Economics and Philosophy, **22** (2006) 1–18
doi:10.1017/S0266267105000684

WHEN ALTRUISM LOWERS TOTAL WELFARE

Kenneth S. Corts*
University of Toronto

Ethical theories grounded in utilitarianism suggest that social welfare is improved when agents seek to maximize others' welfare in addition to their own (i.e., are altruistic). However, I use a simple game-theoretic model to demonstrate two shortcomings of this argument. First, altruistic preferences can generate coordination problems where none exist for selfish agents. Second, when agents care somewhat about others' utility but weight their own more highly, total social welfare may be lower than with selfish agents even in the absence of coordination problems.

1. INTRODUCTION

Ethical theories grounded in utilitarianism are premised on simple idea: social welfare would be improved if everyone acted to promote social welfare rather than his or her own self-interest (i.e., if everyone were altruistic). This apparently obvious conclusion turns out to be less certain than it seems. In this paper I apply game-theoretic reasoning to demonstrate a number of weak links in the connection between individually altruistic agents and the realization of the utilitarian goal of social welfare maximization. Specifically, I demonstrate first that coordination problems may be more severe for altruists than for selfish agents, leading altruists collectively to fare worse than selfish players in some circumstances. I then show that, assuming away coordination problems, altruists do in fact succeed in maximizing social welfare, but only if every agent weights others' utility on par with his or her own.

*105 St. George St., Toronto, Ontario M5S 3E6; kenneth.corts@rotman.utoronto.ca. I thank Adam Brandenburger and Mara Lederman for helpful comments.

Agents who are partially altruistic may fare worse than selfish agents in some circumstances even in the absence of coordination problems.

The primary mechanism by which altruism may lower social welfare is the introduction of coordination problems. Consider the following example. Two students are considering whether to study for an exam. If both study, both will get As (valued at 2). If neither studies, both will get Cs (valued at 0). If one studies and the other does not, the studious one serves to "bust the curve," obtaining an A while the other fails the course (valued at -4). Selfish students have a dominant strategy to study since studying yields a higher payoff than not studying, regardless of the other student's strategy (2 is greater than either 0 or -4). Play of these dominant strategies earns each selfish student an A (and a payoff of 2).

Now imagine that the students are altruistic, with each maximizing the average of his or her own payoff and the other student's payoff. It remains true that each student wants to study if the other studies, since both getting As (an average value of 2) is better than getting a C while the other gets an A (an average value of $0.5^*2 + 0.5^*0 = 1$). However, the fear of busting the curve and causing the other student to fail the course now looms large. In fact, if the other student does not study, the payoff associated with studying and busting the curve ($0.5^*2 + 0.5^*(-4) = -1$) is actually worse than getting a C as a result of not studying ($0.5^*0 + 0.5^*0 = 0$). The fact that each would like not to study (in order to avoid being a curve-buster) if the other is not studying means that a coordination problem has been introduced. It is an equilbrium of this game for neither player to study, resulting in payoffs of 0 for each player, which is worse than the payoffs of the original game. Here, concern for the other player (going out of one's way not to bust the curve) may lead to an inferior outcome for both players.

Philosophers including Regan (1980) and Sobel (1985) have argued that coordination issues present a real and pervasive problem for utilitarian ethical theory. However, arguments in this literature emphasize that altruism does not solve existing coordination problems. My contribution with respect to this argument lies in showing that altruism not only fails to solve existing coordination problems, but actually creates new coordination problems that selfish agents would not encounter. This is the essence of the above example, as the students had dominant strategies and faced no coordination problem until they became altruistic.

More recent work on altruism by economists (Esher, Samuelson, and Shaked 1998) and philosophers and biologists (Sober and Wilson 1998) focuses on the viability of altruistic preferences when subjected to evolutionary processes. This work operates from the assumption that altruists fare well against other altruists, and that groups of altruists collectively fare well, but that altruists do not fare well when their population is infiltrated by selfish agents. My focus is quite different,

https://doi.org/10.1017/S0266267105000684 Published online by Cambridge University Press

as I take the existence of altruistic agents as given and ask whether in fact altruistic agents fare well even when facing only other altruists. The analysis will show that they do not necessarily do so, if one takes coordination problems seriously.

In much of the analysis I employ the solution concept of rationalizability (Bernheim (1984) and Pearce (1984)), which allows coordination problems to flourish.[1] Rationalizable strategies are those that are consistent with each player's rationality, with each player's belief in the other player's rationality, with each player's belief in that belief, and so on. This allows a broader range of behavior than does Nash equilibrium because it does not require that each player is correct in his or her prediction of the other player's strategy. In particular, in the above example with altruistic players it is rationalizable for one student to study and the other not to. This leads to the curve-busting scenario and yields the lowest possible payoff of -1 for each of the altruistic students. This is not a Nash equilibrium outcome since each of the students would improve his or her payoff by changing strategies. However, a rational student might well not study when the other student was studying if the first student (mistakenly) thought that the second student was not. And it is rational for the first student to think that the second student might not study, since the second student could easily believe that the first student was not studying (to which not studying is the payoff-maximizing response). In many simple games with multiple Nash equilibria, the essence of the rationalizability solution concept is that "anything goes."[2]

Regan (1980) makes an important distinction in demonstrating that although utilitarian behavior does not *ensure* the maximization of social welfare (due to coordination problems), it is *consistent* with it. To put it another way, the best-case scenario in a utilitarian world does achieve the maximization of social welfare, though the worst-case scenario need not. I employ a second solution concept (efficient Nash equilibrium) that rules out all coordination problems to demonstrate this result. Specifically, I show that the efficient Nash equilibrium of any game with (fully) altruistic players necessarily maximizes social welfare. For example, in the above example with altruistic students there are two Nash equilibria – one in which both students study and one in which neither studies – and the

[1] In the above example with altruistic players, a coordination problem arises even under Nash equilibrium, as described above. However, rationalizability allows an even broader range of outcomes.

[2] This is true, for example, for two-by-two games (two player games with two strategies for each player) with no ties in payoffs. To see this, note that (absent ties) the only way to have multiple equilibria is to have two equilibrium outcomes along one of the diagonals of the matrix. This implies that each strategy for both players is a best-response to *some* strategy of the other player, which implies that all strategies are iteratively undominated (and therefore rationalizable).

https://doi.org/10.1017/S0266267105000684 Published online by Cambridge University Press

efficient Nash equilibrium (the one in which both students study and get payoffs of 2) does in fact maximize social welfare. I then show that this does *not* hold when agents are only partially altruistic (in the sense that they weight others' utility less than their own) and, moreover, that partially altruistic agents necessarily end up worse off than selfish agents in some games.

Section 2 sets up the basic game-theoretic framework I will employ, which includes a typology of all symmetric, simultaneous two-by-two games. Section 3 evaluates outcomes of these games according to the solution concept of rationalizability, which allows maximal coordination problems. The main result in this section is that altruistic players experience coordination problems in some games where selfish players experience none, which can make altruistic players collectively worse off than selfish agents. Moreover, I show that altruistic players can be collectively worse off than selfish players even in games where selfish players face a Prisoner's Dilemma. Section 4 evaluates outcomes according to the solution concept of efficient Nash equilibrium, which rules out coordination problems. Here, altruistic players always do better than selfish players if they are fully altruistic or if the game is symmetric. However, I show that in asymmetric games they may do worse if they only partially take into account others' welfare, even though coordination problems have been ruled out. Section 5 presents simple examples of each of these results, and section 6 concludes.

2. THE BASIC GAME

Most of the analysis focuses on an exhaustive characterization of the set of symmetric two-by-two games. By symmetric, I mean that payoffs conform to those in Figure 1 with at most two transformations. First, rearrange or rename the strategies so that the Pareto-efficient symmetric outcome is in the X, X quadrant. Second, add or subtract a common term to both payoffs in every quadrant as a normalization that yields 0, 0 in the Y, Y quadrant. Note that together these transformations imply that $a > 0$; however, b and c may be positive or negative. The names of the strategies are arbitrary; this notion of symmetry in no way requires that the players' strategies have literally the same names, though it simplifies the exposition to assume so.

Subsequent sections will evaluate outcomes of this game using two different standard solution concepts: rationalizability and efficient Nash equilibrium. For each of these, an understanding of the interdependences of the players' best responses will prove helpful. All symmetric games described by the above matrix can be classified into one of four types, depending on the relationships of a to b and c to 0. When $a < b$ and $c < 0$, then Y is dominant strategy for both players. Both players' playing their dominant strategies yields the inferior symmetric outcome. This is the Prisoner's Dilemma. When $a < b$ and $c > 0$, no player has a dominant

https://doi.org/10.1017/S0266267105000684 Published online by Cambridge University Press

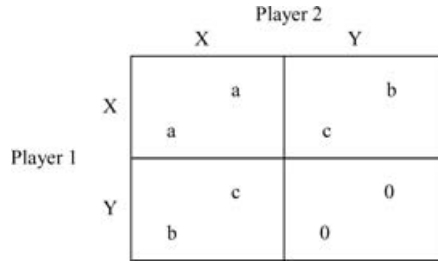

FIGURE 1. A symmetric game.

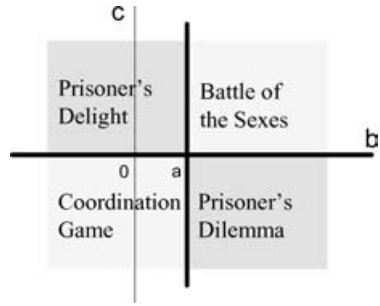

FIGURE 2. A taxonomy of symmetric games.

strategy; each player wants to play Y if the other plays X, but X if the other plays Y. Such games have two (pure strategy) equilibria, which correspond to the quadrants off the main diagonal. In these, the players' payoffs differ, so that which of these equilibria each prefers depends on whether b or c is larger. This is the Battle of the Sexes. When $a > b$ and $c < 0$, no player has a dominant strategy; each player wants to play X if the other plays X, but Y if the other plays Y. Such games have two (pure strategy) equilibria corresponding to the two quadrants on the main diagonal. Since these quadrants yield symmetric outcomes, players both prefer the efficient equilibrium. These are Coordination Games. Finally, when $a > b$ and $c > 0$, X is a dominant strategy for each player. This yields the a, a payoff, which is the efficient symmetric payoff. This kind of game seems not to have a standard name; because of its relationship to the Prisoner's Dilemma (as in that game, both players have dominant strategies; however, playing them now leads to the superior symmetric outcome), I call it the Prisoner's Delight.

For a given value of a, a particular game is determined by a point in $b \times c$ space. Thus, the types of games occupy regions in this space as depicted in Figure 2.

https://doi.org/10.1017/S0266267105000684 Published online by Cambridge University Press

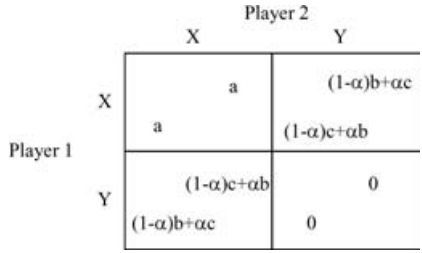

FIGURE 3. A symmetric game with altruistic players.

2.1 Altruistic players

Now assume that the players are altruistic in the sense that they play as if they are maximizing a weighted average of their own payoff and the other player's payoff. This is the standard assumption in the economic modeling of altruistic agents (see, for example, Levine 1998). Specifically, each agent now maximizes the sum of $1-\alpha$ times its own payoff and α times the other player's payoff, where $\alpha \in [0, \frac{1}{2}]$. Note that $\alpha = 0$ is the case of selfish agents and $\alpha = \frac{1}{2}$ is the case of purely altruistic agents who weight all utilities equally.[3] This effectively yields a new game with payoffs as given in Figure 3.

Because the game with $\alpha > 0$ is still a symmetric game in the sense described above, one can apply the same typology of games to the new payoffs to see how the game has changed. The key comparisons are now between $(1-\alpha)b + \alpha c$ and a, and between $(1-\alpha)c + \alpha b$ and 0. Now X is the best response to X when $c < \frac{1}{\alpha}[a - (1-\alpha)b]$. Similarly, X is the best response to Y when $c > -\frac{\alpha}{1-\alpha}b$. These inequalities can be plotted in $b \times c$ space to generate a graph of the regions in which each type of game arises for altruistic players. This is portrayed in Figure 4. The first condition, which determines the best response to X, is labeled BR_X; the second is similarly labeled BR_Y.

2.2 How altruism changes outcomes

The fact that both players maximize weighted average payoffs implies that *total* payoffs do not change within any given quadrant as α changes.[4] No matter how much weight player 1 shifts from its own payoff to the other

[3] This is the usual definition of altruism. It is *not* selflessness ($\alpha = 1$), but rather the maximization of an evenly weighted sum of individual utilities – that is, altruists maximize total social welfare.

[4] I focus the analysis exclusively on total payoffs for two reasons. First, the utilitarian theories that motivate this analysis judge outcomes according to this criterion or, equivalently, total social welfare. Second, unambiguous Pareto comparisons are generally not available. For

https://doi.org/10.1017/S0266267105000684 Published online by Cambridge University Press

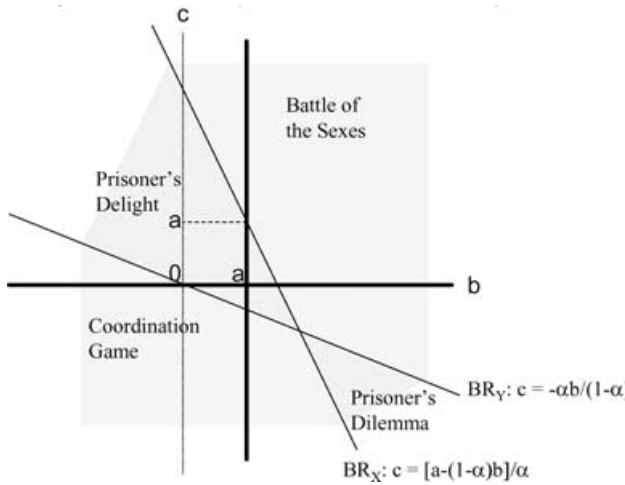

FIGURE 4. The taxonomy in games with altruistic players.

player's, this is exactly offset by the weight that player 2 puts on player 1's payoff. Therefore, regardless of α, total payoffs are always $2a$ in the X, X quadrant; 0 in the Y, Y quadrant; and $b + c$ in the X, Y and Y, X quadrants. As a result, changes in altruistic behavior bring about changes in total payoffs only when they cause the outcome of the game to switch from one quadrant to another (according to some particular solution concept).

In the context of the typology of two-by-two games presented above, this means that total payoffs change only when the game changes from one type of game to another. For example, if the game remains a Prisoner's Dilemma as α increases, then the outcome remains in the Y, Y quadrant and total payoffs remain at 0. However, if the game becomes a Prisoner's Delight, then the outcome shifts to the X, X quadrant and total payoffs increase to $2a$.

To facilitate the analysis of the circumstances under which altruism does change the outcome of the game, Figure 5 overlays the two previous figures. Letters label the regions in which the selfish and altruistic assumptions yield different types of games. These regions will be the focus of the analysis to follow since they represent the cases in which total payoffs may change due to the introduction of altruistic preferences. The dashed and dotted lines are added for purposes of the subsequent analysis. They define the regions in which total payoffs in the asymmetric quadrants (X, Y and Y, X) are better or worse than the symmetric payoffs.

example, note that when the outcome does not change and payoffs are unequal (in the X, Y or Y, X outcomes), one players necessarily benefits at the other's expense when α increases.

https://doi.org/10.1017/S0266267105000684 Published online by Cambridge University Press

All of the labeled regions exist for all $\alpha \in (0, \frac{1}{2})$. As α increases, the BR_Y line rotates (through the origin) clockwise from the b-axis until it coincides with the dotted line when $\alpha = \frac{1}{2}$. The BR_X line rotates (through $b = c = a$) counter-clockwise from the $b = a$ line until it coincides with the dashed line when $\alpha = \frac{1}{2}$. When $\alpha = \frac{1}{2}$, no more Prisoner's Dilemmas remain, and region G extends indefinitely toward the southeast, separating regions C and F.

3. RATIONALIZABLE OUTCOMES

The goal of this section is to demonstrate the prevalence of coordination problems for altruistic players. To this end, I employ the solution concept of rationalizability. An *outcome* is rationalizable if the strategies leading to that outcome are rationalizable, and a *strategy* is rationalizable if it is iteratively not strictly dominated.[5] In the Prisoner's Dilemma and Prisoner's Delight games only the dominant strategy is rationalizable, and there is therefore a unique rationalizable outcome. In the Battle of the Sexes and the Coordination Game both strategies are rationalizable, and therefore all four outcomes are rationalizable.

Intuitively, this is simply a criterion that allows coordination problems to flourish. It says that if neither player can rule out the play of either strategy (because each is a best-response to some strategy the other might play), then any combination of strategies may be realized. Note that there are two types of "coordination problems." There is a lack of coordination in the sense that either of two equilibrium outcomes in, say, a Coordination Game may be played. But there is also a more fundamental lack of coordination in the sense that neither of those equilibria must arise in a particular play of the game. Rationalizability takes seriously the uncertainties inherent in playing games with no dominated strategies.

3.1 Selfish players

To establish the baseline outcomes and payoffs, consider the rationalizable outcomes for the game played by selfish players ($\alpha = 0$). When the game is

[5] Rationalizability is the solution concept that follows from the rationality of players, their mutual belief in each others' rationality, and so on (Bernheim 1984; and Pearce 1984). We will see that this solution concept leaves open the possiblity of coordination problems. Thus, Mackie (1973, p. 291) is incorrect when he asserts that "assuming simply rationality, goodwill, and knowledge both of the causal connections between action-combinations and utility and of each other's rationality, goodwill, and knowledge, [the individuals] will each so act that their combined action will maximize utility." Mackie makes an extended argument about how communication and social norms can resolve coordination problems; however, rationality, altruism, and mutual knowledge thereof do not eliminate them.

https://doi.org/10.1017/S0266267105000684 Published online by Cambridge University Press

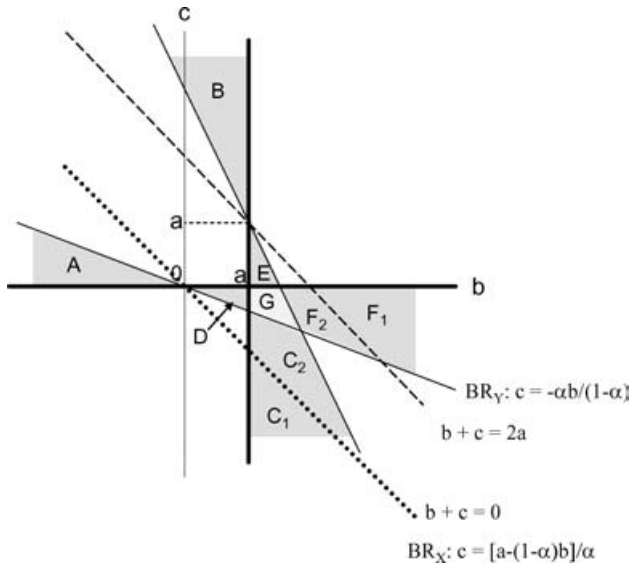

FIGURE 5. When altruism changes the game.

a Battle of the Sexes or a Coordination Game, all combinations of strategies are rationalizable. As a result, total welfare may range from $\min\{0, b+c\}$ to $\max\{2a, b+c\}$. When the game is a Prisoner's Delight, the unique rationalizable outcome involves the play of both dominant strategies, and social welfare is therefore $2a$. When the game is a Prisoner's Dilemma, the unique rationalizable outcome again involves the play of both dominant strategies, and total social welfare is 0.

The total payoffs in each of the seven labeled regions in Figure 5 are recorded in the third column of Table 1. This table shows what type of game prevails in each labeled region and what range of payoffs is associated with the rationalizable outcomes. Here, the entries for the Prisoner's Dilemma and Prisoner's Delight rows are straightforward. Two others require a further observation. Region D yields a coordination game, while region E yields a Battle of the Sexes. Since $2a > b+c > 0$ in both of these regions (i.e., both regions lie above the dotted line in Figure 5, but below the dashed line) and all outcomes are rationalizable, total payoffs range from 0 to $2a$.

3.2 A, B, and C: altruism introduces coordination problems

This subsection and the ones that follow compare the altruistic outcomes to those of the selfish baseline. This analysis is summarized in Table 1, in which the fifth column gives the range of total payoffs achieved in

https://doi.org/10.1017/S0266267105000684 Published online by Cambridge University Press

	Selfish players		Altruistic players		
Region (Fig. 5)	Type of game	Total payoffs	Type of game	Total payoffs	Effect of altruistic players on total social welfare
A	PDlt	2a	CG	[b + c, 2a]	reduces lower bound
B	PDlt	2a	BoS	[0, b + c]	reduces lower bound increases upper bound
C	PDma	0	CG	C1: [b + c, 2a]	reduces lower bound increases upper bound
				C2: [0, 2a]	increases upper bound
D	CG	[0,2a]	PDlt	2a	increases lower bound
E	BoS	[0,2a]	PDlt	2a	increases lower bound
F	PDma	0	BoS	F1: [0, b + c]	increases upper bound
				F2: [0, 2a]	increases upper bound
G	PDma	0	PDlt	2a	increases

TABLE 1. Analysis of rationalizable outcomes.

rationalizable outcomes. In regions A and B, the game begins as Prisoner's Delight, in which the only rationalizable outcome is X, X. With altruistic players ($\alpha > 0$), the game becomes a Coordination Game in region A. Since region A lies below the dotted line ($b + c < 0$), rationalizable total payoffs range from a minimum of $b + c$ to a maximum of $2a$. In the selfish players' outcome of X, X, the total payoff is $2a$, so the introduction of (partially) altruistic agents strictly lowers the lower bound on total payoffs without changing the upper bound.

In region B, the game also begins as a Prisoner's Delight, but it then becomes a Battle of the Sexes. Every outcome is therefore rationalizable in the game with altruistic players. Since region B lies above the dashed line ($b + c > 2a$), the total payoffs range from a minimum of 0 to a maximum of $b + c$, whereas the total payoff for selfish players is $2a$. Again, the introduction of (partially) altruistic players strictly decreases the lower bound on rationalizable payoffs; however, in this case it also increases the upper bound. Note that all other Prisoner's Delights outside regions A and B remain Prisoner's Delights with increases in α; therefore, total payoffs do not change. The results for regions A and B are summarized in the following proposition.

Proposition 1. *Suppose a symmetric game between two selfish players is a Prisoner's Delight. Then making the players more altruistic weakly decreases*

https://doi.org/10.1017/S0266267105000684 Published online by Cambridge University Press

the lower bound on rationalizable total payoffs due to the introduction of new coordination problems.

In region C, the game begins as a Prisoner's Dilemma, in which the only rationalizable outcome is Y, Y. With $\alpha > 0$, the game becomes a Coordination Game, in which every outcome is rationalizable. Rationalizable total payoffs now range from a lower bound of 0 in subregion C2 and a lower bound of $b + c$ in subregion C1 (which lies below the dotted line depicting $b + c = 0$) to an upper bound of $2a$. The upper bound has increased from the 0 total payoff in the unique rationalizable outcome with selfish players. However, the lower bound compares unfavorably in subregion C1.

As bad as the original Prisoner's Dilemma was, the lower bound on total payoffs has become worse for altruistic players in subregion C1. Altruistic players in this region do not have dominant strategies; this introduces uncertainty over the other's strategy. As a result, anything can happen, including outcomes even worse than those experienced in the Prisoner's Dilemma. The analysis of this case is summarized in the following proposition.

Proposition 2. *Suppose a symmetric game has total payoffs minimized in the asymmetric outcomes and is a Prisoner's Dilemma when played by selfish players. Then the lower bound on rationalizable total payoffs weakly decreases as players become more altruistic*.

3.3 D, E, and F: altruism weakly improves social welfare

In regions D, E, and F, the introduction of altruistic players weakly improves social welfare. As this is the least counterintuitive result and the reasoning closely parallels that in the above subsection, I summarize the details only briefly here.

In region D, the game begins as a Coordination Game. In region E, the game begins as a Battle of the Sexes. In both of these kinds of games, all outcomes are rationalizable. Games in these regions become Prisoner's Delights, in which only the best symmetric outcome is rationalizable and total payoffs are therefore $2a$. Since these regions lie wholly below the dashed line defining $2a = b + c$, this unique rationalizable outcome maximizes total welfare. Thus, it is equal to the maximum total welfare achievable in rationalizable outcomes of the original game and the lower bound on total payoffs increases.

In region F, the game begins as a Prisoner's Dilemma, in which only the Y, Y outcome (with payoffs of 0) is rationalizable. It becomes a Battle of the Sexes, in which all outcomes are rationalizable. Since region F lies wholly above the dotted line ($b + c = 0$), the minimum rationalizable payoff is 0. Thus, the change weakly increases total welfare. Whether the upper bound

https://doi.org/10.1017/S0266267105000684 Published online by Cambridge University Press

increases to $2a$ or to $b+c$ depends on whether the game lies in subregion F1 or F2, which are defined by the dashed line representing $b+c=2a$.

3.4 G: altruism necessarily strictly improves social welfare

Region G is the only region in which the introduction of altruistic players necessarily and unambiguously improves total welfare. Here, the game begins as a Prisoner's Dilemma, in which only Y, Y is rationalizable, and becomes a Prisoner's Delight in which only X, X is rationalizable. This strictly increases total welfare from 0 to $2a$. It is perhaps surprising that this strong result on increases in payoffs holds for such an apparently small subset of games.

4. EQUILIBRIUM OUTCOMES

In this section, I use the solution concept of efficient Nash equilibrium in order to eliminate all coordination problems. In a sense, this biases the results in favor of finding that utilitarianism is successful in maximizing total welfare, since coordination problems have been shown above to lead to many inefficient rationalizable outcomes.

4.1 Fully altruistic players

When $\alpha=\frac{1}{2}$, each player acts to maximize total social welfare, defined as the sum of the underlying individual payoffs. In this case, the social welfare-maximizing outcome is always the efficient equilibrium outcome. To see this, note that for any outcome both players have exactly the same payoffs (after accounting for the weighting of underlying payoffs); denote by π^*, π^* the payoffs that maximize total welfare. Since these maximize payoffs over the whole matrix, they must also constitute a best-response to the other player's action that supports this outcome. Since this holds for both players, this is an equilibrium. Since this outcome maximizes total payoffs globally, it clearly must be the efficient equilibrium. This is exactly Regan's (1980) proof that utilitarianism is consistent with maximization of total payoffs though it does not ensure it. There may be other equilibria that are inefficient (consider the game where $a>0>b=c$), but the welfare-maximizing outcome is always the efficient equilibrium, meaning that altruism weakly improves total payoffs when all coordination problems have been assumed away.

4.2 Partially altruistic players in symmetric games

This subsection applies the reasoning of the earlier analysis of Table 1, which applied to rationalizable outcomes, to the solution concept of efficient equilibrium. This is summarized in Table 2. It is easy to see that,

https://doi.org/10.1017/S0266267105000684 Published online by Cambridge University Press

	Selfish players		Altruistic players		
Region (Fig. 5)	Type of game	Total payoffs	Type of game	Total payoffs	Effect of altruistic players on total social welfare
A	PDlt	2a	CG	2a	no change
B	PDlt	2a	BoS	b + c	increases
C	PDma	0	CG	2a	increases
D	CG	2a	PDlt	2a	no change
E	BoS	b + c	PDlt	2a	increases
F	PDma	0	BoS	b + c	increases
G	PDma	0	PDlt	2a	increases

TABLE 2. Efficient equilibrium outcomes.

in all regions, (partially) altruistic players get weakly higher total payoffs in the efficient equilibrium than do selfish players.

In regions A and D, payoffs do not change since Prisoner's Delights and Coordination Games yield the X, X outcome under the solution concept of efficient equilibrium. In region B, Prisoner's Delights become Battles of the Sexes. Thus, total payoffs become $b + c$, compared to $2a$ in the X, X outcome attained by selfish players. This is an improvement in total payoffs since $2a < b + c$ above the dashed line. In regions C and G, Prisoner's Dilemmas become Coordination Games and Prisoner's Delights, respectively. In both cases, efficient equilibrium total payoffs go from 0 to $2a$, which clearly raises total payoffs. In region E, Battles of the Sexes become Prisoner's Delights. Thus, total payoffs become $2a$. This increases total equilibrium payoffs above the $b + c$ attained in the X, Y and Y, X outcomes by selfish players since $2a > b + c$ below the dashed line. In region F, Prisoner's Dilemmas become Battles of the Sexes. Thus, total payoffs become $b + c$, and total payoffs increase since $b + c > 0$ above the dotted line.

4.3 Partially altruistic players in asymmetric games

The previous subsection gives some credence to the view that the only shortcoming of altruism is that it may not solve all coordination problems (the shortcoming emphasized in section 3). Assuming away coordination problems through the choice of solution concept, I showed that in symmetric games introducing even partial altruism always results in weak increases in total payoffs. In this section I introduce an asymmetric game to demonstrate that even when coordination problems are assumed away

https://doi.org/10.1017/S0266267105000684 Published online by Cambridge University Press

(partial) altruism need not increase total payoffs. That is, I show that there are shortcomings that have nothing to do with coordination problems. They do, however, require an asymmetric game.

Full altruism ($\alpha = \frac{1}{2}$) continues to imply that social welfare is maximized in the efficient equilibrium of asymmetric games, as nothing in the proof in subsection 4.1 exploited symmetry. Here, I examine a single example of an asymmetric game, given in Figure 6, with partial altruism ($\alpha < \frac{1}{2}$). In this example, efficient equilibrium and rationalizability both imply the same unique outcome at both levels of α considered (that is, the outcome changes with α, but it changes in the same way under both solution concepts). Thus, coordination problems play no role in the analysis. The example establishes the following fact.

Proposition 3. *Using the solution concept of either rationalizability or efficient equilibrium, total payoffs may fall as players become more altruistic, even in the absence of coordination problems.*

Consider the following game:

		Player 2: x	Player 2: y
Player 1	X	40, 40	0, 16
	Y	56, 36	72, 0

FIGURE 6. An asymmetric game.

Player 1 has a dominant strategy X. Player 2 has a unique best-response x to player 1's dominant strategy. Thus, X, x is both the unique equilibrium and the unique rationalizable outcome.

Now consider the case with partially altruistic players, specifically with $\alpha = \frac{1}{2}$. Recalculating each player's payoffs as $\frac{3}{4}$ (i.e., $1 - \alpha$) times its own payoffs plus $\frac{1}{4}$ times the other player's payoffs yields the game given in Figure 7.

Now player 1 has a dominant strategy Y. Player 2 has a unique best-response y to player 1's dominant strategy. Thus, Y, y is both the unique equilibrium outcome and the unique rationalizable outcome. This yields payoffs of 18 and 54 to the two players, respectively, for a total payoff of 72. This lowers total payoffs compared to the X, x outcome attained by selfish players, which yields a total payoff of 80.

https://doi.org/10.1017/S0266267105000684 Published online by Cambridge University Press

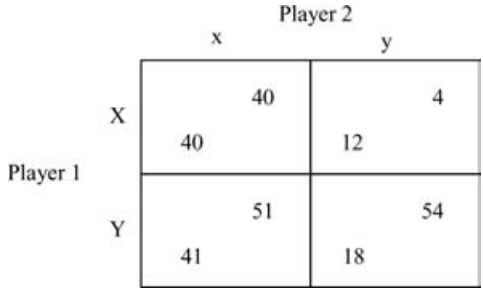

FIGURE 7. The asymmetric game with $\alpha = \frac{1}{4}$.

5. EXAMPLES

In this section I present simple examples to illustrate each of the three propositions. I give numerical payoffs in only the second example, but it should be clear that the logic of each of these three games can be captured in a simple two-by-two game conforming to the assumptions of the above models.

5.1 The O. Henry effect (Proposition 1)

Proposition 1 corresponds closely to the story told in O. Henry's classic fable "The Gift of the Magi" (reprinted in Henry 1984). In this story, Jim secretly sells his prized gold pocketwatch to buy a set of tortoise-shell combs for Della. Della secretly sells her prized long hair to a wigmaker to buy Jim a platinum fob chain for his watch. At the climactic moment in the story, they reveal their gifts and simultaneously discover what each has given up to buy them.

This can be thought of as a two-by-two game in which Jim and Della each have the two strategies of (X) selling their prized possession to buy the gift or (Y) not. If Jim and Della were selfish, each would presumably have a dominant strategy not to part with their prized possessions (Y). This would, by the selfish payoffs, be a Prisoner's Delight, since the outcome in which neither gives up the prized possession is better than the outcome in which both do (that is, play of the dominant strategies leads to the efficient symmetric outcome).

If Jim and Della are altruistic, this is arguably a Battle of the Sexes, in which it is an equilibrium for *either* Jim or Della to make the sacrifice for the other's gift, but *not* both (this corresponds to region B in the figure of section 3, where Prisoner's Delights become Battles of the Sexes). The realized outcome in O. Henry's story demonstrates the problem with Battles of the Sexes: since neither knew whether the other would make the sacrifice, it was rational for each of them to do so, even if they both

https://doi.org/10.1017/S0266267105000684 Published online by Cambridge University Press

agree that the very worst outcome is the one in which Jim sells his watch and Della cuts her hair. That is, a coordination problem has arisen where none existed for selfish players.[6]

5.2 Live and let live (Proposition 2)

Proposition 2 is closely related to proposition 1, the difference being that proposition 2 establishes that these new coordination problems introduced by altruism can lead to outcomes worse than Prisoners' Dilemmas. To see the intuition for this result, consider a more morbid version of the O. Henry story.

Two hikers are both bitten by a poisonous snake. They have with them one full dose of the antidote, but it is in two vials containing two different medicines that must be combined to be effective. Each hiker has possession of one of the vials. Both face a choice between (*X*) hoarding the dose, secretly giving the other water instead and (*Y*) sharing the dose of the medicine contained in their vial. If both hoard, the single medicine will be somewhat effective, putting odds of survival at 60%. If both share, so that each has half a dose of both medicines, the odds of survival for each will be 70%. If one hoards and the other shares, the one who receives some of each medicine (the hoarder) will survive with 80% probability, but the other will surely die. Assume that payoffs are simply the chance of survival and that the total payoff is calculated as the sum of these probabilities.

Selfish players face a Prisoner's Dilemma since one's chances of survival are always better when one hoards (60% is greater than 0%, if the other hoards; 80% is better than 70%, if the other shares). The unique rationalizable outcome and the unique equilibrium outcome is the one in which both players hoard.

Altruistic players, in contrast, face a Coordination Game, since it is better to share if the other shares (2*70% > 1*80%) but better to hoard if the other hoards (2*60% > 1*80%). In this case all outcomes are rationalizable; it is quite possible that rational players will fail to achieve either of this game's two equilibria and find themselves in the worst possible outcome, in which one of them dies and the other has an 80% chance of survival. Note in particular that this is worse than the hoard-hoard outcome (with each surviving with 60% probability) that was achieved by entirely selfish players in the Prisoner's Dilemma.

[6] Of course, whether the realized outcome is in fact a bad one is not clear. O. Henry's final paragraph includes these lines: "And here I have lamely related to you the uneventful chronicle of two foolish children in a flat who most unwisely sacrificed for each other the greatest treasures of their house. But in a last word to the wise of these days let it be said that of all who give gifts these two were the wisest."

5.3 With friends like these . . . (Proposition 3)

Proposition 3 demonstrates that when both players become partially altruistic it is possible that total payoffs fall compared to those realized in the outcome achieved by selfish players. To see this effect, consider the following story, which could motivate the precise payoffs represented in the asymmetric matrix in Figure 6. Players 1 (Al) and 2 (Bill) are two university roommates. Al has a stereo that he may (Y) lend to Bill or (X) not. Bill is considering whether to (y) throw a party at their apartment or (x) not.

If both are entirely selfish, then Al has a dominant strategy not to lend Bill his stereo. Note that Bill wants to borrow Al's stereo since Bill's payoffs are higher in row Y than in row X for both columns. Similarly, Al (who is more studious and less fun-loving than Bill) wants Bill *not* to throw a party since Al's payoffs are lower in column y than in column x for both rows. Finally, note that Bill prefers to throw a party if and only if he can borrow Al's stereo (x is preferred to y given X, but y is preferred to x given Y). The unique rationalizable outcome and the unique equilibrium outcome is the one in which Al does not loan the stereo and Bill does not throw a party.

If both Al and Bill are (partially) altruistic, it is possible that Al will have a dominant strategy to loan Bill his stereo (since it raises Bill's payoffs, which Al now cares about, under both of Bill's strategies). But if Al does loan Bill the stereo, it makes it more attractive for Bill to throw a party, despite the fact that Al would rather he not (and Bill cares about this to some extent). What the numerical example in subsection 4.3 shows is that it is possible for altruism to alter Al's relative payoffs so that he has a dominant strategy to loan Bill the stereo without making Bill so altruistic as to forgo the party for Al's benefit. That is, it is possible for the unique rationalizable outcome and the unique equilibrium outcome with (partially) altruistic agents to be the one in which Al loans the stereo and Bill throws the party; moreover, it is possible for this to yield lower total payoffs than the outcome in which the stereo is not loaned nor the party thrown.

6. CONCLUSION

I have shown two weaknesses of altruism in improving social welfare. First, using the solution concept of rationalizability, which allows coordination problems to flourish, I show that altruism introduces coordination problems where none exist for selfish players. This pushes one step further the observation among philosophers that altruism fails to solve existing coordination problems. Second, I show that although such coordination problems disappear completely under the alternative solution concept of efficient Nash equilibrium, partial altruism (in which agents weight others'

utility less than their own) may nonetheless lead to lower total payoffs than those attained by selfish players.

Certainly altruism does not always lead to worse outcomes in simple games such as those studied here. Figure 5 demonstrates that for many games (those with parameters in the unshaded regions of the figure) introducing altruistic preferences does not change the outcome of the game at all. Tables 1 and 2 point out that in most cases when outcomes do change, they change for the better, leading to increases in total social welfare. When introducing altruistic preferences changes the outcome of the game in a way that makes the players' worse off, it is often by introducing coordination problems; in such games, realized total welfare falls only if in fact the players fall victim to the new coordination trap. Whether this is likely to occur in any particular instance of such a game depends on the context and history that condition each player's expectations about the other's play. The examples presented here demonstrate that it is not implausible that in some circumstances, perhaps surprisingly, altruism may lead all players to fare worse than they would if they were selfish.

REFERENCES

Bernheim, Douglas. 1984. Rationalizable strategic behavior. *Econometrica* 52:1007–28

Esher, Ilan, Larry Samuelson, and Avner Shaked. 1998. Altruists, egoists, and hooligans in a local interaction model, *American Economic Review* 88:157–79

Henry, O. 1984. *41 Stories.* New York: Penguin Putnam

Levine, David, 1998. Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics* 1:593–622

Mackie, J. L. 1973. The disutility of act-utilitarianism. *Philosophical Quarterly* 23:289–300

Pearce, David. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52:1029–50

Regan, Donald. 1980. *Utilitarianism and cooperation.* Clarendon Press

Sen, Amartya. 1987. *On Ethics and Economics.* Blackwell Publishers

Sobel, Howard. 1985. Utilitarianism and cooperation. *Dialogue* 24:137–52

Sober, Elliott and David Wilson. 1998. *Unto others: the evolution and psychology of unselfish behavior.* Harvard University Press