Moral Implications of Rational Choice Theories

J. N. Hooker

Tepper School of Business, Carnegie Mellon University

July 2011

Abstract. Rational choice theories assert that human beings behave rationally, either in the narrow sense of rational self-interest, or in the broader sense that decisions are rationally based on preferences. These empirical theories make no direct ethical claims, but they may have relevance to ethics. Social contract theorists have maintained, for example, that rational individuals can assent to a social arrangement that promotes general welfare in some sense. In particular, self-interested business owners can, under the right conditions, rationally consent to regulation. Social choice theorists have argued in a mathematical mode that if we rationally derive social policy from individual preferences, we will adhere to certain ethical norms, perhaps a utilitarian or Rawlsian maximin principle. However, these arguments are based on strong assumptions, particularly with respect to interpersonal incomparability of utilities. Certain rational bargaining procedures, such as Nash bargaining or Raiffa-Kalai-Smorodinsky bargaining, have been shown to lead to outcomes that likewise have ethical content. The former has seen practical application in industry, and the latter results in a minimax relative concession principle similar to that derived by some social contract theorists.

Introduction

Rational choice theories attempt to explain human behavior as resulting from rational choices, particularly in an economic context. They have been a major part of the Western intellectual landscape since the market system replaced a medieval economy. This historical shift is seen as giving rise to *Homo economicus*—economic man—who is driven by self-interested economic calculation rather than a value system of loyalty and honor. Rational choice theories provide the central explanatory principle of much neoclassical economic theory and have generated a vast academic literature.

Rational choice theories are empirical theories of human behavior and, as such, make no ethical claims. However, they are often seen as having implications for ethics. In particular, they may address the question, "Why should I be moral?" If it can be shown that rational behavior results in ethical behavior, then the question has an answer: I should act morally because it is rational to

do so. Moreover, if people are already disposed to make rational choices, as rational choice theories assert, then there is reason for optimism that people will tend to be moral by following their proclivity to make rational choices.

A key question for linking rational choice theory with ethics is therefore whether rational choices at the individual level are ethical, due to the nature or consequences of rational decision making. This article addresses this question, while making no attempt to evaluate the truth of rational choice theories as empirical claims.

"Rational choice" can actually be interpreted in two ways. One identifies rationality with rational self-interest, the traditional motivator for *Homo economicus*. On this definition, rational choice theories subscribe to some form of *psychological egoism* (at least in economic matters), which is the view that humans act solely out of self-interest. Despite the widespread popularity of psychological egoism, the evidence for it is unclear. Again, however, the aim here is not to determine the truth of psychological egoism, but to ask whether self-interested rational choice is ethical. Many argue that it is, on the ground that self-interested choices make everyone better off in the long run. If so, the question, "Why should I be moral," has a more compelling answer: not only is it rational to be moral, but it serves my interest. Some take this a step further by deducing *ethical egoism,* which asserts that self-interested behavior is ethical not only because of its consequences, but as a matter of principle.

A second interpretation of rational choice does not presuppose self-interest. It assumes only that choices reflect the agent's preferences in a rational fashion, preferences that could be altruistic as well as self-interested. A great deal of research in economics and social choice theory has examined the logical consequences of supposing that people make rational choices in this sense. Social choice theory is particularly relevant to ethics, because it derives consequences from the assumption that people arrive at social policy in a rational way, and some of the consequences have the flavor of ethical principles---such as a utilitarian or Rawlsian maximin principle (see [1] for a survey). Related research has shown that bargaining in a fashion that might be seen as rational from an individual's point of view, such as Nash bargaining or Raiffa-Kalai-Smorodinsky bargaining, results in agreements that are fair in some sense. Because of the rigor of this research, much of which rests on mathematical proof, it deserves close examination for its moral implications.

Rational Choice and Deontology

Before examining the above arguments in detail, it is important to distinguish psychological rational choice theory from rational choice theory as it appears in deontological (duty-oriented) ethics. A line of argument characteristic of Kantian ethics, the dominant tradition in deontology, begins with the proposition that choices must rational because agency itself is rational. The

formal properties of rational agency then logically imply that choices must conform to certain moral principles, such as generalizability.

To expand on this argument a bit, agency must be rational in the sense that there must be some way to explain the agent's choices as based on a rationale. This is important because it distinguishes free action of a human from the behavior of, say, a bumblebee, even though both can be explained as the result of physical and chemical causes. Free choices are distinguished by the fact that one can plausibly attribute to the agent a line of reasoning that the agent sees as justifying the choice. But a rationale can be coherent only if it satisfies certain formal properties. In particular, it can justify an action only if it justifies the action for any agent to whom the rationale applies. This means that the reasons for an action must be consistent with the assumption that everyone to whom the reasons apply performs the action. This is the famous *generalization principle* of Kantian ethics. For example, lying merely because it is convenient for listeners to believe the lie violates the principle. If everyone with this reason for lying lied, no one would believe the lie, which is inconsistent with the reason for lying.

Kantian rational choice theory therefore differs from empirical rational choice theories in that it makes no claim that people in fact act rationally. It asserts only that if they do not, they are not moral agents. It also expects agents to have a conscious rationale for their choices, while most empirical theories have no such expectation; only the choices themselves matter. Yet there is a point of commonality, in that it may be possible to show that rational choices are ethical choices, whether rationality is as defined by Kantian ethics or by empirical rational choice theories.

Psychological Egoism

As noted earlier, one variety of rational choice theory asserts that human choices are based on self-interest. This is often seen as ethically relevant on the ground that self-interested choices benefit everyone in the long run. Adam Smith is routinely quoted as saying that the butcher, brewer and baker provide our dinner out of self-interest rather than benevolence. There is also Smith's famous remark that one who acts out of self interest is “led by an invisible hand to promote an end which was no part of his intention,” namely the welfare of society as a whole [18]. But if self-interested behavior makes everyone better off, then it must be ethical. This gives us a reason to be ethical: ethics coincides with self-interest.

To be fair to Smith, one should note that he explicitly rejected psychological egoism. Following his mentor Thomas Hutchinson, he believed that empathy is an important factor in human motivation and wrote an entire book, *Theory of Moral Sentiments* [17], to elaborate this idea. While he refers to the invisible hand in his book *The Wealth of Nations*, the book views self-interest as a destructive as well as a constructive force in economics and advocates government regulation to tame its excesses.

Nonetheless, if one supposes that individual self-interested actions really do maximize general welfare, one might conclude on utilitarian grounds that self-interest is ethical (even obligatory). People can therefore be ethical by following their inclination toward self-interest. A difficulty with this argument is that the connection between self-interest and general welfare is a question of fact that cannot be resolved by ethical or logical analysis. There is by no means agreement on whether the evidence supports such a connection. Even if the connection exists, it does not establish that self-interested behavior is ethical as a matter of principle. That is, it does not imply ethical egoism. Rather, it simply appeals to another ethical principle, utilitarianism, according to which ethical action must maximize total net utility. At best, the argument shows that self-interested behavior happens to be ethical because it happens to maximize utility.

A second argument for ethical egoism might begin with the principle that "ought implies can." That is, people cannot be obligated to perform acts of which they are incapable. But if people can act only out of self interest, then they cannot be obligated to act in any other fashion. Self-interested action is therefore always ethical.

This argument, like the previous one, fails to establish egoism as an ethical principle, because it rests on a contingent fact of nature that, even if true, could change with time. If self-interested action is ethical as a matter of principle, then it is ethical whether or not people act out of self-interest at any particular time. Ethical principles *judge* how people act, rather than *reflect* how they act.

Social Contract Arguments

One way to connect individual rational action with moral principle is through a social contract argument. The basic idea, which goes back to Thomas Hobbes [11], is that rational individuals will subscribe to a social agreement to behave morally. They will do so because the alternative is anarchy, in which life is "solitary, poore, nasty, brutish, and short." More importantly, they will continue to comply with the agreement, because (according to Hobbes) they will voluntarily install an authoritarian government that gives them strong incentive to do so.

In a modern economic context, a Hobbesian perspective might see commercial firms as lobbying on behalf of business regulation to avoid the chaos of an unregulated environment. Aware that business often has an incentive to flout regulations, they advocate substantial penalties and strong enforcement to make sure it is in their interest to comply. This sometimes occurs, but many would deny that it is required by a firm's rational self-interest.

A social contract argument need not appeal to self-interest. John Rawls, for example, argued in a Kantian vein that it is rational for individuals to agree on a social order that maximizes the welfare of the least advantaged [15]. The argument may be roughly put as follows. If I agree to

a particular social order, then I must have reasons for doing so. These reasons must be equally convincing to any rational person who agrees, including those on the bottom of society. But rational individuals can agree with a social order that puts them on the bottom only if no other social order would make them better off. It follows that a rational individual must endorse a *maximin* solution: one that maximizes the welfare of the worst-off. Rawls, unlike Hobbes, makes no claim that it is in an individual's self-interest to comply with the social contract. The only claim is that rationality requires compliance. It is therefore rational to be ethical, or at least to conform to a Rawlsian conception of distributive justice. In recent terminology, Rawls is a "contractualist," while Hobbes is a "contractarian."

The contractarian David Gauthier [8] argues that a social contract can be based on self-interest if individuals are "constrained maximizers." This means that, in a social contract setting, they choose dispositions to act rather than individual actions. Individuals enter into a social agreement because everyone gains from cooperation, as in the Hobbesean case, but the agreement entails adopting a "disposition" to comply with it. The problem with social agreements, of course, is that one can often do better for oneself by breaking the rules when others are following them. As a result, everyone breaks the rules, and everyone loses. This is known as a *prisoner's dilemma* situation, due to a famous example of it offered by mathematician Albert Tucker. Gauthier argues, however, that if an individual adopts a disposition to comply with a social contract, and makes the contract only with individuals who seem willing to adopt a similar disposition, then it is in the individual's rational self-interest to honor the contract even in a prisoner's dilemma situation.

One might interpret Gauthier's "disposition" as a mild form of Hobbesean government. Having adopted a disposition to comply does not psychologically compel one to comply, but it somehow provides a rational incentive to do so, much as a government provides an incentive to obey the law. In a business context, adopting a disposition might translate to building a corporate culture that favors compliance with industry self-regulation. This may incentivize compliance if the organizational cost of disrupting corporate culture is higher than the gain of breaking the rules.

The success of Gauthier's argument is a matter of dispute (e.g., [13]), but even if sound it does not serve the cause of morality unless rational individuals (or firms) negotiate a social contract with some kind of moral content. Gauthier maintains rational individuals will strive to minimize the *relative* concession they must make to obtain an agreement. That is, each individual's concession is measured relative to that individual's sacrifice that would result from no agreement. Bargaining leads to a *minimax relative concession*, an agreement that minimizes the maximum relative concession of the players. This is essentially the Raiffa-Kalai-Smorodinsky bargaining solution, which is discussed below. The solution can be viewed as incorporating a fairness norm, perhaps due to its resemblance to the Rawlsian maximin principle. It will be seen, however, that the solution can be counterintuitive when there are three or more players.

The normative implications of self-interested rational choice continue to be a subject of interest, particularly for political philosophers. Geoffrey Brennan and Alan Hamlin [4], for example, follow Gauthier's example of analyzing the rationality of dispositional choices, while Russell Hardin [9] takes the more traditional approach of analyzing individual choices. A review of recent work along these lines can be found in [5].

Social Choice Theory

Rational choices have been intensively examined in the social choice theory literature, which generally assumes no particular connection between individual self-interest and social welfare. Rather, it carries out a logical analysis of what it means to agree rationally on social policy when individual preferences differ. It then derives structural characteristics that the resulting social policy must have, some of which may resemble ethical principles (see [7] for a survey of these results). One might therefore conclude, on the basis of purely logical analysis, that it is rational to be ethical, or at least to assent to an ethical social policy.

A perennial issue with this employment of social choice theory is how much normative content is already built into the formal properties of rational agreements. If the negotiators are required to be ethical, then there is little surprise if their agreements are ethical. It is impossible to assess this type of objection without examining the actual derivations in some detail, even if this requires a bit of mathematics. Informal derivations of utilitarian and Rawlsian maximin principle are therefore given below, followed by derivations of two bargaining rules that seem to have ethical content.

Deriving the Utilitarian Principle

Social choice theory begins with individual preferences, which are defined over a set of possible states that society might assume. Each state x provides utility $u_i(x)$ to individual i, who prefers state x to state y when $u_i(x) > u_i(y)$. Individuals may rank the states differently, which raises the question of how to rank states in a way that takes into account everyone's preferences. This is the problem of designing a *social welfare function*, which ranks states socially by assigning them social utilities. More precisely, if $u = (u_1, ..., u_n)$ is a tuple of utility functions, one for each individual, then a social welfare function f_u for u assigns an social utility $f_u(x)$ to each state x. Then x is socially preferable to y if the social utility of x is higher; that is, if $f_u(x) > f_u(y)$.

For example, a utilitarian social welfare function sets $f_u(x) = \Sigma_i u_i(x)$. That is, state x is preferable to y when x generates more utility across the population. This kind of calculation obviously

assumes that the utilities $u_i(x)$ are comparable across persons, at least to the extent necessary to add them up in a meaningful way.

The issue of interpersonal comparability is in fact a central theme of social choice theory [16], and it is analyzed as follows. The key is to ask how the individual utility functions could be altered without changing the social ranking. Suppose, for example, that each individual's utility is multiplied by the same factor $\beta > 0$. One would not expect this to change the social ranking of states, because it simply rescales the units in which utility is measured. Suppose, however, that the units are rescaled and a different constant α_i is added to each person's utility. That is, each individual's utility u_i is changed to $\beta u_i + \alpha_i$. Should this change the ranking of states? It would not alter a utilitarian ranking, because a state x that is socially preferable to y remains preferable after the utilities are altered. That is, $\Sigma_i u_i(x) > \Sigma_i u_i(y)$ if and only if $\Sigma_i(\beta u_i(z) + \alpha_i) > \Sigma_i(\beta u_i(y) + \alpha_i)$. Thus a utilitarian ranking is *invariant* under a utility transformation $\varphi = (\varphi_1, \ldots, \varphi_n)$ given by $\varphi_i(u_i) = \beta u_i + \alpha_i$ for each i. It is convenient to call this transformation a *translated rescaling*.

This means that full comparability across persons is unnecessary for a utilitarian calculation to be meaningful. In particular, *level comparability* is not required: it need not be possible to compare the absolute level of person i's utility with that of person j, because a translated rescaling can make either utility larger, depending on the size of the translations α_i and α_j. However, it must be possible to compare utility *differences* across individuals, because these are unaffected by φ. That is, $u_i(y) - u_i(x) > u_j(y) - u_j(x)$ if and only if $\varphi_i(u_i(y)) - \varphi_i(u_i(x)) > \varphi_j(u_j(y)) - \varphi_j(u_j(x))$. This is called *unit comparability*.

Unit comparability makes a utilitarian social welfare function meaningful, but one can ask if there are further conditions under which the social welfare function *must* be utilitarian. There are. If social choice obeys two formal properties that might be associated with rationality, the social welfare function must have the form $f_u(x) = \Sigma_i u_i(x)$.

One property is *anonymity*, which says that it makes no difference which utility function belongs to which individual. Thus suppose that tuple $v = (v_1, \ldots, v_n)$ of individual utility functions is obtained from $u = (u_1, \ldots, u_n)$ by renaming the individuals. That is, there is a permutation $\pi(1), \ldots, \pi(n)$ for which $(v_1, \ldots, v_n) = (u_{\pi(1)}, \ldots, u_{\pi(n)})$. Then f_u and f_v must define the same social ranking; that is, $f_u(x) > f_u(y)$ if and only if $f_v(x) > f_v(y)$ for all pairs of states x and y.

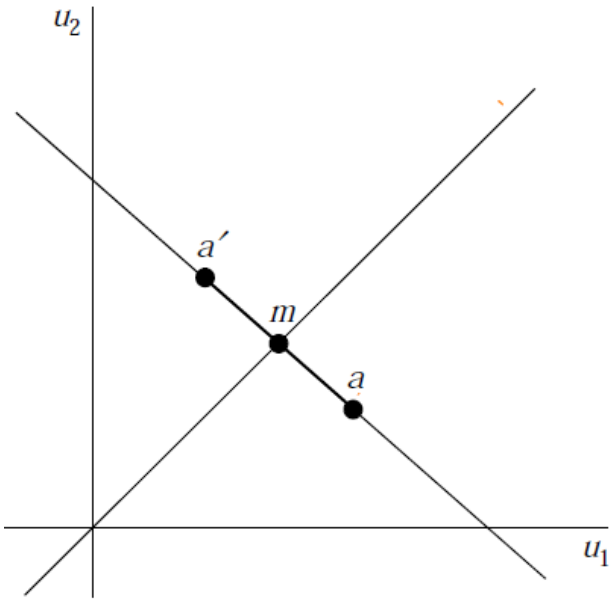

Figure 1. Derivation of a utilitarian social welfare function for the two-person case.

There is also a *strict Pareto* condition, which requires that states preferred by all individuals be preferred socially. Formally, if $u_i(x) \geq u_i(y)$ for all i, then $f_u(x) \geq f_u(y)$, and if in addition $u_j(x) > u_j(y)$ for some j, then $f_u(x) > f_u(y)$.[1]

Given these conditions and unit comparability, the social welfare function must be utilitarian. The essence of the argument may be seen in the two-person case [3], which is graphed in Figure 1. Each state x generates a utility vector $u(x) = (u_1(x),u_2(x))$ that can be plotted as a point on the graph. All states whose utility vectors appear on the same 45^o line running from upper left to lower right have the same total utility. It therefore suffices to show that all such states are ranked equally, because the strict Pareto condition then implies that x is ranked higher than y if and only if $u(x)$ appears on a higher 45^o line, which is precisely the utilitarian ranking. To show this, consider any utility vector $a = (a_1,a_2)$ on a given 45^o line as shown in the figure. Let $a' = (a_2,a_1)$, and let m be the midpoint of the line segment from a to a'. Due to anonymity, a and a' must receive the same ranking. Now suppose, contrary to the claim, that m is preferred to a. The transformation $\varphi(u) = u + b - a$ maps a into m and m into a'. Due to unit comparability, this

[1] An *independence* property is also presupposed, which requires that the ranking of x and y be independent of the utilities of other states. This is implicit in the assumption that one can rank x and y by comparing the values of some social welfare function on x and y.

transformation does not change rankings, and a' is preferred to m. By transitivity, this implies a' is preferred to a, a contradiction. Because a is an arbitrary point on the 45^o line, all points on the line must be ranked equally with m and therefore with each other.

Deriving the Maximin Principle

The maximin principle ranks states according to the utility of the worst-off individual. The social welfare function is therefore $f_u(x) = \min_i\{u_i(x)\}$. The ranking is unchanged when the same monotone increasing transformation is applied to each individual utility. That is, it is invariant under a transformation $\varphi = (\varphi_0, \ldots, \varphi_0)$ for which $\varphi_0(u_i) > \varphi_0(v_i)$ whenever $u_i > v_i$. This means that the maximin principle requires level comparability, because the monotonicity of φ_0 implies that $u_i(x) > u_j(x)$ if and only if $\varphi_0(u_i(x)) > \varphi_0(u_j(x))$.

To derive the maximin principle, one additional property is needed: *separability*, which states that individuals to whom all states look the same play no role in the social ranking. More precisely, let S be a subset of individuals i such that for any tuple u of utility functions, $u_i(x)$ is the same for every state x. Then f_u and f_v give the same ranking if for all individual i not in S, $u_i(x) = v_i(x)$ for all states x.

The claim is that given level comparability and the above axioms, the social welfare function must be the maximin function. Curiously, however, these premises imply only that welfare function is maximin *or* maximax [6]. The latter maximizes the utility of the *best-off* individual; that is, $f_u(x) = \max_i\{u_i(x)\}$. To infer a minimax principle, one must rule out the maximax principle on some other ground.

Again the idea of the argument can be conveyed in the two-person case [3], where separability does not play a role. Let $v = (v_1,v_2)$ as shown in Figure 2 be an arbitrary utility vector, and let $v' = (v_2,v_1)$. Divide the plane into regions about the 45^o diagonal line as shown. Then it suffices to demonstrate that one of two situations must obtain: (a) all the points in regions A, B, C and their reflections A′, B′, C′ (shaded area in Figure 3(a)) are preferable (or indifferent) to v, and all other points are worse than v , or (b) all the points in regions B, C, E and their reflections (shaded area in Figure 3(b)) are preferable (or indifferent) to v, and all other points are worse. Case (a) ranks points relative to v by a maximin criterion, and case (b) ranks them by a maximax criterion. Because v is arbitrary, the social welfare function must be maximin or maximax.

Consider any point a in the interior of region A. It can be shown as follows that if a is preferable to v, then case (a) obtains. A similar argument derives case (b) if a is worse than v. Point a cannot be indifferent to v, because if it were the argument to follow would show that all points in A are indifferent to v, which is impossible because some dominate others due to strict Pareto.

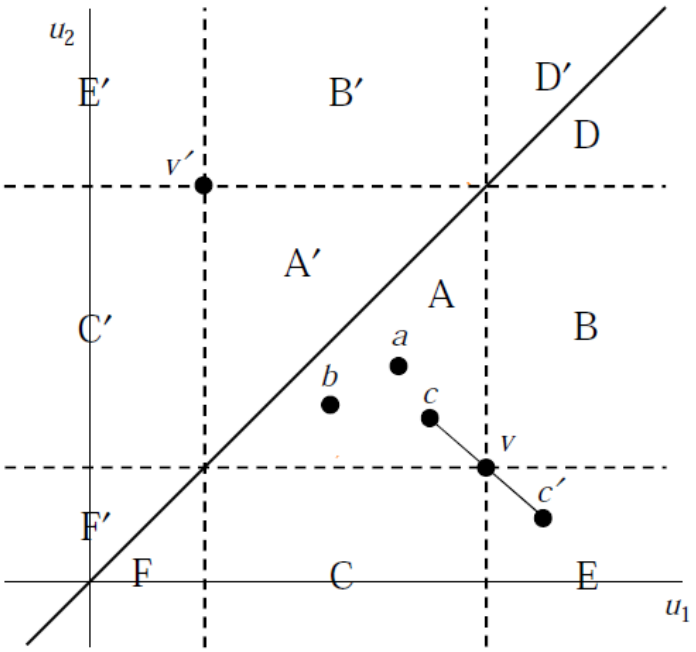

Figure 2. Derivation of a maximin social welfare function for the two-person case.

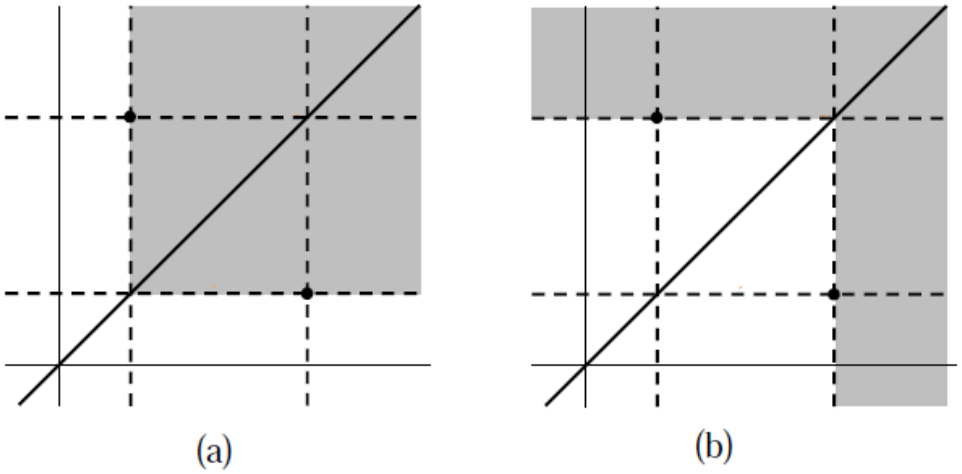

Figure 3. Cases (a) and (b) of the derivation of a maximin social welfare function.

First, show that all points in the interior of A are preferable to *v* by considering any other point *b* in the interior of A. Note that $v_1 > a_1 > a_2 > v_2$ and $v_1 > b_1 > b_2 > v_2$. This means it is possible to design a monotone increasing transformation φ_0 that maps *v* to itself and *a* to *b* (Figure 4). This implies, by level comparability, that *a* and *b* have the same ranking relative to *v*. Thus all points in A have the same rank relative to *v* and are therefore preferable to *v*, as claimed. By

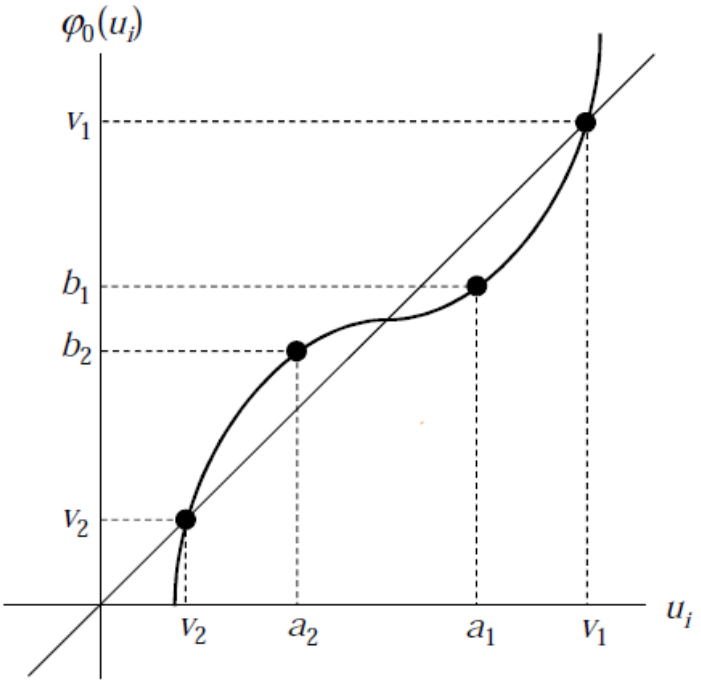

Figure 4. A monotone increasing invariance transformation that maps v_1, a_1, a_2, and v_2 to v_1, b_1, b_2, and v_2, respectively.

anonymity, all points in A′ have the same ranking relative to v', and therefore relative to v because v and v' are symmetric. A similar argument shows that all points in E and E′ have the same ranking relative to v. To show that v is preferable to these points, select any point c in A on the 45° line shown. Then the monotone increasing transformation $\varphi_0(u_i) = u_i + v_i - c_i$ (which is the same for $i = 1,2$) maps c to v and v to c'. By level comparability, v must be preferable to c' and therefore all points in E. Finally, points in B and C are preferable to v by strict Pareto, as are points in B′ and C′ by anonymity. Similarly, points in D and F and their reflections are worse than v. Points on the boundaries of the regions are dealt with in a similar fashion, and case (a) follows.

Assessment of the Social Choice Arguments

One can now examine the normative assumptions implicit in the rational choice axioms and their role in making the proofs go through. The strong Pareto condition is rather innocuous and probably begs no interesting ethical issues. The anonymity condition, however, clearly has normative content because it implies impartiality toward individuals. Yet Kantian arguments

rest on a similar condition by supposing that a rationale justifying one person's action must justify the same action for anyone to who the rationale applies. It seems to be a basic trait of rational action that it should depend only on reasons, not who has the reasons. Rationality might similarly dictate that rational choice should depend only on the utilities that result, not on who has the utilities. Anonymity therefore seems a reasonable starting point.

Yet anonymity is not central to the proofs. One can derive social welfare criteria with a strong utilitarian or Rawlsian flavor without it. For example, unit comparability without anonymity implies a utilitarian criterion with weights: $f_u(x) = \Sigma_i w_i u_i(x)$. This is basically because all points on the line connecting any two indifferent points can be shown to have equal rank, using the invariance transformation much as before, and $\Sigma_i w_i u_i$ is constant along the line for appropriate weights w_i. Also, level comparability without anonymity implies a minimax (or maximax) criterion for all states in which a given individual is worse off, using the same invariance arguments as above. In both proofs, the interpersonal comparability assumption is doing most of the work.

It may seem reasonable to assume unit comparability when deriving a utilitarian function, on the ground that unit comparable makes a utilitarian calculation meaningful. However, unit comparability remains if the ranking is invariant only under a proper subset of translated rescalings, while the proof assumes invariance under *any* translated rescaling. In other words, the proof assumes that utilities have unit comparability and *no more* than unit comparability. This strong assumption is already very close to utilitarianism. A Rawlsian, for example, would immediately object to it because it makes the comparison of worst-off individuals meaningless from the start. If the utility vectors are $(u_1(x),u_2(x)) = (1,1)$ in state x and $(u_1(y),u_2(y)) = (2,0)$ in state y, the Rawlsian prefers x because of the higher utility of the worse-off individual. However, a translated rescaling maps these vectors to (0,2) and (1,1), respectively, in which the Rawlsian preference is reversed. A similar point applies to the derivation of a maximin welfare function from level comparability.

Nash Bargaining

Bargaining theories might be regarded as undertaking a rigorous and mathematical analysis of social contract negotiation. They show that under certain rationality assumptions, bargaining among rational individuals results in agreements that have interesting structural characteristics. These characteristics may also result from a certain kind of bargaining procedure. When the characteristics have an ethical flavor, one might say that there is a reason to be ethical: an ethical arrangement is the result of rational bargaining.

The best-known bargaining theory is due to John Nash [14] and yields the *Nash bargaining solution* (which should not be confused with the Nash equilibrium of noncooperative game

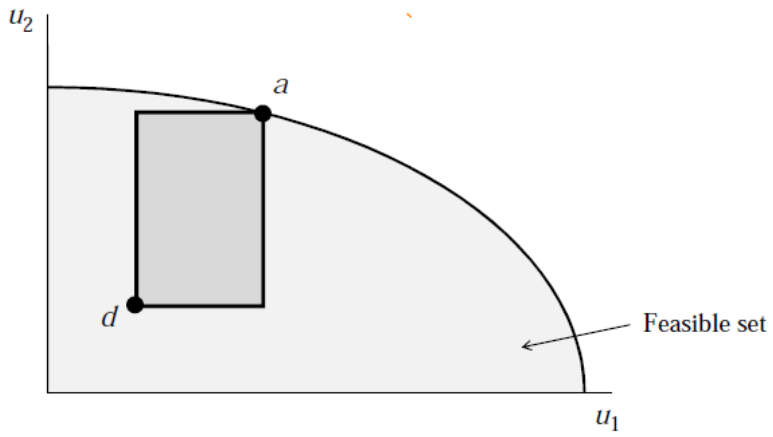

Figure 5. Nash bargaining solution for two persons.

theory). Figure 5 illustrates the Nash bargaining solution for two persons. The point *d* represents the utility vector $(u_1(x),u_2(x))$ of a *default* position *x*. This is the state of affairs before bargaining starts, and the state that remains if bargaining fails. The region under the curve represents the *feasible set*, which is the set of possible agreements consistent with available resources. The point *a* represents the utility vector $(u_1(y),u_2(y))$ after the players arrive at a deal *y*. The Nash bargaining solution selects a deal *y* that maximizes the area of the rectangle shown. In other words, it maximizes the product $(a_1 - d_1)(a_2 - d_2)$. If there are *n* players, the solution selects $a = (a_1, \ldots, a_n)$ to maximize $\prod_i(a_i - d_i)$.

The product formula tends to result in near-maximum utility without overly depriving any one player and can therefore be seen as enforcing some kind of fairness. It has enjoyed a degree of acceptance in practical application and is, for example, widely used to allocate bandwidth to information packets in telecommunications networks. The motivation is to obtain near-maximum throughput while not excessively discriminating against packets from any one source.

The Nash bargaining solution has been defended on both axiomatic and procedural grounds. One axiomatic argument assumes *cardinal noncomparability*, which requires that the ranking of utility vectors be invariant under the transformation $\varphi(u) = (\varphi_1(u_1), \ldots, \varphi_n(u_n))$, where $\varphi_i(u_i) = \beta_i u_i + \alpha_i$. Note that the scaling factor β_i can be different for each individual, whereas in unit comparability it is the same. The argument also assumes anonymity and a Pareto condition, as well as *independence of irrelevant alternatives*,[2] which is necessary if the product criterion is to make sense. It requires that if *a* is the Nash bargaining solution for a given feasible set, it remains the solution if the set is reduced without excluding *a*.

[2] This is not identical to the axiom of the same name used in the proof of Arrow's famous impossibility theorem.

It is impressive that a product criterion could be derived from these axioms, and the proof is quite interesting. However, the premises are again strong. While much attention is focused on the independence axiom, it is rather innocuous in this context. The strongest premise, and the one that does most of the work in the proof, is the assumption of cardinal noncomparability. It leaves room for very little interpersonal comparability, because the ranking must be invariant under transformations that can destroy any conception of quantity in the utilities. One can ask how a theory can account for distributive justice if it denies the possibility of comparing individual outcomes to this extent.

There are results showing that certain negotiating procedures terminate in a Nash bargaining solution, although none of the results are straightforward or easy to state. John Harsanyi, for example, showed that the following bargaining procedure converges to a Nash solution [10]. Suppose again that d is the default position. Player 1 makes offer a, and player 2 makes offer b. If p_2 is the probability that player 2 will reject a, then player 1 will stick with offer a (rather than make a counteroffer) if his expected utility $(1 - p_2)a_1 + p_2d_1$ of doing so is greater than the utility b_1 of accepting b; that is, if his estimated value of p_2 is less than $(a_1 - b_1)/(a_1 - d_1) = r_1$. Player 2 makes a similar calculation, and she will stick with offer b if her estimated value of p_1, the probability that player 1 will reject b, is less than $(b_2 - a_2)/(b_2 - d_2) = r_2$. The key assumption is that it is rational for player 1 to make a counteroffer, rather than player 2, if $r_1 < r_2$. That is, player 1 should counteroffer if the conditions under which he would stick with his offer are stricter than the conditions under which player 2 would stick with her offer. Given this, some elementary algebra shows that each step of the negotiation process improves the product criterion. If there is a minimum distance between offers, bargaining converges to a Nash solution. One must, of course, decide whether the key assumption is reasonable.

Binmore, Rubinstein and Wolinsky [2] derived that under rather complicated conditions, a somewhat different negotiation strategy converges to a Nash solution. An important element of their framework is the time value of utility: a player is willing to accept somewhat less in order to get an agreement sooner. The equilibrium outcome approaches the Nash solution as the time lapse between offers goes to zero.

Results of this kind are difficult to assess but suggest that a Nash solution is not unrelated to a reasonable bargaining procedure. The Nash solution itself, however, has received criticism. The outcome depends heavily on the default or starting position, a subject of much discussion in the literature. If the starting position is already unfair, then the Nash solution is likely to be unfair as well. To this extent, the Nash solution begs the question of rational allocation. However, this is a problem shared to some degree by most bargaining and social contract theories. The Nash solution can also be counterintuitive even when the default position is reasonable. One would expect, for example, that if the feasible set is enlarged, then no player's share should decrease in a rational allocation. There are examples in which a Nash solution violates this expectation. They provided one motivation for the development of Raiffa-Kalai-Smorodinsky bargaining.

Raiffa-Kalai-Smorodinsky Bargaining

The Raiffa-Kalai-Smorodinsky (RKS) bargaining solution is defined with respect to a "ideal" outcome in which each player's utility is maximized [12]. Figure 6 illustrates the two-person situation. There is again a default position *d* and a feasible set. The point *g* represents an ideal outcome in which each player's share is the maximum possible within the feasible set, regardless of the consequences for the other players. The RKS solution is the best feasible point on the line segment from *d* to *g*.

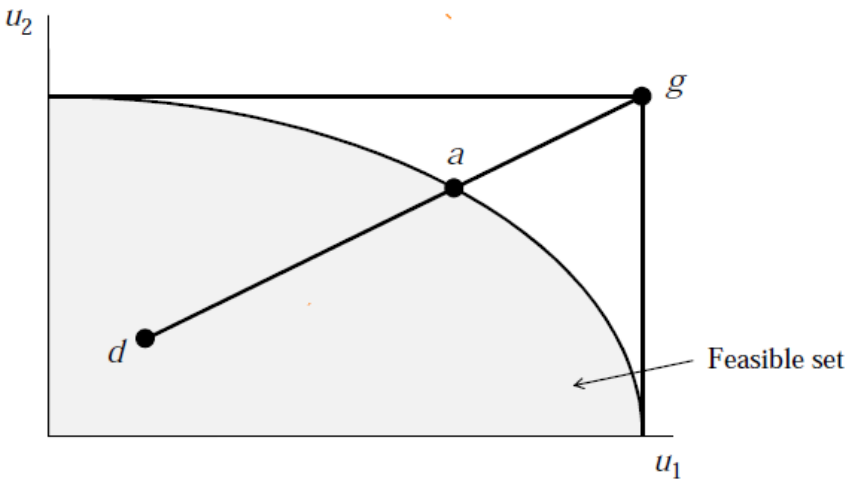

Figure 6. Raiffa-Kalai-Smorodinsky bargaining solution.

The RKS solution can be seen as the result of making proportional allocations to the players. Each player *i* has a maximum gain $g_i - d_i$. First, allot each player 10% of his or her maximum gain, then 20%, and so on until further allocations are infeasible. It is clear that enlarging the feasible set can never cause a player's allocation to decrease (*monotonicity*).

The axiomatic derivation of the RKS solution again assumes cardinal noncomparability, anonymity and a Pareto condition. However, it replaces the independence of irrelevant alternatives with the monotonicity property just mentioned [19]. The derivation is problematic for the same reason as the derivation of the Nash solution: cardinal noncomparability is a very strong assumption that rules out meaningful comparison of utilities as quantities.

A possible bargaining justification is that the RKS solution achieves the minimax relative concession mentioned earlier. If player *i* accepts offer *a*, the player's concession (with respect to the ideal) is $g_i - a_i$. The concession if negotiation fails is $g_i - a$, which means that the *relative* concession is $(g_i - a_i)/(g_i - d_i)$. It might be argued that the players will try to minimize their relative concessions and will reach equilibrium when they minimize the maximum relative

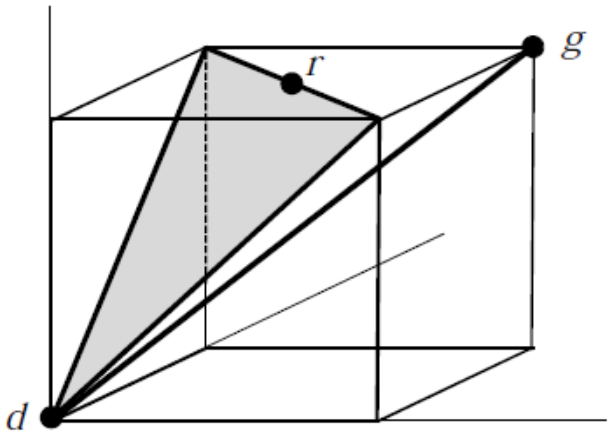

Figure 7. Anomalous example for the RKS bargaining solution, which in this case is the default position d. The feasible set is the shaded triangular area, and the Rawlsian minimax solution is r.

concession among them. But minimizing the maximum relative concession is algebraically the same as maximizing the minimum relative gain $(a_i - d_i)/(g_i - d_i)$, which is what the RKS point accomplishes.

In the two-player case, the RKS solution resembles the Rawlsian maximin solution, except that it maximizes the minimum relative gain rather than the minimum utility. It might be seen as having ethical content on that basis. However, it leads to some anomalies when there are three or more players (Figure 7). The line segment from *d* to *g* leaves the feasible set at *d*, which means that *d* is the RKS solution---scarcely a rational outcome. The Rawlsian maximin solution is *r*.

Conclusions

Although empirical rational choice theories make no ethical claims, they can have ethical relevance. The rational behavior they attribute to humans may lead to ethical choices. If so, rational choice theory suggests that we not only have a "reason" to be ethical---it is rational to be so---but that we will in fact be ethical much of the time.

Some rational choice theories equate "rational" with "self-interested" and assert that people invariably act out of self-interest. If self-interested action contributes to the general welfare, then presumably one can be ethical simply by yielding to one's inclination to act out of self-interest, although this popular view is difficult to defend empirically. Recent refinements of Hobbesian social contract arguments take a revisionist approach, suggesting that it may be in one's self-interest to negotiate agreements that are fair in some sense, and to adopt a disposition to comply

with them. The agreements themselves may resemble a Raiffa-Kalai-Smorodinsky (RKS) bargaining solution.

If rational choices are not assumed to be self-interested, they may nonetheless be ethical simply by virtue of their formal properties. The logical implications of rational choices have been extensively explored in social choice theory, with some interesting conclusions. If people aggregate their preferences in a manner that satisfies certain axioms, then the resulting social policy will, for example, be utilitarian or Rawlsian, depending on the degree to which it is meaningful to compare utilities across persons. The interpersonal comparability assumptions, however, are quite strong because they actually limit the degree of comparability. As a result, they already embody more normative content that, say, the rationalistic premises of Kantian arguments.

A rational bargaining framework may also lead to an agreement that seems ethical in some sense, such as the Nash or RKS bargaining solution. The interpersonal noncomparability assumptions are even stronger in this case, but they perhaps can be circumvented by showing that these bargaining solutions are the outcome of a reasonable bargaining strategy. This may provide a "reason" to make agreements that are ethical in some sense: the agreements would result from rational bargaining in any case. Nash and RKS solutions do not necessarily satisfy standard ethical criteria, however, and they can be quite counterintuitive. Also there is no consensus on how to specify or justify a reasonable default position from which to start bargaining.

One might conclude from all this that the ethical implications of rational choice theory are far from straightforward, but very suggestive. They are much too suggestive to abandon the research program, which might even be profitably expanded. For example, the classical rationalistic assumptions of deontological ethics may describe actual human behavior more than is normally supposed. These assumptions are also minimal, requiring only an anonymity axiom: reasons justify an action for one agent only if they justify it for any agent to whom the reasons apply.

An extremely popular economics game, the ultimatum game, illustrates this point. Half the players receive, say, 100 euros and are given the option of donating any portion of the gift to an anonymous recipient in the other half of the group. If the recipient rejects the donation, both the donor and recipient lose their money. The rationally self-interested choices seem obvious: the donor should give one euro, and the recipient should accept it. However, the average donor gives way about a third of the money, and many give away half. The game is frequently interpreted as showing how irrational human beings can be. However, giving away half the money, and accepting the gift, is the Rawlsian maximin solution. An alternative interpretation of the game is that people are entirely rational but in a different sense: they are rational Kantian agents, and a maximin solution therefore seems reasonable.

Homo economicus has invested heavily in the self-interest motive. We train ourselves to respond to it. Our belief in it approaches irrational ideology, as witnessed by the enormous popularity of psychological egoism despite much evidence of its falsehood, and our insistence on quoting such figures as Adam Smith out of context. The business world, in particular, relies on self-interest instinctively. When a manager wants to induce people to work together toward a common goal, personal incentives are the most obvious and most easily manipulated tool at hand.

At the same time, we rely on the inherent rationality of ethical behavior. In the Western world, at least, we tend to obey laws that seem reasonable to us and break those that do not. Because compliance with the law is largely voluntary in any but a police state, we must construct a rationale that justifies legal and ethical conduct in our minds, and we have been doing so for centuries. Such classic works as Blackstone's *Commentaries on the Laws of England*, and countless closely-reasoned judicial opinions, provide a rational basis for law. Ethical philosophy has rested heavily on rational persuasion, not only in the classical works of antiquity and the rationalistic theories of enlightenment thinkers, but in the Christian and Islamic traditions as well. Even in the business world, some wise managers have discovered that convincing workers of the company's positive contribution to society motivates them as strongly as appeals to self-interest.

An alternative to emphasizing self-interest as a primary motivator, and attempting to show that rational self-interest leads to ethical behavior, is to recognize that we are perfectly capable of acting rationally in a broader sense---a sense that is already ethical because of its anonymity assumptions. More importantly, we can cultivate a disposition to be rational, much as we have cultivated a reliance on self-interest. Empirical rational choice theories can explore this possibility and, in so doing, guide our moral development.

References

[1] K. Arrow, A. Sen, and K. Suzumura, eds, *Handbook of Social Choice and Welfare*, Vol. 1, Handbooks in Economics, Elsevier, Amsterdam, 2002.

[2] Ken Binmore, Ariel Rubinstein, and Asher Wolinsky, "The Nash Bargaining Solution in Economic Modeling," *Rand Journal of Economics* **17** (1986) 176-188.

[3] C. Blackorby, D. Donaldson, and J. A. Weymark, "Social Choice with Interpersonal Utility Comparisons: A Diagrammatic Introduction," *International Economic Review* **25** (1984) 327-356.

[4] Geoffrey Brennan and Alan Hamlin, *Democratic Devices and Desires*, Cambridge University Press, 2000.

[5] Thomas Christiano, "Is Normative Rational Choice Theory Self-Defeating?" *Ethics* **115** (2004) 122-141.

[6] C. D'Aspremont and L. Gevers, "Equity and the Informational Basis of Collective Choice," *Review of Economic Studies* **44** (1977) 199-209.

[7] Wulf Gaertner, *A Primer in Social Choice Theory*, Oxford University Press, 2009.

[8] David Gauthier, *Morals by Agreement*, Oxford University Press, 1987.

[9] Geoffrey Hardin, *Liberalism, Constitutionalism and Democracy*, Oxford University Press, 1999.

[10] John C. Harsanyi, *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge University Press, 1977.

[11] Thomas Hobbes, *Leviathan*, originally published 1651.

[12] E. Kalai and M. Smorodinsky, "Other Solutions to Nash's Bargaining Problem," *Econometrica* **43** (1975) 513-518.

[13] Jody S. Kraus and Jules L. Coleman, "Morality and the Theory of Rational Choice," Ethics **97** (1987) 715-749.

[14] John Nash, "The Bargaining Problem," *Econometrica* **18** (1950) 155-162.

[15] John Rawls, *A Theory of Justice*, Harvard University Press, originally published 1971, revised edition 1999.

[16] K. W. S. Roberts, "Interpersonal Comparability and Social Choice Theory," *Review of Economic Studies* **47** (1980) 421-440.

[17] Adam Smith, *The Theory of Moral Sentiments*, originally published 1759, last revised by Smith 1790.

[18] Adam Smith, *An Inquiry into the Nature and Causes of the Wealth of Nations*, originally published 1776, last revised by Smith 1789.

[19] W. Thompson, "Cooperative Models of Bargaining," in R. J. Aumann and S. Hart, eds., *Handbook of Game Theory*, vol. 2, North-Holland, 1994.