theories of justice based on self-interest and from theories that base justice on other-regarding motives distinct from reciprocity (Rawls, 1993, p. 17, n. 18). Nor is this an idiosyncratic feature of Rawls’ analysis—many political scientists, economists, and historians contend that state schemes in which there is extensive redistribution of income and other resources (“welfare states”) of the sort favored by Rawls are based on reciprocity.

How might we decide whether Gibbard (and Rawls) or Barry is right about the existence of a reciprocity-based conception of justice that is distinct both from justice as mutual advantage and justice as impartiality, when these are understood as described above? And what would such a conception look like? I suggest that one possible strategy would be to make use of Barry’s suggestion that different theories of justice specify or make use of different motives for acting justly. In particular we might ask the following question: is there evidence that motivations having to do with reciprocity, where these are understood as distinct both from motivations based on long term self -interest and from unconditional altruistic or other-regarding motivations not based on reciprocity, play an important role in sustaining cooperation and in the allocation of benefits and costs resulting from cooperation? And if such motives exist, what sort of conception of fairness and justice do they support? I will suggest below that when the notion of reciprocity is properly understood (along broadly Gibbardian lines, but with an important qualification), the answer to the first question is “yes”: There is evidence that the desire to reciprocate is a very strong and robust motive (often but not always stronger and more robust than other sorts of other –regarding motives) in explaining a great deal of human cooperation. Moreover, this motive goes along with a distinctive conception of what one does and does not owe to others in contexts in which cooperation occurs. In other words, my suggestion is that empirical results about human behavior and motivation in situations in which cooperation is a possibility can bear on the question of whether there is a distinctive motive to reciprocate of the sort described by Gibbard. In particular, the empirical hypothesis that people are motivated by considerations of reciprocity makes distinctive predictions about how people will behave in many situations—predictions that are different from the predictions associated with the hypothesis that people are self-interested and the hypothesis that they are motivated by other –regarding considerations that are distinct from reciprocity. These reciprocity-based predictions are often borne out empirically. Of course the existence of a motive to reciprocate does not by itself establish that we are morally right to be guided by it. However, if we follow Barry’s suggestion about the connection between criteria for justice and motives for behaving justice, it is arguable that *if* there is a distinctive conception of justice based on reciprocity, then there should be a corresponding motive of roughly the sort described by Gibbard—in other words, the existence of the motive is arguably (at least) a necessary condition for the viability of justice as reciprocity as a distinct conception of justice. Moreover, as I will also suggest below, to the extent that a conception of justice is based on reciprocity, it is plausible that empirical features of this motive will help to shape normative features of this conception. For example, as an empirical matter, the motivation to reciprocate or not is heavily influenced by the potential reciprocator’s perception of the intentions of the person who is a potential object of reciprocation; reciprocators respond quite differently to those who have intentionally aided them, as compared with those who have produced a similar benefit for them but not intentionally. This points to normative conception of reciprocity according to which such intentions are important. In addition, as we shall see, the existence of large numbers of people who are apparently heavily influenced by reciprocity-based motives matters in other ways as well: it has implications for the stability of alternative principles of justice and more generally for the design of rules and institutions governing cooperative behavior. Indeed, the whole enterprise of designing a theory of justice looks rather different once one begins to pay attention to reciprocity as a motive.

**III. The Structure of Reciprocity: Preliminary Considerations**

Before proceeding it will be useful to flesh out the notion of reciprocity in a preliminary way[[3]](#footnote-1). Like Gibbard, I take the core idea to involve the notion of *conditional* cooperation or provision of benefits, or in the case of negative reciprocity (see below) the conditional imposition of costs. In the most familiar or paradigmatic case, reciprocators cooperate with or aid others who they believe have intentionally cooperated with or aided them in the past and do so (at least in part) for that very reason; in other words they cooperate because they have been so aided in the past, even if cooperation is not the optimal strategy for advancing the reciprocator’s self-interest. It is important to distinguish the motive to reciprocate, thus understood, from self-interest. As an illustration discussed in more detail in Section III, consider a reciprocator who moves second in a “sequential” version of an interaction that in terms of material benefits has the structure of a one-shot prisoner’s dilemma -- an interaction that is sequential in the sense that one player moves first, choosing cooperate or defect and the second mover is then informed of the first player’s choice and then herself chooses whether to cooperate or defect. A reciprocator who learns that the other player has already chosen cooperatively, will herself play cooperatively to reciprocate the first player’s cooperation; by contrast, a self-interested second mover will defect. On the other hand, a reciprocator will think of herself as having no reason to play cooperatively if the first player chooses to defect. An unconditional altruist, in contrast, may think he has a reason to respond to defection with cooperation, if the aggregate pay-off from this pair of choices is (by some relevant measure) greater than the pay-off from (defect, defect). These differences illustrate how reciprocity as a motive is distinguished both from self-interest (conceived as a concern only with material benefits to self) and from unconditional other-regarding motives.

Many readers will be familiar with the idea that in a repeated prisoner’s dilemma, a “reciprocal” strategy like “tit for tat” (begin by cooperating and then do whatever your opponent did on the previous round) can be sustained among self-interested players—the mutual choice of this strategy is a Nash equilibrium of the repeated game, given self-interested players. (A pair of choices in a two person game is a Nash equilibrium if each player’s choice is a “best response” to the other players’ choice; that is, given one player’s choice, the other player cannot do better by choosing a different strategy.) This may encourage the thought that an appeal to self-interested motives in repeated interactions is sufficient to explain whatever reciprocal behavior exists, and that there is no need to appeal to any motive to reciprocate that is distinct from self-interest. For a variety of reasons this is a mistake. For one thing, self-interest cannot explain why people cooperate or behave reciprocally (as in fact many do) in one-shot games such as the sequential prisoner’s dilemma. Moreover, even in a repeated prisoner’s dilemma, the mutual choice of tit for tat is just one possible Nash equilibrium; there are many other possible Nash equilbria (again assuming purely self-interested players) which are much less cooperative and reciprocal, such as mutual choice of “always defect”[[4]](#footnote-2). Appeals to self-interest thus cannot explain why, in such repeated games, reciprocal rather non-reciprocal strategies emerge (to the extent that they do.) Finally, as discussed below, there are many examples of repeated interactions (such as repeated public goods games) in which as an empirical matter, substantial numbers of people behave reciprocally even though such behavior is not even a Nash equilibrium of the repeated game for selfish players. These considerations suggest that while self-interest no doubt plays a role in some situations in explaining behavior that is extensionally equivalent to the behavior that would emerge from purely reciprocal motives, and while self-interest may operate along side of and reinforce reciprocal motives, many examples of behavior that involve returning benefit for benefit cannot be explained purely in terms of self-interest; distinct motives to reciprocate are also involved[[5]](#footnote-3).

Reciprocation involves responding positively to benefits provided by others, but this does not exhaust the notion of reciprocation. There are good empirical reasons (grounded in actual motivation and behavior) to expand the notion of reciprocity to cover decisions to aid or cooperate with others that are not made in response to past cooperative acts but are rather motivated by the desire to cooperate with or provide aid to another in the *expectation* or *anticipation* that such aid will be reciprocated in some way in the future. As an empirical matter, people who respond to cooperation and aid by reciprocating them are also often willing to extend aid to others, before they have reciprocated, in the expectation or hope of in this way *inducing* reciprocation by others. Or at least this is the default behavior of many (but not all) reciprocators, in the sense that they are willing to do this in the absence of specific evidence indicating that that if they perform cooperative acts, these will not be reciprocated. Again as an empirical matter, reciprocators tend to expect a substantial number of others to be reciprocators, and to treat them accordingly, just as more purely self-interested types expect most others to be self-interested, and treat them accordingly.

As an illustration of these ideas, consider the following two games (both variants on “trust games”[[6]](#footnote-4), which are taken from McCabe, Rigdon, and Smith (2003).

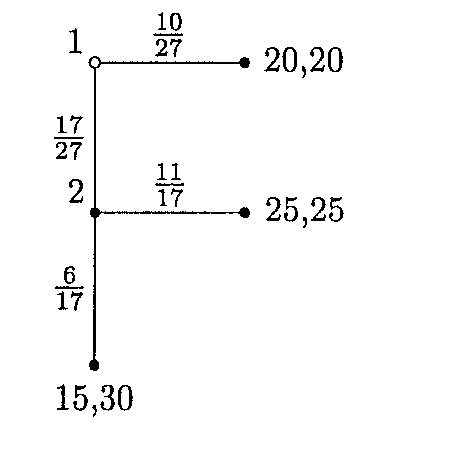
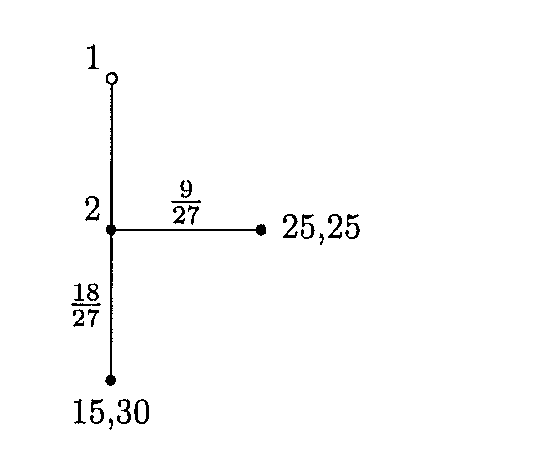
Figure 1 

Figure 2 

In figure 1, player 1 moves first, at choice node 1. Player 1 can choose to move across, ending the game and ensuring a payoff of (20, 20) for herself and player two. Alternatively, player one can move down, in which case player two has a choice of moving across, with payoffs (25, 25, first number is player one’s payoff ) or moving down with pay-offs (15, 30). Player 1’s choice to move down is thus a “trusting” move in the sense that she foregoes a certain pay-off of 20 in order to give player two an opportunity to choose an alternative that is better (25, 25) for both, but also trusting that player two will not choose down, producing an outcome that is still better for player 2 (30 vs 25) but worse for player 1. Put slightly differently, in choosing down, player 1 confers a benefit on player 2 which she expects or hopes will be reciprocated by player 2’s choosing in a way that benefits player 1. To the extent that player 1 chooses in this way, this is an example of the use of the conferral of benefit to induce reciprocation. A natural hypothesis about interactions of this type, described by Philip Pettit (1995), is to suppose that under the right conditions the trust communicated by Player 1’s choice of down can help to induce player 2 to behave reciprocally and that when Player 1 moves down, she acts with this intention or expectation. That is, Player 1 hopes or expects that as a result of her moving down (and the intention to trust communicated by this), player 2 will be more likely to reciprocate by moving across than player 2 would be in the absence of this move and intention. Of course, if player 2 is purely selfish, she will choose down. If player 1 anticipates this move, he will choose across.

Now compare this with the game in figure 2, which is just like the game in figure 1, except that the outside option has been removed—now player 1’s only choice is to move down. Because of this, player 1’s move down cannot communicate to player 2 an intention to trust or help player 2. Supposing that something like Pettit’s analysis is correct, this removes *one* of player 2’s motives to move across: the motive of reciprocating the trust shown by player 1. As a result, one would expect that to the extent that player 2 is the sort of reciprocator who responds to the intentions with which player 1 acts, player 2 would be more likely to choose down in the second version of the game (without an outside option) than in the first version. As the numbers attached to the game trees (which give the proportions of players a choosing each alternative at each node) this is indeed what is found— the difference in choices of down vs. across by player 2 across the two versions of the game is significant at the 0.01 level.

There are a number of features of this result that are interesting but for the moment I want to focus on just the following: the results very strongly suggest that some significant number of players in the second mover’s position (in the first version of the game) behave like reciprocators and are apparently influenced by a desire to reciprocate. At least in the context of this game, those who have been intentionally aided by others are more likely to reciprocate to those others than those who have not been so aided. Moreover, first movers in the first version of the game (accurately) anticipate this response and this expectation of reciprocity plays a role in motivating their choices.

I suggest that to the extent that prospective acts of cooperation (like those exhibited by players in the first movers position in Figure 1, who move down) are motivated in part by a desire to be in a relationship of reciprocation with the person aided and not merely by a desire to advance the player’s own self-interest, we may plausibly think of them as also deriving from the sorts of motives that underlie a conception of justice or fairness as reciprocity[[7]](#footnote-5). It is also worth adding in this connection that if the notion of reciprocation (or a conception of justice as reciprocity) were confined to just the idea that one should return benefit with benefit but said nothing about providing benefits to others when one has not yet received them but there is the reasonable expectation of reciprocation, it is hard to see how reciprocity-based patterns of cooperation could ever get started, since each reciprocator would hold back, not acting cooperatively before he or she benefited from some previous act of reciprocation. The desire for reciprocity, properly understood, goes beyond this.

I spoke above of reciprocators as capable of being motivated *prospectively* by a desire to be in a relationship of reciprocation with others rather than purely by self-interest. As a further illustration of what this means, consider someone who aids another in need by providing some resource for which she expects to receive some form (of what she regards as) reciprocation but where a purely self-interested use of the resource would provide a much greater return: I lend you $ 1000 to start up a small business, with the understanding that you will pay back this principal, along with very modest interest, since that is all you can afford. I might instead have invested the money at a much higher rate of interest. My motivation is to help you, and I am not just trying to do as well as I can for myself in material terms, but the relationship I want to establish with you is not one of unconditional altruism in which I just give you the money, with no expectation of repayment. I expect you to reciprocate by returning the money with interest.

In the cases discussed so far, reciprocation is a dyadic affair, involving just two actors. But in many cases it also seems very natural to extend the notion of reciprocity to cover cases involving the interaction of many actors—for example when these have the opportunity to provide some good either to themselves or others which requires cooperation. Thus, reciprocators will be willing to contribute their fair share to a public good conditional on whether others who benefit from the good and are able to contribute do so as well. Those who accept the good and fail to contribute, free riding on contributions of others, are paradigmatic non-reciprocators. On the other hand, reciprocators differ from unconditional altruists in that they will not continue to contribute to the good if others who benefit and are able to contribute fail to do so.

There are several other features of reciprocity that will occupy us in more detail below, but which will be useful to introduce at this point. First, a clarificatory remark about self-interest, which I have placed in opposition to the motive of reciprocity. Some writer contend that if people cooperate in a game that has the structure of a one-shot prisoner’s dilemma when evaluated in terms of monetary pay-offs, this in itself shows that the players regard it as in their self-interest to cooperate -- if not, why do they cooperate? In other words, cooperation is automatically taken to show that the player’s self-interest includes more than just their monetary pay-off and hence that the game in question (when evaluated in terms of the player’s utilities rather than just their monetary pay-offs) is not really a prisoner’s dilemma. This makes the claim that people always behave in self-interested ways into an unhelpful tautology and makes the claim that people sometimes cooperate in one-shot prisoner’s dilemma false by definition. In order to avoid this result, in what follows, I will confine the notion of self –interested choice to choices that are directed at maximizing material benefits of some kind for the person doing the choosing – monetary pay-offs to self in many of the games discussed below. I will also describe games as having various structures such as that of a prisoner’s dilemma, according to the monetary pay-offs they involve, so that it becomes an empirical question to what extent people behave in a self-interested way in those games.

A second, related issue concerns the notion of “benefit” that is assumed when it is said that reciprocity involves returning benefit for benefit. Here too in order to avoid vacuity[[8]](#footnote-6), I will take it to be a necessary condition for a person to be “benefitted” from a reciprocal exchange that the person receive a benefit as a result of an action by another party that the person benefitted values from a purely self-interested point of view. But even with this understanding, we still face issues about what counts as a benefit. As an illustration, consider an *ultimatum game* with the following structure: one player, A, is given a stake of $10. A second player B then decides whether to accept A’s proposal; if B accepts, he receives whatever amount he was offered by A and A keeps the remainder of his stake. If B rejects, both players get nothing. In the U.S. and other industrialized countries, the median offer by A is around $4 and is usually accepted; offers below $3 are frequently rejected (Camerer, 2003). Suppose that A offers B $2 and B rejects. A’s offer, if accepted, would certainly make B better off in material terms with respect to a reference point which is his total wealth prior to the game, but many players in B’s position would regard A’s ungenerous offer as a violation of norms of reciprocity. B’s rejection of A’s offer (an illustration of “negative reciprocity” – see below) reflects this assessment. On the other hand, experimental results also show that if A and B play a prior competitive game of skill in which the winner is told he or she has “earned” the right to be the proposer, responders will accept much more unequal division of the stakes. (Camerer, 2003) Similarly, if A is constrained to a choice between (10, 0) and (8,2), with the first number indicating A’s pay-off, and A chooses the latter, his offer is much more likely to be accepted, indicating that A’s behavior is much more in conformity with norms of reciprocity— again, despite the fact the offer is exactly the same, in material terms, as previously. (Falk, Fehr, and Fishbacher, 2003)[[9]](#footnote-7).

What these and many other examples suggest is that benefits (and losses) are assessed with respect to a reference or status quo point in considering whether behavior is regarded as providing a benefit and expressing reciprocation—to count as a “benefit” the good offered must meet or exceed what is mandated by the reference point[[10]](#footnote-8). Moreover this reference point seems to vary with context or at least it is determined by considerations that are at least in part exogenous to the particular exchange at hand—that is, they come from some independent source such as cultural norms, assumed structures of rights and so on. In an ordinary ultimatum game, the relevant reference point in the contemporary U.S. (but not necessarily in other cultures) is an equal or not too unequal division—say (6,4) or (5, 5). But if A’s choice is “earned” or constrained, this reference point may shift, so that (8,2) does count as a benefit for B. Of course, a number of reciprocity-based theories of distributive justice also exhibit a similar dependence on choice of a reference point or background expectations—these are typically understood to be non-agreement (or non –cooperation) points or baselines of some kind, with reference to which potential gains from cooperation are assessed.

I have been arguing that to characterize the notion of reciprocity, we require some notion of a reference point with respect to which benefits (or costs) can be assessed. If we wish to go beyond this and assess whether some one involved in a reciprocal interaction experiences a “net gain” from the interaction in the sense that his aggregate benefits exceed his aggregate costs, we need additional structure beyond that of a reference point and a notion of benefit; we need some common metric or currency that allows us to add up benefits and costs for each participant. We need still more structure if we wish to require that net gains in a reciprocal interaction must be equal or, for that matter, stand in some other determinate relation to one another (e.g., must be “proportional”). I take Gibbard’s claim that those involved in a reciprocal interaction must “gain” from the interaction to suggest that he imposes the requirement that those involved in the interaction must enjoy or expect a positive balance of material benefits minus costs. In what follows I will *not* impose any such requirement – this is the one important difference (alluded to above) between my understanding of reciprocity and what I take to be Gibbard’s. Although there no doubt are people who will enter into reciprocal interactions only when they think they have positive grounds for expecting gains, I believe there are good empirical reasons (see Section IX ) for thinking that people often enter into interactions that are plausibly described as reciprocal in circumstances when they do not expect such net gains to self. To the extent that reciprocators pay attention to costs, I believe that a more accurate characterization of the behavior of many reciprocators is that they have a tendency to withdraw cooperation when there is clear evidence that their costs substantially outweigh their benefits, but tend, as a default, to continue the interaction when such evidence is lacking. Moreover, many reciprocators seem to care more about whether there has been some exchange of benefits rather than about whether their material gains from the interaction exceed their costs. This is important because in many reciprocal interactions participants may find it difficult or impossible to assess whether or not they are enjoying such a positive net balance of benefits over costs, either because there is no obvious metric for adding up benefits and costs or because both are very diffuse and hard to track. In such cases, participants will have no unequivocal evidence that their costs outweigh their benefits and so as long as they are getting some benefits in response the benefits they confer, many will continue to reciprocate.

Thus, although this may initially seem counterintuitive, I take the empirical evidence to support the view the mere provision of a benefit or the prospect of such provision can activate motives to reciprocate, in a way that is largely independent of costs, with calculations of whether one is on balance gaining or losing from an interaction coming into play only when the loss is large and obvious. Put differently, the motive to reciprocate can be non-instrumental, arising because simply one wishes to respond to the provision of a benefit by providing a benefit in response, rather than because one wishes to be involved in an interaction in which one gains or comes out ahead, at least as measured in material terms. I will suggest below that these features (cost-insensitivity etc.) are crucial to understanding the role played by motives of reciprocation in sustaining many patterns of cooperation, including those involved in contemporary welfare states. The temptation to suppose that reciprocity requires the expectation of mutual net gain derives, I believe, from conflating interactions based on reciprocity with cooperative interactions based entirely on self-interest, which presumably do require such net gains on the part of all interactors.

So far my focus has been on *positive reciprocity*: a motive to return good with good. However, as both ordinary observation and more careful experiment suggest, reciprocity can also be *negative* in character, taking the form of a desire to punish or inflict a loss on those who have behaved uncooperatively toward us even when the imposition of such punishment is not in our own self-interest. For example, rejection of low offers in a one-shot ultimatum game described above is an illustration of negative reciprocity which is not in the responder’s self-interest, at least when this is narrowly defined in material terms. As this last example suggests, positive and negative reciprocity may not be sharply separable since to the extent that positive reciprocity is truly conditional (as it must be, in my view, if it is to count as reciprocity), it must involve a willingness to distinguish in some way between those who behave cooperatively toward us and those who do not, even if this only takes the form of a “withdrawal of concern” or a refusal to continue to provide benefits to non-cooperators or perhaps simply a desire to aid them less than cooperators, rather than to actively harm them. In what follows I will assume that to the extent that people are willing to punish or impose losses on others or to withdraw cooperation, even when it is not in their self- interest to do so, it is appropriate to think of them as also motivated by considerations of reciprocity.

The cases of reciprocity discussed so far have all been cases of *direct* reciprocity: A confers a benefit or cost on B and B reciprocates by conferring a benefit or cost on A, the person who conferred the benefit/cost on B. In *indirect* reciprocity, A confers a benefit or cost on B and reciprocation occurs, prompted by observation of (or information about) this interaction, but at least one of A or B are not the parties to this act of reciprocation. For example, in response to the receipt of a benefit from A, B may engage in indirect reciprocation by conferring a benefit on a third party, C. Or in response to A’s providing a benefit to B, some third party C, C may provide a benefit to A. Although the notion of indirect reciprocity may seem puzzling to some, there is a considerable amount of evidence from both experiments and field studies (as well as casual observation) that it can be an important motivation in human behavior. For example, Duwenberg et al, 2001 find, in an “indirect” version of a trust experiment in which third parties are given an opportunity to reward a trustor who has given money to a second party, that third parties send almost as much money to the original trustor as is sent by second parties in the original version of the game. Other results show that those who are recipients of benefits but do not have the opportunity to reciprocate to their benefactors are more likely to confer benefits on third parties than those who have not received benefits. I will suggest in Section X that indirect reciprocity plays an important role in sustaining many society wide-benefit schemes. In cases of indirect reciprocity it is even more plausible that the motivation of the parties need not be the expectation of mutual overall gain evaluated in self-interested terms, since the original provider of a benefit need not be the person who benefits from an act of indirect reciprocation[[11]](#footnote-9).

So far I have been discussing reciprocity primarily in a descriptive mode—as an account of motivations and behavior that people actually exhibit. But, as already noted, one might also take reciprocity to be the basis of a normative theory of justice—as a basis for recommendations about how people ought to behave and reason: one has reason to benefit those who have benefitted one (or in the case of indirect reciprocity, some other person) in the past, a reason that is absent when there has been no such provision of benefits. Moreover, these two interpretations (descriptive and normative) are intertwined in the following way: it seems uncontroversial (as a descriptive claim) that most people who behave as reciprocators do so at least in part because they take reciprocity to be a normative ideal of some kind—they reciprocate because they think that they ought to reciprocate or because they subscribe to a normative theory of justice that is at least in part based on reciprocity.

What is the common thread running through the examples of reciprocity-based behavior described above? I take the common feature to be the presence of a motivation in interactions with others that is other-regarding rather than purely self- interested, but where the form of the interaction is conditional or contingent on the past behavior or anticipated future behavior of those others. This contrasts with forms of altruistic or other-regarding motivation that are unconditional in the sense of not making the motivation to aid conditional on the behavior of those aided. Broadly speaking, reciprocators have the following motivational profile: they want to behave cooperatively or to aid others who can benefit from aid, but they also do not want to be taken advantage of by others who behave uncooperatively. Reciprocators want to avoid free-riding or exploiting other’s willingness to behave cooperatively but they also don’t want to be taken advantage of or exploited by non-cooperators. They think that considerations of justice do not require that they expose themselves to such exploitation.

To further illustrate what is distinctive about Justice as Reciprocity (JR) consider an example discussed by both Barry and Gibbard. Two groups of people live on separate islands. Because one island is fertile and the other barren, the people on the first island are wealthier. The two groups do not interact at all. Barry understands Justice as Impartiality (JI) as requiring redistribution from the richer group to the poorer group, on the grounds that the differences in wealth between the two groups have a “morally arbitrary” source. By contrast, as Gibbard notes, if there is no interaction or cooperation between the two groups and no possibility of reciprocation of benefits by the poorer group, JR will not provide a reason for redistribution. (As Gibbard also remarks and as I will argue later, this does *not* mean that there is no reason for redistribution at all, but rather simply that if such reasons exist, they will not be based on considerations of reciprocity.) Note that grounding the reasons for aid or resource-sharing or cooperative behavior in considerations of reciprocity is not tantamount to grounding these reasons in self-interest. Suppose, contrary to the original supposition, that the two groups are involved in some ongoing interaction which benefits both or that they have a reasonable expectation that such an interaction is possible. Then according to JR, this gives the wealthier group reasons to behave cooperatively or reciprocally in this interaction, where this may include contributions and resource sharing that go well beyond what considerations of self-interest would suggest. For example, if the wealthier group could benefit from the interaction in some way that fails to benefit the poorer group (e.g., by using their superior military resources to steal from the poorer), JR would prohibit such behavior even if was in the wealthy group’s self-interest.

**IV. Stability**

So far I have focused on the connection between empirical results about human motivation and normative theory within the specific context of a disagreement between Gibbard and Barry. Viewed in a more general way, we can think of this connection as having to do with the role played by the notion of *stability* within normative theories. It is widely (although by no means universally) accepted that it is desirable that a theory of justice (or more broadly a set of recommendations and principles governing behavior in contexts in which cooperation is a possibility) should be such that they draw on motivations that people actually possess and where these motivations have some tendency to lead people to conform to the recommendations in question. In other words, we want schemes of cooperation (or more broadly rules of justice) to be stable in the sense that given the motivations that people actually have, they will not act in ways that undermine these schemes but rather will be led (at a minimum) to act in accordance with them and, ideally to conform to them for reasons and motives that they can endorse as providing good reasons for conformity. Relatedly, it is desirable that the motives on which cooperative schemes draw be self-reinforcing rather than self-undermining in the sense that when some act on those motives, this encourages others to be similarly motivated[[12]](#footnote-10). Perhaps the best known invocation of this consideration is Rawls’ argument that human motivation being what it is, people can not be expected to conform to arrangements and rules justified by utilitarianism, when those arrangements impose large sacrifices on them in order to increase utility for others.

There are many complex issues about how exactly stability considerations should constrain or enter into reasoning about justice. Here I just want to make the larger point that once stability considerations are recognized as relevant, this provides another obvious avenue by which empirical considerations having to do with human motivation can be brought to bear on normative theory. Whether a set of rules or institutions governing cooperation is stable or not depends on descriptive facts about the motivations (and beliefs – see below) people actually have or can reasonably be expected to acquire and on what happens (again as a matter of empirical fact) when people with those beliefs and motivations interact in various settings.

There are other more diffuse ways in which empirical considerations concerning motivation and behavior can bear on normative claims. On any plausible moral view, there should be some connection between moral requirements and (at least some part of) what we care about —ceteris paribus, it is a consideration in favor of an alleged moral requirement that it furthers, implements, protects, or exhibits respect for what we value. What people value is, if not a purely empirical matter, at least a matter to which empirical evidence is relevant. If, for example, people’s behavior and motivations show that they value reciprocity in the relationships that they have with others, or care about whether others respond to benefits provided them by reciprocating, then this suggests a prima-facie case for moral and political theories that attach some significance to reciprocity and values that go along with it. Of course, this presumption is defeasible—some of what people care about may conflict with other values that they hold more dear, or may be based on mistaken empirical assumptions (we only value V because we believe B and B is false.) However, there don’t seem to be any obvious (or non question-begging) reasons for thinking that reciprocity has these features[[13]](#footnote-11).

**V. Reciprocity as a Motive: Empirical Results**

I turn now to a more detailed exploration of some empirical results bearing on the existence and importance of reciprocity as a motive in cooperation.

Prisoner’s Dilemmas. Consider first an ordinary prisoner’s dilemma in which the numbers in the matrix represent the material pay-offs, say in dollars, to each of the players[[14]](#footnote-12).

Column

Cooperate Don’t Cooperate

Cooperate 7,7 0,12

Row

Don’t Cooperate 12, 0 4,4

If this game is played only once, the dominant strategy for both players (if they are self-interested in the sense of caring only about their own material pay-offs) is to defect. If the game is played repeatedly with the same partner and a fixed probability of the game ending in each round, then, as noted above, if this probability is sufficiently low, cooperative strategies like tit-for tat (that is, beginning with cooperation and then responding to cooperation with cooperation and defection with defection) are among the Nash equilibria of the repeated game, but there are many other Nash equilibria, including many that are far less cooperative (such as always defect.) Experimentally, rates of cooperation are found to be depend on, among other things, the details of the pay-offs, with (unsurprisingly) non-cooperation being more frequent when the gain from non-cooperative choices increases. Nonetheless, somewhere between 30 -60 % of subjects cooperate in typical experimental studies of one –shot PDs, with somewhat higher levels of cooperation when a repeated prisoner’s dilemma is played. For example, Andreoni and Miller (1993) find about a 40 per cent cooperation rate in the first round of a repeated ten round “stranger” PD (in which the game is played ten times but with different partners each time—in this case the dominant strategy for self-interested players is to defect. By comparison in a “partner” condition in which ten rounds are played with the same partner, the first round cooperation rate is 55 per cent[[15]](#footnote-13).

These results are of course consistent with the players not being purely self-interested, although other explanations for their behavior which appeal just to self-interest are possible[[16]](#footnote-14). However, even if a player’s choice to cooperate in a one-shot PD is taken to indicate that she is not purely self-interested, this choice does not distinguish between the hypothesis that the player is a conditional cooperator or reciprocator and the hypothesis that she is an unconditional altruist of some kind, since at least for many choices of pay-offs both hypotheses predict cooperation. A somewhat more sensitive test for the presence of reciprocators is provided by the behavior of second movers in a *sequential* prisoner’s dilemma. In this game, player one moves first, choosing either cooperate or defect, and player two is then informed of player one’s choice and decides how to move. If player two is an unconditional altruist, then depending on the pay-off details and the weight she assigns to player one’s welfare, she may cooperate even if player one has defected. By contrast, if player two is a reciprocator then one would expect whether she cooperates will vary depending on whether player one has cooperated on the first move; player two will reciprocate a cooperative move by player one by cooperating herself, but will be considerably less likely to cooperate if player one has defected. (Player two’s non-cooperation in this case might be due to an active desire to retaliate against player one for her non-cooperation or it might reflect indifference or lack of concern about player one, given her non- cooperation). Of course if player two is purely self-interested, she will defect whether or not player one has cooperated.

In a series of experimental tests relevant to the existence of conditional cooperators, Ahn, Ostrom, and Walker (2003) asked subjects to state their preferences regarding outcomes in one shot double blind prisoner’s dilemmas, both involving simultaneous and sequential moves. In the simultaneous move version of the game, 10% of subjects ranked the mutually cooperative outcome (C, C) higher than the outcome in which they defect and their opponent co-operates (D, C). Another 19% said they were indifferent between these two outcomes, despite the fact that the latter (D, C) outcome is payoff maximizing for them. In the sequential move game, forty percent of subjects ranked (C, C) higher than (D, C) and another 27 per cent said they were indifferent between these two outcomes. Lest these be dismissed as mere survey results, Ahn et al also report results from behavioral experiments that are broadly consistent: 36 per cent of subjects cooperated in the simultaneous move game, and in the sequential game, 56 % of first movers cooperated and 61% of second movers cooperated when first movers cooperated, despite the fact that the latter is not the pay-off maximizing choice[[17]](#footnote-15). By contrast, virtually no second movers cooperated when first movers defected. Ostrom comments elsewhere that these results “confirm that not all subjects enter a collective action situation as pure forward looking rational egoists” and that “some bring with them a set of norms and values that can support cooperation.” (2005, p. 129). However, she also adds that

Preferences based in these [cooperative] norms can be altered by bad experiences. One set of 72 subjects played 12 rounds of a finitely repeated Prisoner’s Dilemma game where we randomly matched partners before each round. Rates of cooperation were very low. Many players experienced multiple instances where partners declined to cooperate (Ahn, Ostrom, and Walker 2003). In light of these unfortunate experiences, only 19 percent of the respondents now ranked (C, C) above (D, C) while 17 percent were indifferent (ibid.). In this uncooperative setting, the norms supporting cooperation and reciprocity were diminished by experience, but not eliminated. (2005, p. 129).

These results strongly suggest that a substantial number of people are motivated by and behave in accordance with a conception of reciprocity like that described by Gibbard [[18]](#footnote-16). As a descriptive matter, motivation by (or behavior in accord with) reciprocity does not seem to collapse into the other two motivations mentioned by Barry—self –interest and impartiality. In addition, however, there is another important feature of the experimental results described by Ostrom that is worth underscoring and to which I will return below: the apparent heterogeneity of motivation across different subjects. Although roughly 60 per cent of second movers cooperate when first movers cooperated in the experiment, this leaves approximately 40 per cent who behaved in a self-interested way in this condition. Other experimental results support a similar picture according to which most populations consist of a mix of different types or varieties of subjects, some of whom seem to be strongly motivated by reciprocal considerations (when these are relevant), and others of whom seem to be motivated mainly by self-interest. A smaller group seems more strongly motivated by non-reciprocal, non conditional other-regarding motives even in circumstances in which considerations bearing on reciprocation seem relevant – for example, they will continue aiding others who are capable of reciprocating but do not do so (that is, they will continue to cooperate even in the face of defection by their partners). As we shall see, this heterogeneity has important implications for issues having to do with the circumstances under which cooperation will be stable.

Public Good Games. I turn now to a second set of experiments, which, in addition to illustrating the importance of reciprocity as a motive also illustrate its important implications for the stability of cooperative schemes. In a linear public goods game, each player *i* from a group of *N* players decides chooses an amount *ci* (from an initial endowment *wi* which is the same for each player) to contribute to a public good. The sum *∑ci* of these individual contributions is then multiplied by some factor *m* (*1/N <m <1*) and this total is then divided equally among all of the players, regardless of how much they contributed. Thus each player *i*’s endowment is changed by *–ci + m∑ ci.* This game may be either played once or repeated, with total contributions announced after each round. In both cases, the dominant strategy for a purely self-interested player is to “free-ride”, receiving his share of the contributions of others, but contributing nothing oneself. Thus in this case, unlike the prisoner’s dilemma, there is no Nash equilibrium in which subjects cooperate in the repeated game. In other words, in this case cooperative behavior cannot be understood as a repeated game effect involving selfish players.

Experimental studies of behavior in repeated public goods games reveal the following “stylized facts” (cf. Camerer, 2003):

(1) Average contributions across all subjects begin at around 50% of the total endowment.

(2) There is considerable variation in individual contributions as a percent of endowment, with some subjects contributing nothing from the first round onward and others contributing a very large portion of their endowment.

(3) As the game is repeated, average contributions typically decline sharply over time, but usually not all the way to zero. Roughly 5-10 % of players continue contributing even as everyone else stops contributing.

Why does this decline in contributions over time occur? Broadly speaking, two explanations have been suggested. One is that all subjects are (and remain) largely or entirely self-interested, but are initially “confused” (by the unfamiliarity of the game or other factors) about what will best promote their self-interest. As they repeatedly play the game and receive feedback about the money they earn from their choices, they gradually learn that the strategy which best advances their self-interest is the dominant strategy of zero contribution. The decline in contributions over time simply reflects this gradual learning. (Binmore, 2007) On this interpretation there is no evidence for reciprocity in the sense employed in this essay.

A second, conflicting explanation is this: not all subjects are self-interested, although a substantial number are. A number of other subjects are reciprocators, and a smaller number are some form of altruist or unconditional cooperator. Reciprocators begin by cooperating in the early rounds of the game, expecting or hoping that others will as well. By contrast, self-interested types contribute nothing from the start. The decline in contributions over time occurs because reciprocators gradually stop contributing in response to the non- contributing behavior of self-interested players—reciprocators are not purely self-interested but they not willing to continue to contribute while others free- ride. Given the structure of the game, the only way reciprocators can respond negatively to free-riding is by not contributing themselves – hence insuring that free riders do not benefit from their behavior. Unconditional cooperators account for the small portion of subjects that continue to contribute even as other subjects withdraw from contributing. Subject’s verbal explanations of their own behavior support this second hypothesis, and so do a number of other pieces of evidence[[19]](#footnote-17).

Restart effects. Here a repeated public goods game is played for e.g. ten rounds and then stopped. The game is then “restarted” -- subjects are told that a new ten round repeated game will begin (with the same players) (Cf. Andreoni, 1988). Contributions exhibit the standard pattern of decline over first ten rounds but the effect of the restart is to temporarily boost contributions -- the average level of contribution much higher in the eleventh round. This is prima-facie inconsistent with an interpretation according to which subjects are gradually learning, through repeated play, that zero contribution is the strategy that best advances their self-interest—if they have learned this by the tenth round of the original game (when contributions are low) why do they suddenly begin contributing at a higher level when the game is restarted? A more natural alternative interpretation is provided by the second hypothesis above: a substantial part of the subject pool are reciprocators who are willing to contribute when the restart begins as part of an attempt to signal anew their willingness to cooperate if others are willing to do so as well; their increased level of contribution is an attempt to get others to change their behavior and no longer free ride.

Punishment. A number of recent experiments have explored the effects of introducing a costly punishment option into a repeated public goods game (cf. Fehr and Gachter, 2000) This allows subjects, at some cost to *c* themselves, to punish non-contributors by imposing a fine that reduces their earnings by some amount *r* where *r>c*. Under this condition, a number of subjects punish non –contributors, even in the final round of the game, in which punishment cannot influence future behavior, so that punishment is clearly not a strategy that increases the punisher’s expected pay-off. Introduction of this option prevents the decline in contributions with repeated play that is otherwise observed. An obvious interpretation is that punishers are reciprocators who are willing to engage in negative reciprocation, even when it is not in their self-interest to do so. In the absence of an explicit punishment option, reciprocators have no way of responding negatively to free-riding except by withdrawing their contributions; when a punishment option is provided, they make use of this instead and do not reduce their contributions.

Comparative Statics*.* If a significant portion of subjects are reciprocators, one would expect the contributions of this group to be positively correlated with their beliefs about the extent to which others are contributing, since reciprocators contribute more to the extent that they believe others are contributing more. This prediction contrasts with the prediction that follows from other hypotheses about the motivations of subjects who contribute. For example, a simple model according to which subjects are unconditional altruists who care, in addition to their own self-interest, only about the aggregate amount contributed by all, predicts that individual subjects will contribute less if they believe that others will contribute more. In an experimental study of this issue employing a repeated public goods game, Croson (2006) found, in accordance with the hypothesis that a large portion of subjects are reciprocators, a strong positive correlation between a subject’s beliefs about how much others would contribute and that subject’s actual contributions, as well as a correlation between individual contributions and the actual level of other’s contribution across different periods of the game. Many field studies show a similar pattern—for example, people’s willingness to fully pay taxes is highly correlated with their beliefs about whether others are similarly complying.

Grouping effects. Some of the strongest support for a general picture of public goods games as involving interactions between reciprocators and more self –interested players with declines in contribution, when they occur, deriving from the adverse impact of the latter on the former is provided by a series of experiments involving grouping effects. A representative example[[20]](#footnote-18) is an experiment conducted by Page, Putterman, and Unel (2005) in which subjects were given information every three rounds about the contributions of other subjects in previous rounds and also an opportunity to express preferences about future partners. Subjects whose mutual rankings were the lowest (with a low ranking meaning that a partner is most preferred for cooperation) were then organized into a first group; those whose rankings were intermediate were organized into a second group and so on, for a total of four groups. The result of this “endogenous grouping treatment” was that the average level of contribution across all groups (70%) was much higher than the average level of contribution in a baseline treatment in which there is no regrouping (38%). Moreover, average levels of contribution varied greatly across the groups, with this following the order of group formation – that is, contributions were highest and continued to be so in the first group formed of high cooperators and are lowest in the fourth group: in the first group, 50 percent contributed their entire endowment in the final period (with the corresponding numbers from the second, third and fourth groups being 43, 18, and 0).

Like the experiments concerning behavior in prisoner’s dilemmas, this experiment also suggests that subjects are heterogeneous and differ in type, with some concerned only to maximize their personal pay-offs and others appearing to be some form of conditional cooperator. Evidence for non-self-interested motives comes, among other things, from the observation that across all four groups, 59% contributed at least something in the last period, even though the dominant strategy for pay –off maximizing is to contribute nothing in the last period. Evidence for subject heterogeneity and the presence of a significant number of conditional cooperators is provided by the great differences in the levels of cooperation across the four groups – if there are no differences of some kind among subject, it is hard to see why the grouping treatment should be so effective in fostering different levels of cooperation. Further evidence for subject heterogeneity is reported in (Carpenter and Mathews, 2009) who find that subjects who behave as unconditional altruists with respect to their contributions in a repeated public goods game are also less likely than conditional contributors to punish low contributors when given an opportunity to do so. At least in this experiment subjects who appear to be motivated by positive reciprocity are also more likely to exhibit “negative reciprocity”, a result that seems to support the unified behavioral profile of reciprocators described in Section II.

This heterogeneity of behavior and motivation across subjects has many important implications for normative political philosophy, including implications for the sorts of principles and institutions that will stably sustain cooperation and the distribution of the benefits it creates. Because reciprocators tend to withdraw cooperation (or to respond with negative reciprocity or punishment) when treated non-reciprocally themselves and because primarily self-interested subjects will behave in non-reciprocal ways when it is in their self interest to do so, the presence of self- interested types, if their behavior is unconstrained and they are able to interact indiscriminately with reciprocators, can cause cooperative behavior to break down, even in populations that contain large numbers of reciprocators and some altruists. The public goods experiments described above illustrate this possibility. More generally, whether cooperation can be sustained in such cases will depend on how the different kinds of subject described above interact with each other. This in turn will depend (among other things) on the rules or institutions that are in place to influence such interactions. For example, rules or institutions can encourage cooperation by allowing reciprocators and unconditional cooperators to identify one another and to interact preferentially with those of similar type and to exclude or limit interaction with self-interested types. Or they can undermine cooperation by allowing interaction between different types to occur indiscriminately. Similarly, the appropriate rules or institutions can encourage cooperation by making it possible to sanction exploitive behavior by self-interested types[[21]](#footnote-19). In addition, cooperation may be favored by setting up institutions that make possible the gathering and dissemination of accurate information about the extent to which others are cooperating or not, since reciprocators condition their behavior on the extent to which others are cooperating. More generally, one consequence of the existence of a substantial group of people who are reciprocators and another who are self interested is that finding the right rules and institutions to govern their interaction matters crucially to successful cooperation. I will return to this theme below.

**V1. Ideal Theory and Partial Compliance**

Yet another consequence has to do with the assumptions about the extent of “compliance” with the requirements of justice and the role of “ideal theory” in moral and political philosophy. As is well –known, Rawls’ theory focuses mainly on the project of working out what the correct principles of justice would be in a society in which it is assumed that most people are motivated to conform to those principles—i.e., under conditions under which there is something close to “full compliance”. Then a non-ideal theory which specifies what should be done under conditions that do not involve full compliance is constructed in light of this ideal, the assumption being that the ideal can usefully guide the construction of a theory for non-ideal cases .

Suppose, following this strategy, that the correct principles of justice, whatever they may be, mandate cooperative behavior in public goods games when many others are cooperating, cooperation in a PD when one’s partner cooperates and so on – that is, that all of this is part of ideal theory. Now consider the following: it appears, from the results described above, that a substantial group of people will not be motivated by considerations of justice alone (at least when justice as understood as involving something more than mutual advantage) to comply with these principles and will fail to comply in the absence of incentives of more self-interested sort. Moreover, as we have seen, this non-compliance will affect the behavior of most of those who are motivated to comply by non self-interested considerations, since most such compliers are reciprocators rather than unconditional altruists. Given this is the case, it is natural to wonder whether the assumption of nearly full compliance is can provide useful guidance for real-world cases of partial compliance. After all, in the presence of nearly universal compliance, there would presumably be no need for rules and institutions designed to prevent the problems that can arise when self-interested types interact with reciprocators—no need for rules and institutions allowing reciprocators to interact preferentially with other reciprocators and unconditional cooperators and to exclude or sanction free-riders. But in the real world providing rules and institutions that address such problems seem absolutely crucial if one’s goal is to have a stable system of cooperation. To assume full or nearly full compliance is to assume away many of the problems that are most central to the design of rules and institutions that govern the behavior of reciprocators and their interactions with other type.

We can also frame this as an issue about the how notion of stability is to be characterized with in a framework like Rawls’. One possible characterization of stability assumes something close to full compliance: a system of justice is stable if people largely conform to its principles under conditions in which all or most are assumed to be motivated by a sense of justice. However, this understanding of stability will not be very useful if, as is apparently the case, there are significant numbers of people who lack aspects of this sense of justice and who will act in such a way as to destabilize the rules, given the motivations of other subjects[[22]](#footnote-20). We need a more robust understanding of stability—one that allows us to talk about stability under conditions in which there is motivational heterogeneity and the absence of full compliance. A satisfactory set of rules and institutions should be stable under these conditions and not merely under conditions of full compliance.

The strategy of focusing primarily or almost exclusively on ideal theory might (arguably) make sense if deviations from full compliance were small, infrequent, and minimally disruptive. It would also be reasonable if results about what we should do under conditions of full compliance bore some simple connection to (or could be straightforwardly transferred to provide guidance about) what to do under partial compliance. It is not clear, however, that either of these assumptions is reasonable. As we have seen, people who are reciprocators or who behave as though they think the correct normative theory in situations involving cooperation is a reciprocity-based theory will not think that what they ought to do in the ideal case in which nearly all are contributing to a public good transfers straightforwardly to what they ought to do in situations in which there is wide- spread free-riding. If most people were unconditional altruists, the task of constructing a non-ideal theory might not be terribly urgent, since unconditional altruists will continue to cooperate even if others do not; this task becomes much more pressing once the existence of reciprocators and motivational heterogeneity is recognized.

To drive this point home, it is useful to compare the actual behavior of people confronted with a public goods problem in the experiments described above with the recommendations of normative theories concerning the provision of aid to others. Murphy (1993, 2000) claims, roughly, that in connection with providing aid to others (e.g., badly off people in other countries), one has a duty to maximize well-being but subject to constraint that one is required to do no more than what one’s fair (proportionate ) share would be if all others were complying with this duty ( “In situations of partial compliance… the sacrifice each agent is required to make is limited to the level of sacrifice that would be optimal if the situation were one of full compliance”1993, p. 280). If we apply this suggestion to a public goods problem, it seems to imply that one should determine what one’s share of the optimal level of contribution to the good would be would be if all were contributing their optimal level and then contribute that amount, regardless of what others are in fact contributing. Murphy’s proposal thus implies that, as a normative matter, one’s contribution should be constant and independent of what others are in fact contributing. Richard Arneson (2004) objects to Murphy on utilitarian grounds. Arneson’s view is the normative counterpart of the descriptive proposal, mentioned above, that attributes contribution (to the extent it occurs) to unconditional altruism. Arneson thinks that, given plausible assumptions, a potential contributor should do more if others are doing less since the less others do, the more good one can do by contributing more. This suggests that, under the right assumptions, your contributions, as a normative matter, should be negatively correlated with total contributions by others. By contrast, if we think of justice as reciprocity as a normative theory, then what it seems to recommend is that, other things being equal, people’s contributions should be positively correlated with other’s contributions, assuming of course that the contributions in question will actually provide benefits. As we have seen, it is this, rather than the pattern of contributions recommended by Murphy or Arneson that corresponds to the pattern that is actually observed.

The recommendation that, as far as considerations of reciprocity go, one’s contributions in a public goods type game should be positively correlated with other’s contributions will no doubt strike some as obviously morally objectionable. Without attempting a detailed defense, let me suggest that this is more to be said in favor of this recommendation than one might initially suppose. First, the recommendation is about what follows from considerations of reciprocity taken in themselves; as we will see, it does not rule out the possibility that there are other sorts of moral considerations, not grounded in reciprocity, that mandate contributions that are not conditional on whether others are also contributing. Second, what the recommendation really amounts to is the claim that whether or not there exists an ongoing scheme of cooperation, in which others are contributing or may reasonably be expected to contribute to some public good makes a difference to one’s own obligation to contribute. In the presence of such a scheme and others’ cooperative behavior, one has obligations to contribute that one may not have if no such scheme is in place. If no functioning scheme is in place and most others are not contributing, then, rather than requiring that one should nonetheless contribute, it may be better that individuals and organizations should instead devote whatever moral energy they have to attempting to create ongoing systems of contribution and cooperation in which most others do (and thus everyone has the obligation to) contribute – better because to the extent that people are reciprocators, they are likely to be resistant to demands that they contribute when most others do not, and because setting up schemes in which many contribute is more likely to result in optimal or acceptable levels of the good in question than insisting that individuals contribute independently of the contribution of others. Finally, to anticipate some remarks below, an obvious normative attraction of a disposition to reciprocate is that in the right circumstances it can encourage cooperative behavior on the part of people who are otherwise not inclined to cooperate, as when selfish types discover that they can do better for themselves by cooperating rather than defecting in repeated prisoner’s dilemmas with partners who are reciprocators. By contrast, a requirement that individuals contribute even when others are not (or that they contribute more as others contribute less) does not discourage and may in some circumstances encourage free –riding behavior. Hence in populations in which many are reciprocators, such a requirement may lead to the unraveling of cooperation. Put more bluntly, those who contribute independently of what others are doing or who contribute more as others contribute less can actually be dangerous to the maintenance of cooperation, in real world circumstances in which some will behave exploitively and others withdraw cooperation in the face of such behavior.

**VII. The Scope of Reciprocity**

I turn now to some remarks on the *scope* of reciprocity as a normative consideration, which I take to have to do, roughly, with question of whether reciprocity is all or only part of justice. What follows is in part conceptual reflection (concerning the contours of the notion of reciprocity, as it figures in everyday moral thinking), but is also in part motivated by experimental results, either already described or described below.

Although there are certainly exceptions, a great deal of recent philosophical discussion seems premised on the assumption or expectation that a successful theory of justice should be monocriterial (or as close an approximation to this as possible) in the sense of postulating a single guiding idea (or a small set of these) that capture all of the considerations relevant to our sense of justice. The conflict between justice as mutual advantage, justice as impartiality, and justice as reciprocity is naturally understood in these terms; for example, Barry assumes that justice can be at most one of these and that if one of these characterizations is correct, then the other approaches do not capture any of the considerations relevant to assessments of justice. As I see it, it is this line of thought that underlies the most common objection to reciprocity based accounts of justice; namely that such approaches are unable to account for the obligations of the able-bodied to provide help and resources to those who are unable to reciprocate. Thus, for example, we find Alan Buchanan characterizing (and objecting to) justice as reciprocity in the following way:

What is common to [different] versions of justice as reciprocity is what may be called the reciprocity thesis: the claim that only those who do (or at least can) make a contribution to the cooperative surplus have rights to social resources. The implications of the reciprocity thesis for the treatment of severely disabled persons are as disturbing as they are obvious. If, for whatever reason, an individual is never capable of being a contributor to the cooperative surplus, then that individual has no right to social resources whatsoever-not even the most minimal support-even in an affluent society. The clearest case would be that of a person who from birth was so severely and permanently incapacitated that he could not contribute (1990, pp 230-31)

In this passage Buchanan is in part simply describing a view about the relationship between justice and reciprocity which he attributes to other writers, such as David Gauthier. As described, this view is indeed objectionable, for reasons that Buchanan sets out in his essay. It is worth underscoring, however, that the view described in the passage quoted above does not, as an empirical matter, capture very well the role played by appeals to reciprocity in ordinary moral thinking and motivation[[23]](#footnote-21). Instead, the empirical evidence seems to strongly support a picture according to which a number of distinct considerations and motives (of which desires to reciprocate are just one) influence decisions about aid and cooperation and hence people’s normative assessments of these. When reciprocity is possible (see below), it seems to be a very powerful other-regarding motive, although not the only such motive. When reciprocity is not possible, other sorts of other-regarding motives often come into play.

As an illustration, return to the two trust game variants discussed in Section II (Figures 1 and 2). We noted there that the empirical results from this game provide evidence that a substantial number of people are reciprocators. Here, however I want to focus on a distinct point. This is that a substantial minority of players choose to move across even in the second version of the game, —these subjects forgo a larger payoff for themselves in order to provide a larger pay-off for the other player even when the motive to reciprocate is absent. So while the desire to reciprocate is one motive that influences second movers to choose in an other-regarding manner, it is not the only such other-regarding motive. Many other games show a qualitatively similar pattern of other regarding behavior in circumstances that do not involve reciprocal interaction[[24]](#footnote-22) .

In understanding these results, the following distinction will be helpful. Some people may not be capable of reciprocation of any relevant sort at all -- e. g., people who suffer from serious handicaps or very young infants. In other cases, people may have normal adult capabilities but the contingencies of their situation may prevent reciprocation – as when you provide one-time help to a stranger who for some reason will not be able to identify you for purposes of reciprocation. Similarly, the rules for the game in figure 2 do not permit subjects in the position of player one to reciprocate kind acts by player two, even if they wish to. Let us say that in these cases the subjects are *incapable of reciprocation*. In other cases, those involved in an interaction will be perfectly capable of reciprocation but, for one or another reason, will decline to reciprocate—let us call these *non-reciprocators*. These will include, for example, those who move down in the version of the trust game in Figure 1.

With these distinctions in mind, I believe that empirical results support the following picture of the scope of reciprocity as a motive – a picture which I believe also comports with our commonsense understanding. Many subjects (but, as we have seen, far from all) are influenced, to different degrees, and in addition to self-interest, *both* by motives to reciprocate *and* by other –regarding or altruistic motives of a more unconditional sort. In situations in which reciprocation is not possible – e.g., when, as in the game in Figure 2, subjects in the second position must choose among different possible distributions of money which will be divided among themselves and other subjects who have no way of reciprocating these choices-- choosers are guided not just by self-interest but by other-regarding motives of an unconditional or non-reciprocal sort, establishing that such motives are sometimes operative in human decision-making. On the other hand, as we also see from Figures 1 and 2, when reciprocation is possible, motives and preferences tied to reciprocation seem to exert a powerful independent influence on behavior, although not to the complete of exclusion of motives for helping of a more unconditional sort. Other experiments show that reciprocators tend to treat those can be aided but are incapable of reciprocation quite differently from those who are capable of reciprocation but fail to provide it – reciprocators are much more likely to aid the former and respond with either hostility or withdrawal of concern toward the latter[[25]](#footnote-23).

This in turn suggests a somewhat different way of thinking about the significance (and limitations) of reciprocity than what is suggested by Buchanan’s critique (and the treatments of justice as reciprocity on which he focuses). Instead of understanding reciprocity-based conceptions of justice as implying that those not capable of reciprocation are due nothing as a matter of justice, it is much more in accord both with ordinary moral thinking and with what is suggested by experimental results to think of a reciprocity –based conception as capturing a set of considerations that are relevant to decision-making when dealing with people who are capable of reciprocation, but which do not apply when dealing with those incapable of reciprocation. In saying that considerations of reciprocity “do not apply”, I mean that in such cases considerations of reciprocity generate no obligation to aid in themselves, but that as far as reciprocity goes, it is left open whether or not other sorts of moral considerations, not based on reciprocity, provide reasons to aid or share. In other words, the idea is that considerations of reciprocity (Justice as Reciprocity) have a distinctive but limited domain of application (having to do with relationships between people who are capable of reciprocating). They capture part of (or one set of considerations relevant to) justice but not all of justice—not the part having to do with obligations to aid that are not conditional on the choices or behavior of those aided. These unconditional obligations may require aid in circumstance in which reciprocity is not possible[[26]](#footnote-24). Put slightly differently, reciprocity tells us something about how to act toward others in circumstances in which reciprocal behavior by those others is possible; it does not tell us that we should not help when reciprocal behavior is not possible.

This seems to better correspond to how many people think about the significance of reciprocity than the monocriterial picture suggested by Buchanan’s critique. My conjecture (which I take to be the consistent with empirical results described above) is that most people who assign weight to reciprocity in their dealings with others do not think that this requires them to suppose that they owe nothing to those with whom they are not in a reciprocal relationship – particularly if the reason for this is that those others are not capable of reciprocating in any relevant way. For example, most of us would regard someone who refused to provide aid to a newborn baby left on his doorstep on the grounds that the baby is incapable of reciprocating, and that one can have only have obligations to help potential reciprocators, as a sociopath, rather than as someone who showed a proper appreciation of the role of considerations of reciprocity in moral decision-making. Similarly, as I see it, it is not an implication of how we think about reciprocity that the state should not provide support of various kinds for those who are incapable of reciprocity. More generally, it is not part of how we ordinarily think about reciprocity that obligations to aid can only be based on reciprocity. Thinking about the significance of reciprocity in this way allows us to avoid the mistake of arguing that because reciprocity-based considerations are not the whole of justice, they do not matter (much) even in circumstances which fall into the proper domain of reciprocity-- that is, circumstances in which we are interacting with people who are capable of reciprocating.

**VIII. The Role of Intentions and Relationships in Reciprocity**

Another feature which one finds in the empirical literature that is of considerable importance to understanding the structure of reciprocity is that reciprocators care not just about the outcomes which they receive as a result of other’s choices but also about the *intentions* with which others act and the attitudes toward other potential reciprocators expressed in their actions. A choice by A which confers a material benefit or loss *x* on B may be viewed by B in one context as revealing a kind or cooperative intention on the part of A and in another context as revealing a hostile or uncooperative intention and this will in turn influence how B responds to A’s choice. In other words, Bs care not just about the material benefits or costs they incur as a result of A’s choices but also about what those choices seem to reveal about A’s attitudes and motives. (This is no doubt in part because those motives and attitudes are often taken to reveal something about A’s likely future behavior toward B, but A’s motives and intentions also seem to matter to B independently of this consideration, since they influence B’s response even in one-shot games.) The pair of ultimatum games discussed in Section II again provides an illustration. A proposer’s choice of (8,2) when the only alternative is (10,0) is not regarded by most responders as expressing a hostile or uncaring intention toward them, while a choice of (8,2) when divisions like (6.4) and (5,5) are also possible is taken to express such an intention and is much more likely to provoke rejection. Similarly an unequal division of the stake in an ultimatum game which is known to be generated randomly by a computer is much more likely to be accepted than the same unequal division when it is chosen by a human proposer—again presumably because the latter but not the former is taken to reveal a hostile or uncaring intention or attitude. The same pattern holds for acts of positive reciprocity. If, say, I provide you with a benefit but only as a side-effect of some action that I undertake for self-interested reasons (or the choice of which is fully explainable in terms of my self-interest) you will be much less likely to reciprocate than if I seem to choose the action because I wish to benefit you, even though some alternative which is more optimal with respect to my self-interest is available.

A related point is that, insofar as they are reciprocators, people seem to care about *who* performs acts of reciprocation and not merely about the benefits or costs conferred by those actions. If A has benefited from some kind action of B, then insofar as A is motivated by a desire to reciprocate, this will be a desire that *A herself* intentionallyprovide some benefit in response to B’s prior action (typically, of course, this will be a benefit to B, but as the possibility of indirect reciprocity illustrates, this benefit may be directed to someone else). Insofar as A’s desire is a desire to reciprocate it will not be just a generalized desire that B receive a benefit somehow or other, perhaps as a result of the action of some third party. Similarly, B will care about whether it is A who acts so as to benefit B in response to her (B’s) prior kindness, and not just that she receive a benefit from some source or another that “compensates” her for whatever “costs” she may have incurred in benefitting A[[27]](#footnote-25). By contrast, some unconditional altruists may care only about the aggregate level of well-being of themselves and others and not, independently of this, about whether actions that confer benefits are reciprocated or who confers benefits to whom[[28]](#footnote-26). For example, if A and B are such altruists and B provides a benefit to A at cost *c* and B then receives from C some benefit that more than offsets *c*, A may not regard himself as under any obligation to reciprocate to B and may instead think it more appropriate to direct his benevolent impulses elsewhere, where they will do more good. Benefit-conferring actions motivated by unconditional altruism thus may exhibit “crowding out” effects that are not present when such actions are motivated by reciprocity, in the sense that altruists may see no reason to confer benefits on those who have already been sufficiently benefited by the actions of others (and perhaps no reason to act altruistically at all if all relevant potential recipients have been sufficiently benefitted) while reciprocators may think they have an obligation to provide benefits to those who have previously helped them, regardless of whether they have also been helped by others and regardless of how their overall level of well being compares with others.

More generally, reciprocity does not merely have to do with the exchange of material benefits or costs or with the equitable allocation of benefits and costs among interacting parties; it also has to do with the *relationship* between the intentions or motives with which participants act in the course of their interactions with one another; with whether interactions are characterized by a kind of mutuality (or fit or correspondence) of intentions in which participants act with the intention to respond to kindness or cooperative behavior with kindness or cooperation and to uncooperative behavior with hostility (or at least withdrawal of support). One result of this is that the notion of reciprocity seems to resist full characterization in purely outcome based or “consequentialist” terms, in the sense that it has to do not just with material pay-offs that subjects receive but with the beliefs and intentions with which subjects act, their beliefs about one another’s beliefs and intentions and with the establishment of the right kind of relationship between these. In the importance that it assigns to intentions and to relationships among participants that resist easy characterization in purely outcome- based terms, reciprocity thus incorporates elements that are often thought of as “deontological”. The fact that reciprocators care about these other elements in addition to the material benefits they receive from reciprocal interactions provides another reason why models of reciprocity based on (just) mutual material gain are inadequate.

**IX. Does Reciprocity Require Mutual Gain?**

I turn now to a more detailed look at an issue discussed briefly in Section III. Positive reciprocity involves returning a benefit for a benefit conferred or providing a benefit in the expectation of a benefit. We noted above that what counts as a benefit must be assessed relative to a reference point of some kind such that that are favorable with respect to that reference point count as gains and unfavorable changes count as losses or costs. Suppose then that X provides to Y a benefit *B1* (where as before *B1*is a “material” benefit to Y assessed relative to some appropriate reference point) and that it costs X amount *C1* to provide this benefit (where *C1* is a material cost also is assessed relative to some appropriate reference point) and that Y responds by providing some benefit *B2* to X at cost *C2* (that is, to Y). In Section III, we distinguished several different possibilities or conditions for reciprocation: 1) X and Y might think of themselves as in a relationship of positive reciprocation to one another as long as they believe that *B1* >0 and *B2* >0. Satisfaction of this condition (or the expectation that it will be satisfied) might be sufficient to activate motives to reciprocate in both, independently of any costs associated with this action. 2) More strongly, X and Y might require, in addition to (1), that (2) for reciprocation both “gain” over where they would be in the absence of this interaction, taking into account both the benefits they receive and their costs. That is, X and Y might require that (they believe) *B1*- *C2*>0 and that *B2*-*C1*>0 for them to regard their interaction as reciprocal and for motivations to reciprocate to be operative. Finally, as a third possibility, both might require, in addition to (1) and (2) that (3) for a reciprocal interaction, the overall gains of both must be positive and equal (*B1*- *C2*)= (*B2*-*C1* ) or perhaps “proportional” in some way: (*B1*- *C2*)= *k*(*B2*-*C1)*. (1) is a plausible necessary condition for a relationship being “reciprocal”, but what about (2) and (3)?

The first thing to note is that it is often straightforward to determine whether each interactor has benefitted materially from the other’s action—that is, whether (1) *B1,2 >0* is satisfied. It is often harder, in real life cases, to assess whether there has been overall gain from an interaction in the sense demanded by (2) and even harder to assess whether (3) holds. Determining whether (2) is satisfied requires, among other things, a common “currency” or metric for adding up benefits and costs as well as information about the magnitudes of these. Many reciprocal interactions involve “exchanges” of very different kinds of aid or benefit and the incurring of very different kinds of costs and it may be very unclear how to assess these along some common scale. This seems particularly clear in interactions among friends. Suppose I drive you to the airport and that you invite me over to dinner when you return. My “costs” may include the time and expense involved in the drive; the dinner is among the benefits I receive. Your benefits includes the ride to the airport, your costs the time and expense of making dinner. If we ask which if either of us came out “ahead” or experienced a net “gain” from this interaction, it may be unclear how to answer this question. For one thing, there typically will be many other more diffuse benefits involved in addition to those described above—the interesting conversation we have on the way to the airport, the pleasure we take in each other’s company, the mutual good will engendered and the sustaining of a relationship that may be beneficial to one or both of us in the future in unanticipated ways, perhaps reputational benefits involved in signaling to others that you and I are good prospects for reciprocal relationships and so on. As we try to take into account how the probability of various future benefits and costs (suitably discounted) associated with our relationship may be affected by our present interactions, the calculation becomes, to say the least, uncertain and difficult. The fact that reciprocators care not only about material benefits and costs but the intentions and attitudes that underlie these and about who provides such benefits makes the notion of any exact cost/ benefit analysis even more problematic[[29]](#footnote-27). More fundamentally, if we are friends, we will not, at least consciously, engage in such calculations concerning exact costs and benefits, even though our relationship involves ongoing reciprocation. Of course we might, in the spirit of revealed preference theory, take the very fact that we voluntarily engage in the activities described above to show that both of us gain on balance from them, since otherwise we would not engage in them. However, as already noted, this trivializes the requirement that for reciprocation both parties must “gain” from the interaction.

In view of these considerations, it is not surprising that empirical studies of reciprocal interactions involving friends and acquaintances in real-life contexts tend not to find the kinds of precise quid pro quos (or expectations of such quid pro quos) that are suggested by models of reciprocity as mutual gain or equal exchange (see, e.g., Silk, 2003). Instead, what empirical studies (as well as ordinary experience) suggest is a much looser and vaguer pattern of accounting: often (and particularly when there is some degree of closeness between the parties) those involved in reciprocal interactions don’t even try to keep track of the precise balance of costs and benefits (both because this is difficult or impossible and because it seems contrary to the character of the relationship the reciprocating parties would like to maintain between themselves) but nonetheless are sensitive in the long run and in the aggregate to large imbalances, especially insofar as these result from failures to reciprocate in situations in which reciprocation is possible. If I invite you over to dinner a number of times and you invite me back a fewer number of times, but nonetheless fairly frequently, I may pay little attention to this imbalance, but if I invite you over repeatedly and you never reciprocate, I am likely to eventually stop inviting you.

A natural reaction to these observations is that while precise tracking of costs and benefits may not be characteristic of friendships and other close personal relationships, it nonetheless may be characteristic of reciprocity insofar as it involves more impersonal interactions and large scale patterns of redistribution. However, even this is far from clear. One relevant example is provided by anthropological studies of patterns of food distribution in human groups of a few hundred people. Such studies show strikingly large imbalances in food “flow” among both individuals and subgroups (Kaplan, Hill, Lancaster, and Hurtado, 2000, Kaplan and Gurven, 2005). In particular, young and middle aged males generate many more calories from hunting and other activities than they consume, with the surplus going to women, children, and the elderly. Indeed, according to these authors, the ability to sustain such imbalances played a crucial role in human evolution, since it provides large amounts of protein, essential to brain development, to children who could not obtain it from any other source[[30]](#footnote-28). Moreover, there are large differences in hunting and food gathering skill among men in this age group, so that they generate differing amounts of food resources. In addition, there is considerable variation in food consumption across families with some having more children and other dependents than others. Kaplan and Gurven comment that

therefore, even among men of the same age, there must be net transfers over the long-term from families producing a surplus to families producing a deficit. (2005, p. 93)

They add that

In terms of reciprocal altruism [that is the sort of tit-for –tat exchange based just on mutual self- interest] there is no incentive for young or older adults with small families to support older adults with large families. Those older adults with large families will never produce a surplus to ‘pay back’ those subsidies. The older adults are likely to die before the young adults reach the age in which they need assistance to support their families (2005, p. 94)

Nonetheless young and middle aged adults do provide such assistance. Why? One possibility is that they receive benefits in some form besides food such as increased mating opportunities and more off-spring, with the consequence that food-sharing produces an overall gain in terms of increased fitness for them (as well as for those with whom they share food)[[31]](#footnote-29). However, it has been difficult to establish that such benefits are present in sufficient magnitude to offset the costs associated with hunting and food-sharing—another illustration of the difficulties facing treatments of reciprocity that demand precise systems of accounting of costs and benefits. Another possibility that probably captures at least some of the factors at work is that food sharing functions as system of indirect reciprocity in the sense described in Section 1II: young men contribute food to older men who are no longer able to hunt effectively in the expectation that when they are in a similar position, young men from the next generation (although in many cases not the particular people they have helped in the past) will provide support for them. Arguably, they would be less likely to receive such support when old if they did not provide it to others when young. Finally, to anticipate a possibility to be discussed below, it may be that the societies in question have simply been successful in establishing norms that lead people to share food even when they do not benefit materially (either in the form of more offspring or in some other way) from doing so—people who conform to such norms are praised and acquire high prestige, those who violate the norms are criticized and sanctioned in various ways—and this is enough to generate compliance with these norms. If societies with such norms do better on average in terms of the survival of their members than societies without such norms, and norm compliance can be enforced cheaply and effectively, this may be enough to sustain such arrangements, even if the most successful food producers do not obtain material gains from this arrangement.

Given the features just described, some readers may doubt that food sharing is based on “reciprocity” in any sense, rather than on other forms of other-regarding preferences. Yet the sharing in question has many of the features of reciprocity-based distribution. In particular, sharing is to a large extent conditional on the behavior of those who receive food and sanctions are employed in response to behavior that is seen as exploitive or advantage-taking. Thus if someone is viewed as lazy in the sense of not contributing to the group even though he or she is capable of contributing something (e.g. by constructing tools, providing child care etc.) generous sharing of food will be withdrawn, the person in question will be the object of gossip, may be avoided socially and so on. Thus the relevant norm of reciprocity is understood in terms of the idea that those who benefit from food-sharing should contribute something in response, insofar as they are able. (Kaplan and Gurven, 2005) Note, however, there is no necessary requirement that such contributions must outweigh or be equal to whatever benefits are received, since some of those benefitted may not be able to make contributions of such magnitude. I suggest that this is a recognizable idea of reciprocity even if does not conform to models of reciprocity that stress the notion that all must on balance “gain” from the reciprocal interaction. It is a notion of reciprocity based on (1) above -- that reciprocity requires only the exchange of some benefits-- rather than on (2) or (3).

Turning now to societies like our own, many interactions seem to have a similar character. In many human interactions, some will have resources or skills that others do not (you are wealthy and I am not, you can drive, I cannot; I can cook, you cannot; I have a house suitable for entertaining numbers of people and get pleasure from doing so, while you do not, you are an accomplished professional or expert in some area and I am a novice just beginning to learn) and this will very likely lead to asymmetries or inequalities in the values of benefits exchanged and costs borne. Yet there are recognizable notions of reciprocity (and failure to reciprocate) which apply to such relationships. If, say, you always drive when we do something together (because I cannot drive) and I lack the money to pay for gas (and you wouldn’t accept my money even if I were to offer it), there will nonetheless be other things I can do which will signal a desire to reciprocate—I can call regularly to see how you are doing, send you a birthday card, include you in my circle of friends and so on. My behaving non –reciprocally would involve such actions as accepting your generosity in driving without making any effort to respond to it with kindness and gratitude, exploiting your willingness to be helpful while ignoring your welfare whenever it suited my purposes, and so on. Note this is connected to the observation, made above, that the intentions and attitudes with which people act and not just the material value of the benefits they confer make a difference to whether their actions are viewed as reciprocal—my behavior can count as reciprocal because of the attitude it expresses toward your welfare even if the material benefits I am able to offer you are small in comparison with what you provide to me. Of course at some difficult-to-identify point the distribution of benefits and burdens in a relationship may become so uneven or one-sided that it is no longer “reciprocal” in any plausible sense, but my claim is that this does not automatically happen as soon as one party’s material costs exceeds his benefits from the relationship[[32]](#footnote-30) [[33]](#footnote-31).

**X. Reciprocity as a Basis for Distributive Justice**

With these ideas in mind, let us turn to the idea that reciprocity can serve as a basis for a notion of distributive justice realized by a modern welfare state (conceived as a state in which there is significant transfer of resources from the better to the worse-off). I have already observed that justice sometimes requires forms of unconditional altruism. Insofar as some policies of welfare states derive from such unconditional motivations, it will not be true that all such policies are founded on a concern with reciprocity. Let us, however, put this consideration aside and ask how, insofar as the notion of reciprocity can be used to provide a partial justification for some of the institutions of the welfare state, such a justification might go.

If one adopts a conception of reciprocity like that associated with (2) above (reciprocity requires mutual gain), there is a prima-facie puzzle about how to construct such a justification. Suppose we are guided by an account of justice according to which there should be a net transfer of resources from the well-off to the less well-off. It is obvious how the less well-off gain from this transfer but how does it benefit the well-off? If it does not, and reciprocity requires mutual gain, in what sense is a relationship of reciprocity present between the less and better well-off? How, from the perspective of the better off, can such transfers be seen as grounded in a concern with reciprocity?

As Gibbard (1991) suggests, there is a possible answer to this question, arguably due to Rawls (or at least attributed by many to Rawls[[34]](#footnote-32)). This attempts to explain why the principles of Rawls’ theory, such as his difference principle (and for that matter, any other theory which mandates redistribution from the rich to the poor) follow from a conception of justice based on reciprocity, conceived in terms of mutual gain. First, reciprocity requires, as noted above, a reference point or baseline with respect to which benefits and costs are to be assessed—take this to be an equal division of resources of some kind. Then it is obvious why, under the difference principle or some other principle that mandates transfers to the worst –off, this group gains, on balance. But (the argument continues) the better-off groups gain too if in exchange for the transfers mandated by the difference principle, the worse-off groups do not withdraw their willingness to behave cooperatively toward the better –off and/or do not withdraw their contribution to the social product. In Gibbard’s version, this cooperative/contributive behavior on the part of the worst –off includes eschewal of violent measures like appropriation of whatever resources are possessed by the better-off. In this respect, both groups gain from the implementation of the difference principle or a similar set of transfers and this is what makes their relationship one of reciprocation.

While I do not dispute that this is one consideration that may motivate the better off to support transfer payments, I think that it is strained to think of it as a reciprocity –based motive. Suppose you invest your time and energy growing apples, while I do not, for whatever reason. I do, however, make it clear that I will not unilaterally expropriate your apples, conditional on your giving me some of them. Should we think of this as a relationship involving reciprocation, in which my non-expropriation is regarded by you as the provision of a benefit of the sort that invites or requires reciprocation? My guess is that, as an empirical matter, you are unlikely to see things in this light. My expressed willingness not to take the apples, conditional on receiving some of them, is much more likely to strike you as a threat or as extortion, rather than as the provision of a benefit[[35]](#footnote-33). Nor is this merely a semantic observation about the word “reciprocation”. The deeper point is that, as an empirical matter, my failure to expropriate is not, in itself, the sort of behavior that is likely to awaken in you an impulse toward reciprocation, understood as the adoption of intentions toward me that reflect kindness and goodwill. Instead, to the extent that you and I reach an understanding in which I give you some apples and you don’t unilaterally take what you wish, your motivation is likely to involve a calculation that it is in your self-interest to give in to my threat, rather than resist, likely accompanied by feelings of hostility toward me. In turn, this motivation is likely to lead to a different kind of interaction between us than one motivated by a desire to reciprocate. I think the evidence supports a similar conclusion about what happens on a larger scale when society-wide transfer payments are at issue—it matters whether public support for these, especially among well-off and powerful groups, is motivated by a calculation that it is in their best interest to pay off less well-off groups in order to preserve public order and avoid expropriation or whether instead the motivation is more genuinely reciprocal.

This conclusion is reinforced by another consideration. Although Rawls and others often write as though better-off groups would be worse off if less well-off groups were to withdraw their cooperation or their contribution to the social product, there are many circumstances in which, as an empirical matter, it is far from clear that this is correct. In some circumstances better-off groups may do better for themselves if they were able to avoid interacting with less well-off groups at all (especially if such interaction requires transfer payments) and instead engaged in cooperative activities only among themselves. If such withdrawal is impossible, it still might better serve the interest of the better-off if they were to respond to any perceived threats from the less well-off by adopting strong self-protective measures—by hiring more police, building more prisons and so on (as is increasingly done in the contemporary U.S.) More generally, if we say that reciprocity between two individuals or groups requires that both “gain” overall from the interaction and if we appeal to the considerations described above to show how the better-off gain from transfer payments to the less well-off, we open ourselves to the possibility that circumstances may be such that the better-off could “gain” even more by adopting other less benign measures in their dealings with the worse-off. The tendency of the better –off to search for such alternatives (and to think themselves justified in adopting them) is likely to increase to the extent that they think of their only reason for behaving “reciprocally” is the threat of expropriation.

For these reasons, I think that it is worth exploring another possible way in which the notion of reciprocity might be used to provide a basis for the welfare state. This alternative understanding has already been described above—the key idea is that it rejects the claim (2) that reciprocity requires mutual overall gain. Instead, it rests on the idea that reciprocity requires satisfaction only of condition (1) above: that benefit be returned for benefit or for the expectation of benefit. At the level of individuals this means that two people may be in a reciprocal relationship (and both may be motivated by considerations of reciprocity in their interaction) even though one of them does not gain on balance from the interaction, at least when “gain” is evaluated in material terms. Similarly, in the case of groups, the less and more well-off can be in a relationship of reciprocity even though, especially in the presence of transfer payments, the better-off may not benefit materially from this relationship. The food-sharing schemes described above may illustrate this possibility. By understanding the notion of reciprocity in this way, we avoid the need to invent spurious ‘benefits” that the well-off must be receiving if their interactions with the less well-off are reciprocity-based.

Applied to groups or large-scale organizations like the state, this proposal raises at least two issues. First, what might (or should) the reciprocal behavior of the less well-off groups consist in or involve, if it does not involve providing a net material benefit to the rest of society? Second, is it really psychologically realistic to suppose that well-off groups may be motivated by considerations of reciprocity to support programs that benefit the less well –off but at a net cost to themselves, as long as the less well-off reciprocate in some relevant sense? Under what conditions is this likely to happen?

I will attempt to answer both questions together. I begin with the observation that in both the United states and western Europe support for the welfare state and redistribution is strongly correlated with beliefs about whether poverty is the result of laziness and irresponsibility or whether instead it is the result of factors that are outside of the control of the poor (such as “bad luck”). In addition, such support is also very strongly correlated with whether respondents believe that programs providing for redistribution are being exploited by able-bodied people who accept assistance but fail to work hard, behave responsibly, or develop skills that will make them no longer dependent on state aid. (For details, see Fong, Bowles and Gintis, 2005. I emphasize that here and in what follows, I am reporting attitudes and motivations people seem to possess, rather than endorsing these attitudes.) Even within the United States (which is a less generous welfare state than virtually any other industrial democracy) large majorities say that they would favor increased aid to those who are poor as long as this aid goes to those whose poverty is not the result of their own choices (the elderly, the sick, the handicapped and so on) or in the case of those who are able to work, as long as the aid goes to those who are willing to work and /or to develop skills which will make them more employable or more able to contribute to the social product. What seems to create resentment, indignation, and unwillingness to provide help is not (or not mainly or simply) the costs of the aid (or whether aid is seen as a net gain or loss for those providing it) but rather whether the aid is seen as permitting or even encouraging (what is considered to be) free –riding or exploitive or otherwise “bad” behavior. It is this sort of behavior which is viewed by aid-providers as a failure to “reciprocate” and which awakens indignation or at least withdrawal of concern. Someone who accepts aid and is able but unwilling to work or behave responsibly is viewed as taking advantage and behaving non-reciprocally. Someone who accepts aid but contributes something back in the form of work or at least attempts to do so or who attempts to develop skills that would allow for such contributions is seen as reciprocating in the relevant sense, even if the value of the aid received exceeds what is returned. In their study of the role that considerations of reciprocity play in support for the welfare state, Fong, Bowles and Gintis, 2005 summarize their conclusions this way:

Our interpretation… is that because of strong reciprocity[[36]](#footnote-34), people wish to help those who try to make it on their own, but for reasons beyond their own control, cannot. People wish to punish-- or withhold assistance from-- those who are able but unwilling to work hard (2005, p. 285)

They add:

The task of politically viable egalitarian policy design might thus begin by identifying those behaviors that entitle an individual to reciprocation. Among these in the U.S. today

would be saving when one’s income allows and working hard and taking risks in

both productive endeavors and schooling (2005, p. 297)

I am well aware that many readers will regard proposals that make benefits or income transfers conditional in some way on the choices or behavior of those aided as deeply morally unattractive. This is not the place to discuss this issue in detail; however, some very brief comments may be helpful. First, I assume that to the extent one regards reciprocity in the sense under discussion as a suitable basis for (or important element in) welfare or income transfer programs, one will also be committed to ensuring that those who are potential recipients have meaningful opportunities to reciprocate—this means, among other things, serious state support for education and job/vocational training that enable the development of skills that makes positive contribution possible, as well as the availability of state sponsored forms of employment when other forms are unavailable. Second, given that many people are motivated by such impulses, it seems plausible that programs of aid organized in this way are likely to attract more public support and to be more generous toward many of those in need of aid, than programs that disregard the significance of reciprocal motives. In other words, even if one is tempted to argue that in an ideal world, citizens would be motivated by unconditionally altruistic impulses to provide aid that are independent of considerations of reciprocity, the fact remains that in the real world, many people will not be so motivated. Adopting proposals about distributive justice that assume otherwise may result in outcomes for those aided that are morally less desirable than proposals organized around reciprocity-based considerations[[37]](#footnote-35). Finally, to the extent that one continues to insist that it is simply unjust that welfare schemes should depend in any way on the choices of those aided or on their willingness to behave reciprocally, it is arguable that, in light of the considerations described above, one is abandoning the idea that motivation or rationale for such schemes is based on reciprocity. Instead, one is implicitly drawing on some other distinct conception of justice such as justice as impartiality.

Let me conclude this section by drawing together and elaborating on some of the strands in my discussion. I argued above that Gibbard was right, against Barry, in holding that there is a distinctive motive for justice based on reciprocity and associated with this a way of thinking about justice that is distinct from both JMA and JI. I have also argued, however, that Gibbard is wrong to hold that reciprocity necessarily requires mutual gain. I also think that Barry is right that it is at best dubious that a mutual-gain-based conception of reciprocity provides support for the redistributive policies of modern welfare states or for principles like Rawls’ Difference principle. Barry’s response is to adopt a version of the idea that justice is founded on impartiality, while I have suggested that an alternative possibility is to adopt an different way of thinking about reciprocity which does allow for reciprocity-based considerations to play an important role in justice , without constituting the whole of it. Finally, in anticipation of an argument in the following section, I believe that it also is important to distinguish between the following two claims about the role of reciprocity: (1) considerations based on reciprocity can provide an important motive for supporting or conforming to the redistributive institutions of contemporary welfare states and (2 the difference principle is singled out or uniquely supported over alternative distributive principles by considerations to reciprocity. My argument in this section has been that there is a plausible case for (1). For reasons that will emerge in section X1, I think that (2) is much more dubious.

**XI. Additional Implications and Conclusion**

I turn now, by way of conclusion to two further points. The first revisits the role of norms and institutions in the interactions of reciprocators. The second revisits the normative significance of reciprocity.

There has been considerable interest recently (particularly in connection with issues of international distributive justice) in the role of institutions and more generally rules and norms in connection with obligations of distributive justice.[[38]](#footnote-36) How if at all does the fact that a functioning set of norms and institutions concerning just distribution are in place and generally adhered to affect participant’s obligations? What are individual’s obligations when there are no such functioning rules and institutions in place or to other groups which are governed by very different institutions from our own?

In any account of justice in which reciprocity is important, institutions (and associated norms) will play an important role. We have already noted several reasons why reciprocators, particularly in heterogeneous populations, require rules norms and institutions for successful cooperation. These play the roles of allowing reciprocators to successfully identify and preferentially interact with each other (rather than with non-reciprocators) and they can also provide information about the general level of cooperation in a group, which influences the decisions of reciprocators regarding cooperation. In addition the appropriate norms and institutions can provide for sanctioning of those who behave non-cooperatively. For these reasons alone, people governed by motives of reciprocation will think that their obligations are heavily influenced by the rules and institutions governing cooperation in their group that are actually in place and functioning.

There is, however, an even more fundamental reason why norms and institutions are of central importance to reciprocators. This is that in their absence it is often highly indeterminate what someone who wishes to behave reciprocally should do. Reciprocity involves responding to benefits with benefits, but as we have seen, what counts as a benefit and still more what is an appropriately reciprocal response to the provision of a benefit seems highly dependent on context and on which background norms and institutions. The bare notion of reciprocity does not supply these.

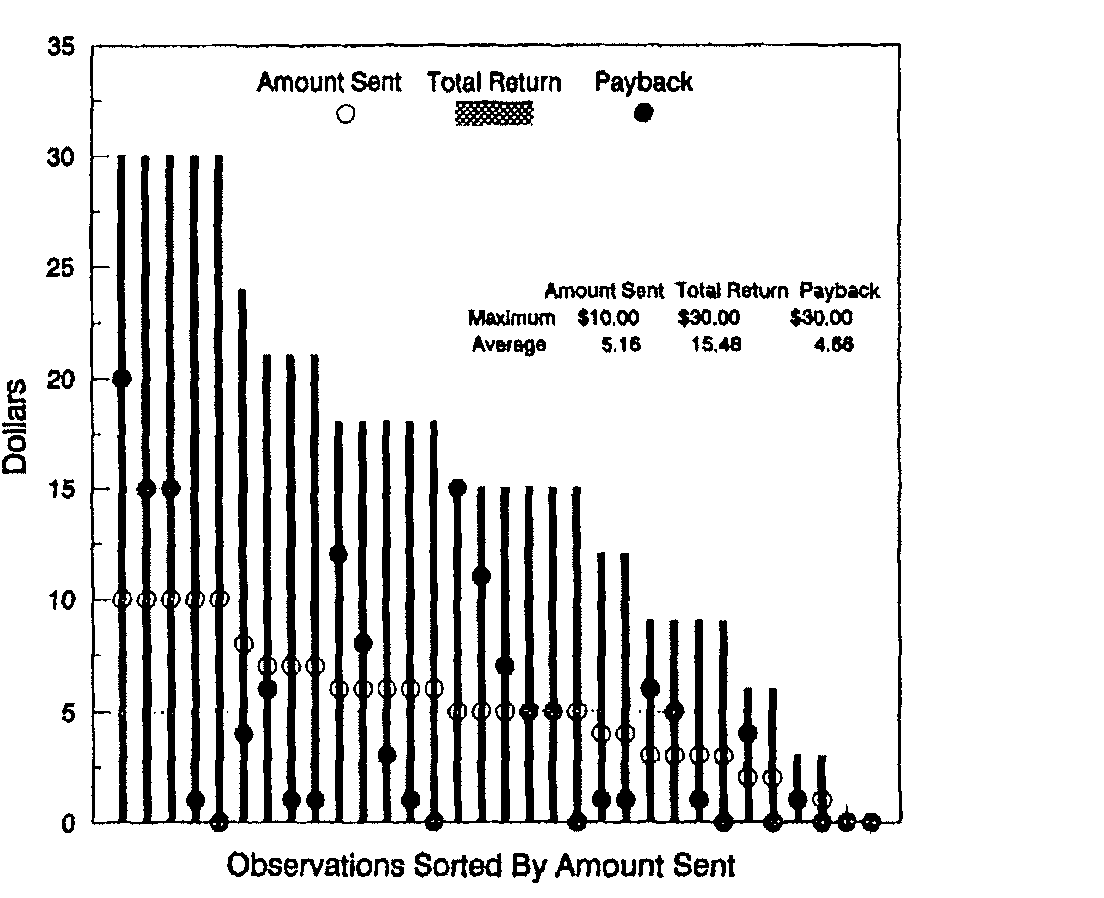
As an illustration, consider the original version of the trust game introduced in Berg, Dickhaut, and McCabe, K. (1995) , which is more general than the trust games described section III. In this game, each player is given a stake of $10.00. Players in the first movers position (trustors) have the option of transferring some or all of this amount to a second player (the trustee). The amount transferred, x, is then tripled by the experimenter, and the trustee can transfer any portion of his total stake (now $10 + 3x) back to the trustor. Presumably reciprocal behavior requires that trustees return some portion of what they receive back to the trustor and trusting behavior (or behavior motivated by the anticipation of reciprocation) on the part of the trustor involves transferring some amount of money. Consider, then a trustor who is a reciprocator and who wishes to behave kindly toward the trustee. Although this is an issue that needs to be investigated empirically, my conjecture is that at least among contemporary American subjects, there is no specific norm governing the amount of money to be sent by trustors. The absence of such a norm might be revealed by, for example, a lack of consensus or considerable variability in this group in response to questions about what trustors “should” send (or, alternatively, a lack of consensus about the amounts that will lead trustees to regard trustors as cooperative). I think that it is even more likely that there is no clear norm about the amounts that should be returned by trustees, given the amount they have received from the trustor . This again should show up in the absence of a consensus regarding parallel questions. For example, is a trustee behaving reciprocally as long she returns more than she is sent? Or does reciprocal behavior instead require a more generous response—e.g., an even split of the amount generated by the investment, so that if the initial endowment is *e* and the amount invested *i*, the trustee returns *3i/2*, with the trustor getting *e-i+3i/2= e+i/2* and the responder *e+3i/2*? Or should the responder return *2i,* so that each player ends up with an equal amount *(e+i/2)*? Or should the responder return even more, since, after all, the risk associated with the interaction is borne entirely by the trustor? In the absence of norms or institutions prescribing appropriate behavior in this sort of situation, we should expect a great deal of variability among reciprocators in the amount sent and a great deal of variability among reciprocators in the proportion of this returned. This is consistent with the observed results from one version of the game reproduced in Figure 3. 

Figure 3

If, for the sake of argument, we regard any trustor who sends a positive amount as a reciprocator, then, in this version of the game 30 out of 32 proposers are conditional cooperators, but the amount they send shows a great deal of variation – amounts sent and number of subjects sending that amount are as follows ($10,5), ($8, 1), ($7, 3), ($6,5), ($5,6), ($4,2), ($3,4), ($2,2), ($1,2). Looking at the response of the trustees, we see that a sizable portion of trustees are free riders or near free riders - of the 28 trustees who received more that $1, 12 returned either $0 or $1. However, in the same group of 28, 11 returned more than they received. Taking the latter to be reciprocators [[39]](#footnote-37) figure 3 shows there is considerable variation in the amount returned both in absolute terms and as a percentage of the amount invested even among this group[[40]](#footnote-38). I take this to reflect, in part, the absence of any clear norm about what the right amount to return even among those who wish to behave reciprocally.

When there are no clear norms providing guidance about what constitutes reciprocal behavior, even well-intended reciprocators may find it difficult to sustain cooperation. In the absence of such guidance, subjects who intend (by their lights) to behave cooperatively will often make choices that strike others as uncooperative (or insufficiently cooperative) prompting non- cooperative responses, failures to coordinate in optimal ways, and so on. This will be particularly likely to the extent that (as we may safely assume is usually the case) even subjects with preferences that incline them toward reciprocation will also be influenced by self- interested preferences which will incline them to interpret cooperative requirements in self-serving ways, so that while they recognize, e.g., that if they have received a positive amount in a trust game, they should provide some return, they will find it easy to persuade themselves that it is appropriate to return less rather than more. By contrast, when a background system of norms and associated institutions is already in place or, if not, when people are able to explicitly agree on these, it will be much clearer to the participants what counts as reciprocal behavior and what sort of behavior signals an interest in reciprocation.

What do these considerations imply about the project of grounding the notion of justice in reciprocity? Consider Gibbard’s question of whether “ Rawls's difference principle [could] be accepted by people whose moral motivations are ones of fair reciprocity”? (1991). In view of the discussion in Section X, I believe that a *conditional* version of the difference principle *could* be accepted by people motivated by reciprocity (where a conditional version is one in which the resources provided to the worse off are not completely independent of their willingness to reciprocate ). This claim should be distinguished, however, from the claim that the motive of reciprocity, by itself, uniquely picks out the difference principle as the only acceptable candidate for a principle to govern the distribution of fruits of cooperation among reciprocators. In view of the discussion immediately above, I think it is unlikely that the notion of reciprocity, in itself, can yield a conception of justice that is as determinate as the notion embodied in Rawls’ theory. As we have seen, the bare notion of reciprocity is in itself, relatively non-specific; it requires filling in by some more determinate set of rules and norms prescribing just what counts as reciprocal behavior in various circumstances, how to deal with violations of reciprocity and so on, where these rules have some other source than just the notion of reciprocity taken in itself. Thus while there may well be other considerations that single out the difference principle or something like it as the most morally appealing principle governing distribution, it is doubtful that we can find these considerations just in the bare notion of reciprocity. Instead, it seems that a range of different principles, (and norms and institutions) are consistent with the motive of reciprocity, in the sense that reciprocity can motivate us to act in accord these and we can think of them as embodying ideals of reciprocity. This is by no means to say, however, that considerations based in reciprocity are completely toothless or unconstraining. Such considerations do narrow down the range of possible theories of justice, for example, by excluding both JMA and JI as completely general theories of justice. Despite the indeterminateness of reciprocity, accounts of justice that violate or ignore considerations or reciprocity will seem normatively unappealing to many people and for this reason may lead to wide-spread non –compliance and other unattractive consequences if we try to realize them.

In this essay I have not attempted a full scale defense of the normative significance of reciprocity aside from arguing that people do seem to value reciprocity in their relations with others and do so for reasons that are not purely instrumental (that is, they do not seem to value reciprocity purely as a means to some further end, involving completely distinct values). This is turn has implications for the stability of various theories of justice. There is, however, an additional point, implicit in some of my remarks above, that can also be made about why reciprocity matters morally. It is that it is a point in favor of reciprocity as a motive, that, in comparison with more unconditionally altruistic motives, it is less exploitable and can help to influence people who are motivated by self-interest to behave in more morally acceptable ways. That is, an important limitation of motives and principles that are unconditionally altruistic is that these can be taken advantage of by people who are willing to behave in exploitive or morally objectionable ways out of self-interested motives—defecting when others cooperate and so on. Someone who acts from reciprocal motives will not be exploitable in this way. Indeed, as we have seen, an unwillingness to behave cooperatively toward those who do not themselves cooperate can influence even self-interested types to behave in more morally acceptable ways[[41]](#footnote-39).

References

Ahn, T., Ostrom, E. and Walker, J. (2003) “Heterogeneous Preferences and Collective Action” *Public Choice* 117: 295-315.

Andreoni, J. (1988) “Why Free Ride? Strategies and Learning in Public Goods Experiments” *Journal of Public Economics* 37: 291-304.

Andreoni, J. and Miller, J. (1993) “Rational Cooperation in the Finitely Repeated Prisoner’s Dilemma: Experimental Evidence.” *Economic Journal*, 103: 570—585.

Arneson, R. (2004) ["Moral Limits on the Demands of Beneficence?",](http://philosophy.ucsd.edu/faculty/rarneson/limitedbenevolence3.pdf) in *The Ethics of Assistance: Morality, Affluence, and the Distant Needy,* ed. by Deen K. Chatterjee (Cambridge: Cambridge University Press, 2004), pp. 33-58.

Barry, B. (1989) *Theories of Justice: A Treatise on Social Justice*, Volume 1. Berkeley, CA: University of California Press.

Berg, J, Dickhaut, J. and McCabe, K. (1995) “Trust, Reciprocity, and Social History” *Games and Economic Behavior* 10: 122-42.

Binmore, K. (1994**)** *Game Theory and the Social Contract*: Volume 1: *Playing Fair*. Cambridge: MIT Press.

Binmore, K. (2007) *Does Game Theory Work? The Bargaining Challenge*. Cambridge, MA: MIT Press.

Binmore, K. (2005), “Economic Man-or Straw Man” *Behavioral and Brain Sciences.* 28:23-24.

Bochet, O., Page, T., Putterman, L., 2006. “Communication and punishment in voluntary contribution experiments”. *Journal of Economic Behavior and Organization* 60, 11-26.

Boyd, R. and Richerson, P. (1988) “The Evolution of Reciprocity in Sizable Groups” Journal *of Theoretical Biology* 132 337-356.

Burlando, R. and Guala, F. (2005) "[Heterogeneous Agents in Public Goods Experiments](http://ideas.repec.org/a/kap/expeco/v8y2005i1p35-54.html)," [*Experimental Economics*](http://ideas.repec.org/s/kap/expeco.html)*,* Springer, vol. 8(1), 5-54,

Camerer, Colin (2003), .*Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

Carpenter, J. and Matthews, P. (2009) “What norms trigger punishment?”

*Experimental Economics* 12: 272–288.

Clark, K. and Sefton, M. “The Sequential Prisoner’s Dilemma: Evidence on Reciprocation” *The Economic Journal* 111: 51-68.

Cox, J., Fredman, D. Gjerstad, S. (2007) “A Tractable Model of Reciprocity and Fairness” Games *and Economic Behavior*. 59, 17-45

Croson, R (2007) “Theories of Committment, Atruism and Reciprocity: Evidence From Linear Public Goods Games” *Economic Inquiry* 45, 199-216.

Dufwenberg, M., Gneezy, U., Güth,  W., Van Damme, E. (2001) “Direct vs Indirect Reciprocity: An Experiment” *Homo Oeconomicus* 18,.

Fehr, E and Schmidt, K. (1999) “A Theory of Fairness, Competition, and Cooperation” *Quarterly Journal of Economics* 114, 817-868.

Fehr, E. and Fischbacher, U. (2004) “Third-party punishment and social norms” *Evolution and Human Behavior*. 25: 63–87.

Fehr, E. and Gachter, S. (2000) “Cooperation and Punishment in Public Goods Experiments” American Economic Review 90, 980-994

Fehr, E., & Fischbacher, U. (2003). “The nature of human altruism”. *Nature*, 425, 785–791.

Fong, C. Bowles, S. and Gintis, H. (2005) “Reciprocity and the Welfare State” in Gintis, Bowles, Boyd and Fehr (eds.) *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*. Cambridge, MA: MIT Press.

Freeman, S. (2007) *Rawls.* London*:* Routledge.

Gibbard, A. (1991) “Constructing Justice” *Philosophy and Public Affairs*, 20, 3, 264-279.

Gintis, H., Bowles, S. Boyd, R. and Fehr, E. (2005) *Moral Sentiments and Material Interests: The Foundation of Cooperation in Economic Lif*e. Cambridge, MA: MIT Press.

Gunnthorsdttir, A. McCabe, K. and Houser, D. (2007) “Disposition, History, and Public Goods Experiments” [*Journal of Economic Behavior & Organization*](http://econpapers.repec.org/article/eeejeborg/), 62, 304-315.

Henrich, Joseph, Boyd, Robert, Bowles, Samuel, Camerer, Colin, Fehr, Ernst and Gintis, Herbert (eds. 2004), *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies.* Oxford: Oxford University Press.

Kaplan, H. and Gurven, 2005 “The Natural History of Food Sharing and Cooperation: A Review and Multi-Individual Approach to the Negotiation of Norms” in Gintis, Bowles, Boyd and Fehr (eds) *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*. Cambridge, MA:MIT Press.

Kaplan, H., Hill, J., Lancaster, J. and Hurtado, A.. 2000. “A Theory of Human Life History Evolution: Diet, Intelligence, and Longevity”. *Evolutionary Anthropology.* 9, 156−185.

Kavka, G. (1986) *Hobbesian Moral and Political Theory*. Princeton, NJ: Princeton Univeristy Press.

McCabe, K, Rigdon, M. and Smith, V. (2003) “Positive Reciprocity and Intentions in Trust Games,” *Journal of Economic Behavior and Organizations*, (52) 2003, pp. 267-275.

Murphy L. (2000) *Moral Demands in Nonideal Theory*. Oxford: Oxford University Press.

Murphy, L. (1993) “ The Demands of Beneficence” *Philosophy and Public Affairs* 22: 267-292.

Ostrom, E. (1990) *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Page, T., Putterman, L. Unel, B. (2005) “Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency” *The Economic Journal* 115: 1032- 53.

Pettit, P. (1995) “The Cunning of Trust” *Philosophy and Public Affairs* 24: 202-25.

Rabin, M. (1993) “Incorporating Fairness Into Game Theory and Economics” *American Economic Review* 83: 1281-1302.

Rawls, J. (1971) *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Rawls, J. (1993) *Political Liberalism.* New York: Columbia University Press.

Samuelson, L (2005) “Foundations of Human Sociality: A Review Essay” *Journal of Economic Literature* 43: 488-497.

Sangiovani, A. 2007. “Global Justice, Reciprocity, and the State”. *Philosophy & Public Affairs* 35: 2-39.

Silk, J.B. 2003. “Cooperation without counting: the puzzle of friendship”. In *The Genetic and Cultural Evolution of Cooperation* (P. Hammerstein, ed.) Cambridge, MA: MIT Press.

1. I will follow these writers and many other philosophers in thinking of the purview of theories of justice in a very expansive way, so that any set of rules or principles prescribing how (or on what basis) benefits and burdens associated with cooperative schemes are to be distributed, or prescribing obligations to provide aid or share resources in circumstances in which there is actual or potential interaction with others can be viewed as drawing on a theory of justice. Similarly, I will regard recommendations about the institutions that best implement these ideas as potentially falling within the purview of a conception or theory of justice. [↑](#footnote-ref--1)
2. This departs from Barry’s own characterization of JMA in terms of **"**the constraints on themselves that rational self-interested people would agree to as the minimum price that has to be paid in order to obtain the cooperation of others "-- -- a conception that he associates with Gauthier. However, Barry also notes that the constraints self -interested persons would “agree to” needn’t be constraints that they would actually follow, if it would not be in their self-interest to do so. Barry takes this to show that the motive of self-interest need not motivate people to conform to the constraints associated with justice as mutual advantage, and concludes that this an objection to JMA, as understood by Gauthier, which claims to rely on such a motive. While I agree with this criticism, I think that Gauthier’s conception of JMA is highly idiosyncratic and that it is at least equally natural to take the core commitment of JMA to be that people always ought (it would be “rational” for them ) to do what best advances their self-interest. On this conception, Gauthier is simply mistaken about what JMA implies, since Gauthier’s conception has people keeping agreements and commitments that it not be rational for them to keep, given that what is rational for each person is to advance his or her self-interest. [↑](#footnote-ref-0)
3. The notion of reciprocity that I describe is closely related to the notion of “strong reciprocity” employed by a number of economists. According to Fehr et al ( 2002), “[a] person is a strong reciprocator if she is willing (i) to sacrifice resources to be kind to those who are being kind (= strong positive reciprocity) and (ii) to sacrifice resources to punish those who are being unkind (= strong negative reciprocity)”. My understanding of reciprocity adds the idea that reciprocators also may be motivated by the *prospect* of reciprocation rather than just by a desire to respond to kindness with kindness. I also make explicit an idea that seems to be implicit in Fehr et al’s characterization: reciprocators need not be motivated by the expectation that they will experience a net gain from a reciprocal interaction, in the sense that their material benefits exceed their costs—see section IX. [↑](#footnote-ref-1)
4. See, for example, Binmore (1994) for discussion. [↑](#footnote-ref-2)
5. It is worth remarking that it is consistent with the existence of reciprocators, so characterized, that reciprocators often or generally do better for themselves (in self-interested terms), in ecologically realistic circumstances, than those who do not act reciprocally—“do better” when the full range of benefits they receive because of their reciprocal behavior, including various sorts of indirect benefits and reputation effects are taken into account. (I take no stand here on whether or not this is the case). According to the characterization I adopt, what makes someone a reciprocator has to do the motives with which that person acts rather than the consequences of acting on those motives. That those who act for reciprocal motives often do well for themselves (in terms of self-interest) does not mean that their motives for action are exclusively self-interested or that when motives of self-interest and reciprocity conflict, they will be guided only by the former. Relatedly, I do not deny that the recommendations of self-interest and reciprocity may often coincide or that both motives may be present when people act. [↑](#footnote-ref-3)
6. The original version of the trust game, due to Berg et al. (1995), is described in more detail in Section XI. In the original version, subjects (trustors) can choose to send any positive amount from a stake of $10 to a second party (trustees). The amount sent is then tripled by the experimenter and the trustee can send whatever amount he or she wishes back to trustor. Most trustors send positive amounts and a significant number of trustees return more than they receive, a result which is interpreted by Berg et al as suggesting that “ reciprocity exists as a basic element of human behavior” (1995, p. 122) [↑](#footnote-ref-4)
7. More specifically, we think of the general category of reciprocator as encompassing a range of possibilities, depending upon exactly how the reciprocator conditions his cooperative behavior on the behavior of others or his expectations about others’ behavior. Some reciprocators may initiate cooperation as a default; they begin interactions with others by cooperating unless they have specific reasons to expect that they are interacting with non-cooperators and they switch to uncooperative behavior only if they encounter it in others. Other reciprocators may be more cautious, in the absence of information that others will cooperate, they may initially fail to cooperate (or not interact at all, if that is an option) and only begin cooperating if they first benefit from a sufficient amount of co-operation by other subjects. Moreover, the thresholds levels of cooperation or non-cooperation by others that prompts such behavior may vary across subjects; some default reciprocators may continue cooperating in the face of a few instances of non cooperative behavior, withdrawing cooperation only as these become frequent. Other subjects may have lower thresholds and begin behaving uncooperatively when they encounter only a few instances (or a single instance) of uncooperative behavior. Some previously cooperative people may begin defecting indiscriminately (even against those with whom they are interacting for the first time) whenever they have experienced enough defections by others, judging that they need not cooperate if the probability that others will cooperate is sufficiently low and that their past experience is good evidence for this probability. And so on for many other possibilities.

   [↑](#footnote-ref-5)
8. To see the need for this restriction consider the proposal that whenever A helps B at a material cost to himself, A derives non-material benefits of some kind from helping B (since otherwise A would not help); hence that A as well as B “benefits” from the interaction and hence the interaction counts as reciprocal. I take this to trivialize the notion of benefit (and reciprocation). This proposal counts any case in which A voluntarily gives a gift to B as a case in which A and B both benefit and as a case of reciprocation, even if B does nothing affecting A in response to A’s gift. Genuine reciprocation requires an intentional response or requital on B’s part which provides some benefit to A. [↑](#footnote-ref-6)
9. Put slightly differently, people involved in reciprocal interactions often seem to care not just about the benefits they receive from those with whom they interact, but about how much it costs their partners to provide those benefits. If A fails to provide a benefit which would be very expensive to provide (he chooses 8,2 in a ultimatum game when 2, 8 is the only alternative, this will be interpreted quite differently by his partner than if he could have provided 5,5 but chooses 8,2 instead. This is another respect in which reciprocal interactions often seem to involve more than purely self-interested motives, since, at least on a naïve analysis, there is no obvious reason why self-interested interactors should care about the costs to their partner of providing benefits, in addition to the magnitude of the benefits themselves. [↑](#footnote-ref-7)
10. For descriptive theories of behavior incorporating this idea in contexts involving “social preferences” or a concern with ‘fairness”, see, e.g. Rabin, 1993 and Cox, 2007. [↑](#footnote-ref-8)
11. Again it may very well be that the causal explanation (or part of the explanation) for why practices of indirect reciprocity exist is that those who exhibit such behavior often receive reputational or indirect benefits of various kinds. For example, perhaps Cs who benefit As in response to their having benefitted Bs acquire valuable reputations for generosity (indeed, reputations for *discriminating* generosity, since they reward the worthy.) These Cs may also be benefitted in response to their generosity by the indirect reciprocity of fourth party Ds. Similarly, Bs who reciprocate to Cs when benefitted by As may acquire good reputations. Nonetheless it does not follow that the motivation for engaging in acts of indirect reciprocation is always self -interest or that people will not engage in such acts when it is not in their self interest to do so. [↑](#footnote-ref-9)
12. As noted above, reciprocal motives seem, as an empirical matter, to be self-reinforcing in this sense: B’s reciprocation of A’s kindness will encourage A to continue to behave kindly toward B. By contrast, when some are disadvantaged in order to maximize aggregate utility it seems unlikely that they will be motivated to act in accord with utilitarian principles. [↑](#footnote-ref-10)
13. A similar conclusion holds if we think that one task of moral and political philosophy is to produce some sort of rational reconstruction of the judgments that people make or if we think that such judgments should serve as defeasible constraints on normative theory, perhaps via some process of reflective equilibrium. Assuming that such people’s non- verbal behavior, as well as what they say, is often evidence for the normative judgments they endorse, the results of empirical studies of behavior of the sort described below are relevant to normative theory via this route as well. [↑](#footnote-ref-11)
14. This is the pay-off matrix used in Andreoni and Miller, 1993. [↑](#footnote-ref-12)
15. Broadly similar results are reported by Kiyonari et al. (2000) who found that, in one shot PDs with Japanese university students, 38 percent engaged in mutual cooperation. [↑](#footnote-ref-13)
16. Binmore (e.g. 2007) and others have suggested several alternative hypotheses which would explain cooperation in a one shot PD even among purely self-interested players. For example, the players may be “confused” and mistakenly think that cooperation will best promote their self-interest. [↑](#footnote-ref-14)
17. Also relevant in this connection are additional results reported in Kiyonari et al. (2000) These experimenters ran three different versions of a prisoner’s dilemma. The ﬁrst was a standard “simultaneous” Prisoner’s Dilemma and the second was a sequential version of the game in which the second mover was told that the first player had already chosen to cooperate. The third was a “ﬁrst- player” treatment in which the subject was told that his decision to cooperate or defect would be made known to the second player before the latter made his own choice. The experimenters found that 38% of the subjects cooperated in the simultaneous treatment, 62% cooperated in the second player treatment, and 59% cooperated in the ﬁrst-player treatment. This suggests that many subjects were reciprocators or conditional cooperators when in the second mover’s position and were also expected that their partners would cooperate when the partners knew that they would not be victimized by defection. The contrasting figure for cooperation in the simultaneous move version of the game (38%) suggests that a fair amount of non-cooperation is due to “fear” (to lack of confidence that one’s opponent will cooperate), rather than “greed” ( a desire to exploit one’s opponent). [↑](#footnote-ref-15)
18. They also suggest that the expression of reciprocal motives is self-reinforcing: people respond to actions that accord with positive reciprocity by continuing to positively reciprocate. They respond to non-reciprocal behavior by failing to act cooperatively, even toward third parties [↑](#footnote-ref-16)
19. In addition to what is described below, there are other kinds of evidence which I lack the space to describe in detail that seem supportive of the hypothesis that some of the participants in public goods games are reciprocators, rather than being purely self-interested. For example, allowing pre-play communication (and especially the making of promises) among the players boosts contributions in public goods games, even when agreements that are made are unenforceable. (cf. Sally, 1995, Bochet, Page and Putterman, 2006) It is hard to see why this factor should make a difference to purely self-interested players since any communication will just be “cheap talk”; on the other hand communication can play a role in solving coordination problems among players who wish to cooperate, but only conditional on others cooperating as well. [↑](#footnote-ref-17)
20. There are a number of other experiments reporting results that are consonant with those reported by Page et al., including Gunnthorsdttir, McCabe, and Houser (2000), and Burlando and Guala (2005 ) [↑](#footnote-ref-18)
21. Such preferential grouping not only allows conditional cooperators to interact preferentially with each other. It also provides an incentive for self-interested players to mimic conditional co-operators (at least until the final round of play) – that is, to play as if they are cooperators – and in this way to obtain the benefits of interacting with conditional cooperators, who would otherwise exclude them. [↑](#footnote-ref-19)
22. One way of making this issue vivid is to ask about the assumptions that the parties in the original position make about compliance. Rawls seems to assume that the parties choose the principles of justice under the assumption that there will be nearly full compliance. But if this means that they assume that no one will behave as a free-rider in connection with public goods or that subjects will not respond to non-cooperative behavior by withdrawing cooperation themselves, this assumption seems empirically unwarranted. If the parties are informed about basic results in social science, why should not they instead assume that the principles that they choose will govern individuals who are motivationally heterogeneous in the manner described above? [↑](#footnote-ref-20)
23. In addition, as already explained, and as discussed in more detail below, I do not agree with Buchanan that relations of reciprocity can exist only among “net contributors”. However, I will put this consideration aside in what follows. [↑](#footnote-ref-21)
24. Dictator games in which one subject (the dictator) unilaterally chooses an allocation between himself and a second subject, with the latter having no say in this allocation are the simplest example. In such cases, there is no opportunity for reciprocal interaction but dictators often allocate non-trivial amounts to the other subject. For example, on average U.S. subjects allocate 20 percent to the other player and the median allocation is around 40-50%. Dictators thus seem to be influenced by other regarding but non-reciprocal motives. Other studies have explored games which are “dictator-like” in the sense that a subject chooses a division of a resources between herself and a several others, where this choice is neither preceded by or followed by any act of reciprocation. In one well- known study (Charnass and Rabin, 2002), many subjects seem to exhibit preferences that involve a weighted average of self-interested concerns and other- regarding concerns, including“utilitarian” concerns with efficiency (maximizing total pay-off) and “Rawlsian or maximin” concerns with increasing the pay-off the worse off group. This suggests, reassuringly, that the criteria to which familiar philosophical theories of justice appeal are indeed at work in people’s choices but that most people are not guided by a single such criterion to the exclusion of all others. It also suggests, however, that when opportunities for reciprocation are present, these same subjects are also influenced by reciprocal motives which influence their willingness to act on other-regarding motives: when, for example, X has the opportunity to sacrifice to help others, but fails to do so, many other subjects will no longer be willing to share resources with X.

    [↑](#footnote-ref-22)
25. For example, in the ultimatum games discussed in Section III, failure to make a generous offer to another in the constrained choice case in which it is extremely costly or impossible to do so provokes much less hostility than an ungenerous offer in an ordinary ultimatum game in which such an offer is possible and not unusually costly. If we think of a generous offer when possible as “reciprocal” behavior (in the sense that it anticipates positive reciprocation in the form of agreement of the other party to the offer) and someone for whom it is impossible for very costly to make such a generous offer as “incapable of reciprocation”, this suggests that people respond much more harshly to those who are capable of reciprocation but do not reciprocate than to those who are incapable of reciprocation. Similarly although the relevant experiment has not to my knowledge been done, one would expect that in a public goods game with a punishment option, non-contributors who are capable of contributing would be much more likely to be punished than non-contributors, who because of the rules of the game, are unable to contribute or for whom contribution would be extremely costly. Results like this support a picture according to which there is a distinctive domain to which reciprocity considerations apply—those capable of reciprocating. Those who do not fall into this category are regarded as subject to different rules. [↑](#footnote-ref-23)
26. For that matter, JR does not exclude the possibility of obligations to aid those who are capable of reciprocation and have previously failed to reciprocate, although in such cases the obligation to aid will of course not be based on considerations of reciprocity. [↑](#footnote-ref-24)
27. Evidence from experimental games allowing for “punishment” of uncooperative players suggests that a similar point holds for negative reciprocity; many subjects desire that they themselves punish non-cooperators, and do not just have desires that non-cooperators be punished by someone or other or that cooperators suffer some cost. Moreover, many subjects behave as though they hold a non-instrumental or retributivist theory of punishment; that is, as though they value punishment for non-cooperators as desirable in itself and not just for the benefits it may bring in deterring non-cooperators in the future, since they punish in circumstances in which no such deterrent effects are possible. [↑](#footnote-ref-25)
28. These are not the only possibilities. Some subjects may be altruists rather than reciprocators but care whether they or some one else are the ones who confer benefits, perhaps because they get a “warm glow” from their altruistic actions that they wouldn’t get if others were to confer the benefits. Other altruists may not care about this. My point is that it is built into the notion of reciprocity that reciprocators care about who is providing benefits to (or imposing costs on) whom and also how these benefits and costs are connected to previous behavior; not so for a more general notion of altruism. [↑](#footnote-ref-26)
29. Suppose that I drive you to the airport and when I ask you to do the same for me on some later occasion, you decline but offer to compensate me for my earlier expenses or to provide me with taxi fare for my trip. In the first case, I may come out even when my overall gain is evaluated in material terms; in the latter I receive a benefit (a ride to the airport) which is equal in value to the one that I provided you. Needless to say, in assessing your behavior and whether it is appropriately reciprocal I am unlikely to focus just on these facts about the magnitudes of the material benefits involved. I will also care about the intentions and attitudes expressed by your behavior, and this will complicate (or render indeterminate) any calculation about whether I have “gained” overall from the interaction. [↑](#footnote-ref-27)
30. For purposes of comparison, chimps and other primates do not engage in anything like this pattern of food-sharing. [↑](#footnote-ref-28)
31. A closely related possibility is that reciprocal behavior in the absence of mutual gain plays a signaling role of some kind that is indirectly beneficial to the participants independently of any direct material gain. Perhaps, for example, some one who contributes more to reciprocal relationship than he receives signals or advertises his generosity and his suitability as partner for future cooperation to third parties and indirectly benefits from this. Perhaps some one who reciprocates a smaller benefit than the benefit he receives (rather than reciprocating nothing at all) advertises that he is grateful, that he appreciates the value of reciprocation and so on, thus appearing to be a more attractive partner for future cooperation than some one who is a complete non-reciprocator. I take no stand on whether such hypotheses are empirically defensible; the point is that there are many possible mechanisms that might be invoked to explain why forms of reciprocity that do not involve mutual gain are operative. [↑](#footnote-ref-29)
32. Another possible illustration of a reciprocal interaction that does not involve mutual gain or an exact quid pro quo is provided by the use of small gift items to encourage charitable donations. In response to a donation, many charities will provide an item of much smaller value to the donor—e.g., a mug or a ticket to a raffle. This practice is prima-facie puzzling: the donor comes out behind when this exchange is evaluated in material terms, and if the donor is willing to donate $n in exchange for a mug that she values at $x, why isn’t it simpler for the donor just to donate $n-x? The answer may be that the exchange alters the interaction so that it is no longer viewed by the donor as a pure act of unreciprocated altruism; instead it is a now viewed as a reciprocal interaction, and motives associated with reciprocation, which are often stronger than those associated with pure altruism, are awakened, leading to a higher level of donation. [↑](#footnote-ref-30)
33. I conjecture that requirements like (2) mutual gain and (3) equal mutual gain are most likely to apply to reciprocal relationships in circumstances in which the parties have roughly similar abilities and resources or face roughly equal costs in providing a given level of benefits to one another. In such “symmetrical” circumstances, the parties will be much more likely to think that there is no obvious reason why one party should experience a positive benefit when the other does not. (Perhaps overgeneralization from this particular case accounts for some of the appeal of (2) and (3)). By contrast, when there is an obvious explanation/justification for why one party experiences an overall material gain from the interaction and the other does not—e.g. one party has a much greater ability to provide a benefit or can provide it at a much lower cost—insistence on (2) and (3) is less likely.) [↑](#footnote-ref-31)
34. Gibbard attributes this argument to Rawls, as does Freeman (2007). [↑](#footnote-ref-32)
35. I recognize that talking of “my” apples and your “stealing” them, may seem question-begging in this context, since it presupposes a system of property rights and other institutions of the sort that Rawls’ framework is introduced to justify. On the other hand, it is hard to see how one can appeal to reciprocity at all, without presupposing some notion of “benefit”: certainly Gibbard and Rawls appeal to such a notion in claiming that distributive justice (or aspects of it) can be grounded in a notion of reciprocity connected to mutual gain. It thus seems reasonable to ask whether foreswearing of violence constitutes provision of a “benefit” in the relevant sense. [↑](#footnote-ref-33)
36. As noted in Section III, by “strong reciprocity” these writers mean what I have been calling “reciprocity”. [↑](#footnote-ref-34)
37. It is worth noting that the situation under discussion here is interestingly different from the situation assumed in the usual discussions of ideal theory and partial compliance. In the usual discussions it is often assumed that partial compliance is the result of deviations from what justice requires for reasons of self-interest. But when people fail *because they are reciprocators*, to follow norms mandating unconditional aid, their deviations are not due (just) to self-interest, but rather to their holding an alternative view of what others are owed that is recognizably a moral view, even if some regard it as a mistaken view. So in this case, it is harder to argue that if we adopt principles of justice or design institutions that make allowance for the fact that people are reciprocators we are allowing people’s willingness to behave in bad or selfish ways to compromise our conception of justice. [↑](#footnote-ref-35)
38. See, e.g., Sangiovani, 2007. [↑](#footnote-ref-36)
39. Obviously this is a very permissive stipulation about who counts as a conditional cooperator. My guess is that many trustors will not regard trustees who return only slightly more than they receive as kind or cooperative. [↑](#footnote-ref-37)
40. Out of the 11 trustees who returned more than invested, 6 returned *2i* or more, with the remainder returning an amount intermediate between *i* and *2i* that bore no obvious relationship to *i*. [↑](#footnote-ref-38)
41. This is another example of a normative consideration whose significance will be overlooked if we focus just on conditions of full compliance. [↑](#footnote-ref-39)