Economics and Philosophy, **26** (2010) 189–216
doi:10.1017/S0266267110000192

THE MORAL BASIS OF PROSPERITY AND OPPRESSION: ALTRUISM, OTHER-REGARDING BEHAVIOUR AND IDENTITY

KAUSHIK BASU
Cornell University

Much of economics is built on the assumption that individuals are driven by self-interest and economic development is an outcome of the free play of such individuals. On the few occasions that the existence of altruism is recognized in economics, the tendency is to build this from the axiom of individual selfishness. The aim of this paper is to break from this tradition and to treat as a primitive that individuals are endowed with the 'cooperative spirit', which allows them to work in their collective interest, even when that may not be in their self-interest. The paper tracks the interface between altruism and group identity. By using the basic structure of a Prisoner's Dilemma game among randomly picked individuals and building into it assumptions of general or in-group altruism, the paper demonstrates how our selfish rationality interacts with our innate sense of cooperation. The model is used to outline circumstances under which

This paper has been in incubation for a long time and I have accumulated more intellectual debts than I can inscroll in a footnote. While the risk of omission is, in this case, a certainty, I must record my gratitude to Geir Asheim, Talia Bar, Alaka Basu, Anindya Bhattacharya, Anil Bisen, John Bone, John Gray, Matt Jackson, Hyejin Ku, Ashok Mathur, Dilip Mookherjee, Bhiku Parekh, Amartya Sen, Richard Swedberg and Maria Monica Wihardja. Some of the initial ideas were part of my Vera Anstey Lecture, delivered at the 88th Annual Conference of the Indian Economic Association. I have subsequently presented this paper at conferences at York University, UK (on 'Identity, Responsibility and Justice', May 2007) and the University of California, Riverside (on 'The Non-Welfarist Basis of Welfare Economics', October 2007), and at the Indian Statistical Institute, New Delhi, and the paper has benefited greatly from the discussions that followed. Finally, I am grateful to an anonymous referee and an editor of the journal for extensive and valuable comments on the paper.

cooperation will occur and circumstances where it will break down. The paper also studies how sub-groups of a society can form cooperative blocks, whether to simply do better for themselves or exploit others.

1. CELEBRATING SELF-INTEREST

That the butcher, the baker and the bee-keeper, pursuing nothing but their self-interest, can bring about social order is not an easy idea to grasp. Hence, the proposition that it is possible for meat to arrive on the diner's table, bread to get delivered to the street corner deli, and honey to travel from the remote Tasmanian farm to the Edinburgh restaurant without anybody intending to help anybody else was a stunning intellectual insight. It is not surprising that when, on 9 March 1776, Adam Smith's book, *The Wealth of Nations*, containing this proposition was published it was quickly recognized as a classic.[1] So enamoured were the political economists of that time and their progeny, the economists, that this became the central tenet of economic theory. That individuals would be self-seeking was not just taken to be a fact, but celebrated. Development and growth were attributed to the actions of such atomistic selfish individuals. This, in turn, has tended to obscure the fact that rapid growth and successful development may also require individual integrity and altruism, and the ability of individuals to forego some personal advantages for reasons of societal benefit.

In the early 1990s I used to take a team of research students to a cluster of villages in one of the most anarchic and poor regions of India – now in the state of Jharkhand. Seeing the economic inefficiencies and lack of development in the region, it would be natural for a budding economist to proffer the popular advice that what is needed is for less government and for individuals to be left free to pursue their self-interest. But such an advice would be quite absurd in this case. There was no trace of any government for '*less* government' to be a feasible option. And there was no dearth of individually selfish behaviour either. What was lacking was the fauna and flora of social values and the cooperative spirit that make economic efficiency and development possible. Contrary to what many textbooks teach us, the regions of the world which are economically the biggest disasters are often the ones which are models of the free market, with amoral individuals seeking nothing but their own self-aggrandizement.

Ever since Adam Smith's classic, methodological individualism has become such a deeply entrenched foundation stone of economics that the

[1] It is interesting to note that, in the first two or three decades after the book came out, Smith was considered a renegade thinker (Rothschild 2001). He would become the voice of orthodoxy and be claimed by the conservatives only after the safety of his death in 1790.

predominant tendency has been to refuse to admit that a person can and often does act in his national interest or class interest or caste interest or interest based on some other collective identity.[2] When the need to recognize such collective interests was felt, the acceptable method among economists has been to *derive* the social or cooperative behaviour from the primitive of self-interest.

The aim of this paper is to break away from this individualistic tradition and to treat as a primitive the fact that individuals have hard-wired in them, admittedly to varying extents, the 'cooperative spirit', which allows them to often work in the collective interest, even when that may not be in their *self*-interest and to make sacrifices for the sake of fairness and integrity.[3]

It must be noted however that, though the central tendency in economics has been to deny the cooperative spirit, there is now a body of writers who have recognized this and have even constructed models to make amends for it (see, for instance, Knack and Keefer 1997; Fehr and Gachter 2000; Fehr and Falk 2002; Hoff *et al*. 2006; Benabou and Tirole 2006; Ellingsen and Johannesson 2007).[4] And sociologists, cognitive psychologists and moral philosophers have for long written about the importance of trust and altruism among people, and how these are important for more complex relationships to thrive and for a group or a nation to progress economically (e.g. Luhman, 1979; Gambetta, 1990; Fukuyama, 1996; Hauser, 2006). That human beings have innate social and normative values is increasingly recognized in our formal social-science

[2] This remark must not be read as equating individual selfishness with methodological individualism. For one, the latter pertains to what the *analyst* does, whereas the former concerns the agent *being analysed*. It is true that, if individuals are motivated by nothing but self-interest, this *facilitates* the use of methodological individualism, but the relation does not run the other way around. I will remark later on methodological individualism but that is secondary to the paper. This paper is about social outcomes when individuals are motivated by collective concerns. What methodological label one uses to describe such an analysis is of some intellectual interest (hence my occasional remarks on this) but not germane to my analysis.

[3] This is not to deny that many interesting questions of fairness and justice can be raised within the domain of self-interested players. I explored some of this in Basu (2000). Recently, Myerson (2004) has developed the ingenious approach of modelling justice as a method of selecting equilibria in contexts where there are multiple equilibria and, left to anarchy, agents can end up in the equilibrium where everybody is worse off.

[4] One area where the cooperative spirit has been an accepted assumption is the analysis of the household and economists and sociologists have tended to take a relatively common approach (see Blumberg and Coleman 1989; Basu 2006). Zelizer (2005: 165) observes, '[The] mixture of caring and economic activity within households takes place in a context of incessant negotiation, sometimes cooperative, other times full of conflict.'

models, thanks to the new literature on 'behavioural economics'[5] (though it is a bit alarming that social-scientists needed a large literature to realize this).

In the light of this existing research, the main objective of the paper is not just to acknowledge that human beings have these traits, but to track their consequences in an area that has received little attention – the interface between altruism, identity and welfare. I will assume that the 'utils' that measure a person's welfare need not coincide with the 'payoffs' that the person seeks to maximize through his or her acts of choice. Altruism, in particular, can cause a divergence between the two.

The formal analysis begins by demonstrating a converse of the celebrated 'invisible hand theorem' of economics, which asserts that, even though each individual may be innately selfish, the collection of such selfish behaviour, mediated through the market, leads to socially optimal outcomes. I shall here argue that human beings are innately social and other-regarding and, while these traits typically aid cooperative behaviour, there are situations where, despite each individual's instinctive cooperative spirit, social optimality breaks down. In other words, the invisible hand is not always benevolent; it can work in reverse, whereby a group of innately altruistic individuals behave in a way that is collectively ruinous for them. I do not think that the villagers of Jharkhand I mentioned above are innately any different from the citizens of more prosperous and well-organized communities. They are caught in a malevolent equilibrium. The possibility of such malevolent equilibria is, in itself, well-known and can be illustrated by lots of standard games. My plan is to take this forward in two ways. It is first illustrated how such an equilibrium can be pervasive with incomplete information and how there can be a domino effect. Then I study the consequences of in-group altruism.

The paper is focused largely on positive analysis. I deliberately do not take a normative stand on the cooperative spirit. This is because the same spirit of cooperation that promotes progress can be, and in the long history of mankind there are many instances where it has been, turned against other groups, usually minorities, but also majorities that are disorganized and unable to promote their own cooperative spirit of resistance. This is an important problem of the interface between identity and altruism. When a people's altruism is confined to some in-group, this can lead to even greater oppression than oppression by selfish but atomistic individuals.

[5] See, for instance, the paper by Loewenstein and O'Donoghue (2005), which summarizes some of this, with emphasis on how our rational selves combine with our other selves to guide what we ultimately choose.

For one, group oppression allows for the free-riding of guilt among the oppressors.

2. WORLDLY CONTEXTS

In one of his recent books, Amartya Sen (2005: 335–336) asks an interesting question. Why did British investment, which came so plentifully to sectors like tea, coffee, railways and jute, of its prize colony, India, nevertheless fail to come in, in any substantial measure, into cotton textile, iron and steel? He goes on to point out that the latter were central to the old established industries of Britain in Manchester and elsewhere. But this still does not explain why the bureaucrats of the Raj, who had no direct interest in these industries, would deprive India of capital. To close the argument, Sen notes that we have to recognize that a 'general sense of social identity and priorities, which are known to play a considerable part in economic decisions in general, exerted significant influence on the pattern of British investment in India'. The British bureaucrats were working not in their self-interest but in the interest of the group that they identified with.

The converse was also true. When, in the early twentieth century, insurgency and uprisings against British rule started in many places in India, especially Bengal, there was some puzzlement on the part of the British, as was evident from the Rowlatt Sedition Committee Report of 1918, about the fact that the leaders of these insurgencies were usually from among the English-educated elites (known among the British as 'gentlemanly terrorists'), employed by the British and the ones to gain most from the persistence of British rule (Ghosh 2005). The answer once again lies in identity. These elites identified themselves more with the Indian masses than with the British rulers and were willing to make personal sacrifices in order to promote the group interest. Winston Churchill was known to be puzzled and irritated by this deficiency of narrow self-interest among the Indian elites fighting for the nation's freedom. This is not to deny that people also have strong self-interests and, with even more skilfully designed incentives, the British rulers may have been able to keep the Indian elites behaviourally loyal to the Raj. But, in fact, to design such incentives right, one has to keep in mind that self-interest is often mediated by one's collective-identity interest, and the analytics of this, as we shall presently see, can be complicated even in the simple world of the Prisoner's Dilemma.

In addition to examples from out in the world, there is now plenty of evidence from controlled experiments that people can work in their collective interests, even when that entails making personal sacrifices and, moreover, trust and altruism can be conditional on who they are interacting with, even when they are all strangers. That human beings have these additional 'moral preferences', like the desire to reciprocate

and win approval in the eyes of others and, at times, of one's own conscience is now well-documented.[6] An interesting set of experiments was conducted by Fershtman and Gneezy (2001) on students from the University of Haifa, Academic College of Tel Aviv and Tel Aviv University. They were made to play the 'Trust game', in which trust can generate wealth but it requires each player to curb his or her self-interested behaviour.

Fershtman and Gneezy found that, not only is trust widespread, but a large number of agents are willing to go all the way in trusting others so as to achieve the efficient outcome.[7] What is more, they showed that trust can be conditional on identity. Close to 60% of the individuals playing this game chose to be trusting when their opponent was of Ashkenazic origin; but only 20% chose to be trusting when the opponent happened to be of Eastern origin. Similar results of conditional trust have been reported from other experiments by other researchers (Eckel and Wilson 2002; Burns 2004).

The objective of the next sections is to take some of these ideas – of our innate cooperativeness and also our ability to vary the extent of cooperativeness depending on the identity of the person we are interacting with – to an abstract analytical model and track their consequence for efficiency and development. Under what conditions does the cooperative spirit result in cooperation? And when does cooperation break down, despite individuals having an in-built cooperativeness?

3. GAMES AND ALLEGORIES

3.1 Basic framework

Instead of assuming that human beings are selfish and they 'cooperate' only when 'cooperation' is a derivative of selfish behaviour,[8] as most economics models do, I shall here assume that the cooperative instinct is innately human. Just as self-interest creates drive and ambition, so can these other social concerns. But, more importantly, it is these other social characteristics – mainly the cooperative instinct – that provide the glue to hold society together and prepare the ground for markets to function

[6] See, for instance, the discussion by Fehr and Falk (2002). They show, interestingly, that not only are these other traits a part of the human psyche but, at times, monetary incentives can actually backfire because they can weaken one of these other motivations for human action. Our morals can also take the form of wanting to punish cheats, even when that is costly to oneself (Hoff *et al*. 2006). In an earlier work, this, coupled with the instinct to cooperate with others is described as 'strong reciprocity' (Gintis *et al*. 2003).

[7] There is now a substantial literature that reports similar findings of trust and altruism in experimental situations from around the world. See Ensminger (2000) and Heinrich *et al*. (2004).

[8] One may legitimately wonder why the word cooperation should be used in such cases.

efficiently (Granovetter 1985; Elster 1989; Arrow 1998; Nee and Ingram 1998; Platteau 2000; Basu 2000; Francois 2002). Turning this argument around, we could claim that economies can fail when the cooperative instinct breaks down. Traditional economics, rooted in methodological individualism, tends to make little room for our innate cooperative spirit and so is handicapped in commenting on its breakdown.

Digressing momentarily, I would like to point out that whether we view the present paper as breaking away from methodological individualism or not depends on how narrow a view we take of methodological individualism. The step taken here may be best thought of as a partial breakaway. If, for instance, we went further and assumed that the altruism that one individual feels for others and the intensity of it depends on how society as a whole behaves and on societal outcomes, then it would be impossible to fully describe an individual's preference without describing the behaviour of the whole group and we would clearly have moved away from the purest form of methodological individualism (Basu 2008). As Pettit (1993: 117) puts it, 'the question is whether the whole is something more than a sum of the parts or [. . .] whether the parts are transformed through belonging to the whole'. I am arguing that, in an important sense, the answer to this last question is yes.[9]

There are many different kinds of games that can be used to understand the connection between trust, altruism and identity – for instance, the Trust Game, the Ultimatum Game and the Traveller's Dilemma (Basu 2000; Bowles 2004; Heinrich *et al*. 2004). But let me here use what is, arguably, the most familiar game in the social sciences – the Prisoner's Dilemma. Moreover, the Prisoner's Dilemma has been used very elegantly by Sen (1974) to motivate the dilemma individuals face between their selfish wants and innate value judgements. The game is illustrated below in Table 1.

[9] If one took the negation of methodological individualism to be the view that there are laws pertaining to the collectivity which are *sui generis*, and how we describe the individual has to be deduced from these laws, then clearly what I am doing here has to be described as just a form of methodological individualism. However, some would take the stand that the method just described is not the complement of methodological individualism but its polar opposite – at times called 'methodological holism' (Watkins 1952) and that there are many other methods which occur between these two extremes. Pettit (1993: 165), for instance, distinguishes between two alternative ways of moving away from strict individualism – 'the vertical', which recognizes that individuals in society are 'affected from above', that is, aggregate social outcomes influence individual agency, and 'the horizontal', which admits that individuals are affected by not higher level forces but one another. A narrow definition of methodological individualism is the one taken by Arrow (1994), which allows him to argue that neoclassical general equilibrium theory is not methodologically individualistic. If the method used in this paper has to have a label it is best thought of as belonging to Pettit's category of horizontally holistic.

		Player 2	
Player 1		C	D
	C	6, 6	0, 8
	D	8, 0	3, 3

TABLE 1. The Prisoner's Dilemma.

Though its mathematical structure is standard, it will be *played* differently than in most textbooks. Hence, it needs some explanation. What is illustrated above are the dollar payoffs and I shall take it (purely for expositional ease) that each number represents an index of each person's overall well-being, for instance, units of utility or 'utils'. It is convenient to assume that utils match one-on-one with dollars. So in this game player 1 can choose between *C* and *D* and likewise for player 2. It is a useful mnemonic to think of *C* as 'cooperative behaviour' and *D* as 'defection'. If player 1 chooses *C* and 2 chooses *D* – something that can equivalently be described as 'if players 1 and 2 choose (*C*,*D*)' – then 1 earns \$0 and 2 earns \$8. If they choose (*D*,*C*), they earn 8 and 0 dollars, respectively, or (8,0), in brief. And so on. This entire information is summarized in Table 1.

The standard analysis of the game goes as follows. Note that no matter what the other player does, it is better for a player to choose *D*. Hence, the outcome will be (*D*,*D*) – both players will choose defection – and so they will earn \$3 each. It is an unfortunate outcome since they could have earned \$6 each if both chose *C*, the cooperative strategy.

In reality, people do not just maximize their own dollar incomes or even their own utilities. People typically have fellow feeling, altruism, and sense of fairness. To keep the analysis as simple as possible, I shall allow for one kind of other-regardingness in the formal analysis, which will be referred to as altruism.[10] This will be captured by assuming *as if* \$1 (or, what is the same in this paper, 1 util) earned by the other player is valued by this player as equal to α dollars of his own, where $0 \leq \alpha \leq 1$. Later I shall allow the possibility of α varying depending on who the other player is. Thus, α may be 1 for kin, $^1/_2$ for kith, and 0 for an alien; and so on. But

[10] That people do more for one another than would be dictated by purely selfish considerations is widely noted from various walks of life. Laborers typically work harder than can be explained purely in terms of their direct self-interest (Fehr and Gachter 2000; Minkler 2004). Caregivers often give more care than they are required to give in terms of their job requirements (Zelizer 2005). For a recent statement on how, in the classical writings of economists, the idea of self-interest is more capacious and accommodative of interpersonal welfare concerns than is typically made out to be, see Medema (2009).

https://doi.org/10.1017/S0266267110000192 Published online by Cambridge University Press

let us, for now, treat this as fixed. Hence, now if player 1 plays *C* and 2 plays *C*, player 1's behaviour is predicted by treating her (*effective*) *payoff* as $6 + 6\alpha$. So the game that the players actually play is described below.

		Player 2	
		C	D
Player 1	C	$6+6\alpha, 6+6\alpha$	$8\alpha, 8$
	D	$8, 8\alpha$	$3+3\alpha, 3+3\alpha$

TABLE 2. The Behavioural Payoffs.

It is possible to argue that the altruism that I feel for another person depends on the altruism he or she is expected to feel for me or how nice she is to me (Rabin 1993; Levine 1998; Gintis *et al.* 2003). Bringing in such interdependent altruism parameters would also allow us to talk about trust and other kinds of social behaviour. But I leave such complexities out of the present paper. Taking a rather novel route, Sen (1974) argued that our morals may be viewed as a meta-ordering, that is, an ordering over the orderings of all the possible outcomes of the Prisoner's Dilemma. It is interesting to see that, using $\alpha = 1$, amounts to creating a moral *ideal* ordering over the four outcomes of the game; and setting $\alpha = 0$, amounts to capturing the selfishly best (or least morally-tainted) ordering. So there is an implicit meta-binary relation suggested by the approach taken here.[11]

Two important clarifications are worth placing on record. First, one question that may arise in the reader's mind is about the meaning of selfishness. It appears at first sight that, once the α is treated as a part of a person's preference, she can, then on, be thought of being perfectly selfish, since it is *her preference* to give a weight of α to others' income. So, it seems arguable that, given her preference, she is just as selfish as a person who values only his own dollars.[12] The problem with this line of argument

[11] Beyond this we know little about the binary relation. It may be incomplete and also violate transitivity. It is arguable that when we try to rank alternatives that can be evaluated by different yardsticks, intransitivity and incompleteness are more likely to occur. We may then need to use other kinds of relational concepts such as 'being on par' or reconcile to the conundrums that arise with transitivity (Qizilbash 2002; Basu 2007). These are more likely to happen in contexts of moral binary relations. Fortunately, the route I am taking here, via the simple use of α keeps us clear from these kinds of philosophical intricacies.

[12] This refers to a much larger problem, namely, that of interpreting the payoffs in a game. We can of course write down the number that each player will earn but there is no easy way of representing what this means to the player, who may 'correct' the number psychologically to take account of fairness, altruism and so on. Not surprisingly, this problem arises more seriously in sociological games and one of the earliest discussions

https://doi.org/10.1017/S0266267110000192 Published online by Cambridge University Press

is that it reduces selfishness to a tautology; selfishness then becomes impervious to criticism.[13] To counter this, what has to be kept in mind is that, contrary to what many economists claim, it is not a tautological definition of selfishness that economists *end up* using. Economists would not have been able to derive any testable proposition if they did so. Because all behaviour would then be compatible with selfishness, the selfishness assumption would not be able to predict any behaviour.

Hence, the way I view α here is not as an innate part of a person's utility but simply as a guide to a person's *behaviour*. Indeed, it may not be a part of our preference; it could be simply that we behave *as if* we valued other people's dollars by that amount. A player's welfare or level of utility is throughout measured by the utils shown in Table 1. To prevent ambiguity, the reader may think of this as reflecting the person's *economic* well-being or welfare. What is being argued, in that case, is that people do not play to maximize their economic well-being but a hybrid of that and social and moral values, captured by α.

Consider a person who gives $1000 to a charity. It would be reasonable to say that he preferred to give this money (that would be pretty normal use of English). But would we say that he is better off by giving the $1000 to the charity? Many mainstream economists would say yes, but I would contest this and argue that the person *is* worse off, in terms of most reasonable interpretations of well-being (and, if we restrict attention to economic well-being, this is even more obvious) but that he, nevertheless, prefers to make that little sacrifice for a good cause.[14] Otherwise, 'making

of this problem occurred in Bernard (1954) – see also Swedberg (2001). Weibull (2004) encounters the same problem when analyzing the problem of interpreting results from experimental games.

[13] Some economists would counter this by pointing out how, even if we were pure revealed preference theorists, who *defined* preference by choice, this would imply certain restrictions on a person's choice function and so the theory would not be tautological. To this I have two responses. While a tautology is a zero-one concept, it is possible to develop a cogent concept of 'near tautologies' and we could argue that revealed preference theory is a near tautology. Secondly, and more powerfully, when most economists are given actual examples of violations of consistency axioms of choice functions, such as the Weak Axiom of Revealed Preference or the Chernoff condition (see Basu 2000), they tend to argue how the change in the feasible set of alternatives changes the meaning of each alternative and so there is no violation of choice behaviour over the same alternative in different situations, because the alternatives are not the same in the different situations. The point is that mainstream economics suffers from a natural human tendency – to defend one's theories to the point of rendering them unfalsifiable.

[14] In a paper focused wholly on this subject, we would distinguish between two kinds of other-regarding behaviour. When a person makes a sacrifice for her child, for instance, it is arguable that this behaviour is an extension of a person's selfishness, since a child's welfare is often internalized by us. But when one makes a contribution to some social charity or helps a person one does not know, it is arguable that this entails personal sacrifice. One does it not to gratify oneself but because one believes that one *should*

a sacrifice' would have to be deleted from our lexicons. This divergence between the index of individual well-being and what guides individual behaviour needs some getting used to since it is alien to traditional choice theory. Fortunately, there is a small literature in game theory that inclines towards this: see Weibull (2004) and Battigalli and Dufwenberg (2005).[15]

Another way to get to the same conclusion is by a slightly unusual use of the familiar mathematical method of 'proof by contradiction'. Here we do not really reach a contradiction but an unacceptable conclusion. Assume that a person's choice always exactly reflects his or her utility or welfare. Economists often try to prevent excessive government intervention by arguing that, if an exchange or trade enhances the utilities of the buyer and the seller, and has no negative fall-out on a third party, then there is no moral justification for stopping the trade or exchange (this is the Paretian argument).[16] Suppose now a politician bans the sale of houses. The standard argument that economists use against such an ill-conceived intervention is to point out that if an adult wants to sell his house and another adult wants to buy it, it is reasonable to expect that they will be better off by this, and since this is no one else's concern, this is a Pareto improvement; and so government should not ban it. But note that the politician can turn around and argue that since she is choosing to stop the sale, and choice reflects utility, the sale is no longer a Pareto improvement.

So, by this argument, no government intervention can ever be stopped on the ground that the intervention impedes a Pareto improvement, because the mere fact of the politician choosing to stop a transaction makes the transaction a non-Pareto-improvement. This somewhat absurd conclusion arises from the supposition that choice always reflects the chooser's welfare. Indeed it seems entirely plausible to me that a politician's policy choice is not something that should be equated with the politician's own utility.

To sum up, there are three indicators associated with each person – the dollars earned by her, the utility she gets and what I call her 'effective payoff'.[17] In this paper I treat the first two as the same – they are at times

this. Behaviourally the two cases may look the same but they are prompted by different internal processes and therefore would be evaluated differently when we normatively compare the outcomes. In this paper I am considering the latter kind of model for 'other-regarding' behaviour.

[15] Sen (2006: 21) discusses the standard question economists ask 'If it is not in your interest why did you choose to do what you did?' and observes: 'This wise-guy skepticism makes huge idiots out of Mohandas Gandhi, Martin Luther King, Jr., Mother Teresa, and Nelson Mandela, and rather smaller idiots out of the rest of us . . .'

[16] There are critiques that have been aimed at this – see, for instance, Sen (1983).

[17] Henceforth, a reference to payoff will mean effective payoff. And when I want to refer to a person's direct well-being (that is, the kind of numbers shown in Table 1), I shall speak of dollars or utility.

called 'material payoffs' in the literature (see Sethi and Somanathan 2001). This is an innocuous assumption, made for expositional convenience. However, I treat the third as distinct from the other two. This is a significant assumption – one that is crucial to this paper. Hence, what is being assumed is that the effective payoff numbers are guides to human behaviour. People behave as if they are maximizers of those numbers. Their well-being however is related to but distinct from those numbers. The well-being numbers are given in Table 1 and the effective payoffs shown in Table 2 are the numbers we get by making the α-based corrections to them. In a paper trying to develop criteria for moral decisions, Dietrich (2006) draws a useful distinction between (among other things) welfarism and preferencism. Using his language, moral evaluation based on Table 1 amounts to welfarism, whereas that based on Table 2 is preferencism. It must be evident that this paper is making a case for welfarism.[18]

Second, while formally what I am modelling is altruism rather than trust, it is reasonable to think of the model as an idiom for trust or other indicators of a person's sense of society. As will be evident soon (from Figure 1 below), a person's likelihood of cooperation depends on her expectation that the other person will cooperate. Hence, we could think of the player's decision as follows. If she trusts that the other person will cooperate, then she will be more inclined to cooperate. Hence, the analysis that follows, while explicitly that of altruism, could also be thought of as a model of mutual trust.

We could similarly, introduce stigma into the model by assuming that there is some stigma attached to being selfish and playing D. Of course, the person who chooses D need not be selfish but could be playing this in anticipation of the other player choosing D. But one of the functions of stigmatization, as pointed out by Gans (1972), is to scapegoat individuals in order to maintain certain norms of behaviour. Further, in a more sophisticated and realistic model we may wish to allow for the fact that the α I attach to the other player's utility would generally depend on how she achieved it. I may attach a higher α to her income if she achieves it through (C,C), than if she achieves it through (C,D). But I shall here stay away from such complications.

Let me conclude this sub-section with a digression to address a question of language that arises here and later. It was argued above that, when people are asked to play the Prisoner's Dilemma game as described

[18] The favoring of welfarism over preferencism is at times a sign of paternalism. This, however, does not apply to the kind of criterion being recommended in this paper. It is not as if a person's choice is being over-written on the ground that we, the analysts, know better what is good for the person. All that is being suggested is that, in evaluating a person's welfare, we should discount what a person does for reasons of moral and social commitments. There need be no disagreement with the person about what constitutes his or her welfare. Hence, while there may be basis to an indictment of 'maternalism' in this paper, the charge of paternalism would be unfounded.

in Table 1, they often play it differently from what standard analysis suggests. It is possible to argue, however, that it is not as if people play the Prisoner's Dilemma game differently from what textbooks suggest, but that when they play the game in Table 1, they are not playing the Prisoner's Dilemma at all but a mentally re-interpreted 'new' game, that is described in Table 2. In that case, instead of saying that people play the Prisoner's Dilemma in a non-standard way, we would have to say that they end up playing another game in a standard way. As soon as we differentiate between what denotes a person's utility and what represents a person's behaviour, this dilemma becomes unavoidable, since what the players are playing may be the Prisoner's Dilemma when we look at the utilities but quite another game when we look at the behavioural payoffs. This is the same problem that Weibull (2004) confronts and standard analysis does not have to contend with this problem because it assumes that players interpret games literally and utilities and behavioural payoffs are always identical.[19]

Given these alternative conventions of language, I prefer to say that when two players play the game in Table 1 they are playing the Prisoner's Dilemma. More generally, the name of the game will be determined by the utilities and not the behavioural payoffs. This has the advantage that the name of the game does not depend on the mental processes of the players. It is just how they play the game that changes (with how they mentally interpret the game). This nomenclature has the added advantage of not rendering the claim of Nash equilibrium play into an unfalsifiable proposition.

3.2 Homogeneous society

Suppose we have a society with n individuals and players are randomly matched with each other and made to play the Prisoner's Dilemma. Note that a society in which players manage to cooperate a lot will become richer and better off over time. And if we append to this simple model a larger economy so that people can save a part of their income (over and above what they need to consume) and earn interest on that, then a society that manages to reach the outcome (C,C) often could become many times more prosperous than a society that always reaches the outcome (D,D). If, for instance, 3 is subsistence consumption, then the latter society will,

[19] The general idea of associating each outcome of a game with more than one number for each single player (or, equivalently, the idea of having multiple payoff functions for each player), as is being done in this paper, is a useful methodological step. Even in evolutionary game theory our penchant to have one payoff function that represents both utility, which explains behaviour, and fitness, which captures the propensity of an agent or strategy to have more progenies and so have better long-run survival properties, has done us disservice. Allowing for separate representations of these would give us fewer results but those would be more robust.

presumably, have no savings, whereas the former will not only earn more, but save and become even richer in the long run.

Keeping in mind that the cooperative spirit, captured here by the altruism parameter, is natural to human beings, I want to locate conditions under which cooperation will occur and conditions where it will collapse into individualism and totally self-seeking behaviour.

Let us begin by considering the case where a player is uncertain about how her opponent will play. Suppose λ is the probability that the other player will play cooperatively, that is, choose C. Then, if this player plays C, her expected (effective) payoff, denoted by $u(C)$, will be given by:

$$u(C) = \lambda(6 + 6\alpha) + (1 - \lambda)8\alpha.$$

And, if she chooses D, her expected payoff, $u(D)$, is given as follows.

$$u(D) = \lambda 8 + (1 - \lambda)(3 + 3\alpha).$$

These are easily derived from Table 2. Hence, she will choose C if and only if $u(C) \geq u(D)$, or

$$\lambda \geq \frac{3 - 5\alpha}{1 + \alpha} \tag{1}$$

Strictly speaking, if $u(C) = u(D)$, she is indifferent between C and D. In order to keep the language of discourse simple, I am using a harmless tie-breaking assumption here, namely, that, when a person is indifferent between cooperation and defection, she chooses cooperation.

Equation 1 can be used to draw a line in an (α, λ)-space which marks the zone where a player will choose to play cooperatively. In Figure 1, the line AB is the graph of (1), with the inequality sign replaced by an equality. Hence, if, for some α, the λ happens to be on or above the line AB, then a player will choose to play C. In other words, if a player's altruism parameter, α, and her expectation that the other player will cooperate, captured by λ, are such that (α, λ) lies on or above the line AB, then and only then will she choose to cooperate.

This does not as yet tell us how this society will behave. This is because, while the society's altruism parameter may be exogenously given,[20] λ cannot be exogenous. Each individual's decision on how to play the game determines what fraction of society will play C and this determines what λ will be. Hence, we have to *derive* the value of λ.

This is easily done. If α is to the left of A, that is, $\alpha < 1/3$, then no matter what the value of λ, a person will choose D. If everybody does this, λ will in fact be 0. Likewise consider the case where α is to the right of B, that is $\alpha > 3/5$. Then, no matter what value λ takes, each player will choose C. Hence λ will be 1.

[20] In a more detailed work even this would be derived from more basic assumptions of biology and psychology.

https://doi.org/10.1017/S0266267110000192 Published online by Cambridge University Press

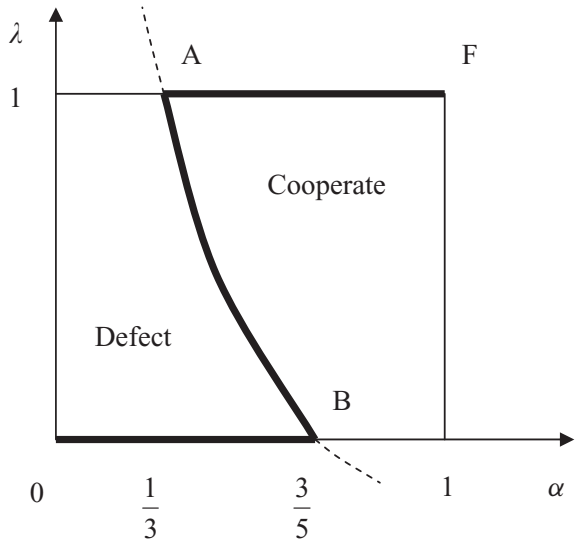

FIGURE 1. The Zone of Cooperation.

Finally, consider the case where $\frac{1}{3} \leq \alpha \leq \frac{3}{5}$. Let me use $\lambda(\alpha)$ to denote the point on the line AB. That is, $\lambda(\alpha) \equiv (3 - 5\alpha)/(1 + \alpha)$. If $\lambda > \lambda(\alpha)$, then a player confronting this λ will choose C. Since all players are identical, all will choose C under such circumstances; hence $\lambda = 1$. If, on the other hand, $\lambda < \lambda(\alpha)$, by a similar reasoning λ will be 0. In other words, we have multiple equilibria. This will lead to threshold effects and tipping behaviour, as in Granovetter and Soong (1983) and Schelling (1972), whereby behaviour can swing over from one extreme to another once it goes over a critical line. Finally, if $\lambda = \lambda(\alpha)$, then each player is indifferent between C and D. Hence, it is, in principle, possible to have λn players play C and $(1 - \lambda)n$ players play D. Hence, $\lambda = \lambda(\alpha)$ is also an equilibrium, albeit a precarious one.

Gathering the above derivations together, we have the following possible societal equilibria. If $\alpha < 1/3, \lambda = 0$. If $\alpha > 3/5, \lambda = 1$. If $1/3 \leq \alpha \leq 3/5$, then $\lambda = 0$ or 1 or $\lambda(\alpha)$. This information is summed by the correspondence illustrated by the thickened line in Figure 1, denoted by FAB0.

If we ignore the points on AB, which depict unstable equilibria (a slight perturbation will have society spiralling away to one of the two other equilibria), then we see that if altruism is very high ($\alpha > 3/5$) cooperation will be automatic. If altruism is very low ($\alpha < 1/3$), there can be no cooperation. But with intermediate altruism there are multiple equilibria. The same society can behave totally cooperatively or totally non-cooperatively. By seeing one society behaving cooperatively and

https://doi.org/10.1017/S0266267110000192 Published online by Cambridge University Press

getting richer and another that is anarchistic, selfish and poor we cannot conclude that there are innate differences between the people of these societies. It could simply be the case of both behaviours being self-sustaining in equilibrium; and so two *ex ante*-identical societies could exhibit very different kinds of outcomes.

Some useful policy wisdom emerges from the above model. What we have modelled here as altruism is part of the general idea of other-regardingness and the social spirit. There are situations in life – for instance, in starting a business – where we have to take the risk of vulnerability for the business to work. This is akin to playing *C* in the Prisoner's Dilemma. If your business partner (player 2) is cooperative (chooses *C*) you both do well, but if he betrays you, you will do badly (get 0). Hence, what this model shows is that altruism and other-regardingness are critical ingredients for a society to do well and prosper. In the present model we have treated α as exogenous. But we know at an intuitive level that people (especially children) can be taught or inspired to be more altruistic, trustworthy and generally other-regarding. Now, one person being more altruistic (having high α) would not help that person economically. In fact, he would be vulnerable to being cheated. But, if *at a societal level* all individuals were more altruistic, for instance, with α going from less than 1/3 to over 1/3, then there would be the possibility of greater cooperation and, if α went above 3/5, cooperation would occur for sure, with all the attendant economic benefits of higher income and higher utility, as shown in Table 1.

Hence, greater altruism among a people is like a public good. How exactly a government or an educational institute can create and nurture a more altruistic society we do not fully understand, but, at the same time, we do know that these traits change and can be changed. People can be taught not to litter the streets. Societies can cultivate habits of charity. Corporations can become environment conscious. Even if we do not as yet understand how these things happen, it is important to recognize that (a) unselfishness, altruism and trustworthiness are traits that are innately present in human beings and so can, potentially, be modified and nurtured, and (b) such traits are valuable for economic development and efficiency.

3.3 Heterogeneous society

All this time I have dealt with a society where all individuals have the same level of altruism. But some of the more interesting and complex issues arise when we recognize that the 'cooperative spirit', while innate, can vary across individuals.

What we are interested in understanding is what generates greater cooperative behaviour among citizens. The degree of altruism, α, is

an instrument towards this. In a homogeneous society, our aim would be to raise α if we wished to make cooperation more likely. But in a heterogeneous society the relation between the distribution of altruism and the possibility of cooperation can be complex. Interestingly, a tiny change in α can cause huge changes in behaviour. For instance, the addition of a small number of selfish individuals in a society can, like adding culture to milk, transform the character of the entire society, in this case to a non-cooperating one. Hence, the cooperative outcome can be a fragile equilibrium.

To understand this, let us suppose that person i has an altruism parameter of α_i. If we number individuals from the most selfish (person 1) to the least selfish (person n) – and clearly there is no loss of generality in this – then we have:

$$\alpha_1 \leq \alpha_2 \leq \ldots \leq \alpha_{n-1} \leq \alpha_n \tag{2}$$

An individual's altruism parameter is not visible. It will be assumed throughout that n is large; and that, when a player faces an opponent, she assumes that his altruism parameter is α_1 with probability $1/n$, α_2 with probability $1/n$ and so on.

Consider first a case where all n persons have altruism parameters in the interval $[1/3, 3/5]$. It is then easy to see that everybody playing C is an equilibrium and everybody playing D is another equilibrium. This is obvious. Since each person's α lies between $1/3$ and $3/5$ each person will choose C if he expects everybody else to choose C; and each person will choose D if he expects everybody to choose D.

What is interesting is that the introduction of one person can cause a breakdown in the cooperative equilibrium. Indeed, the introduction of *one* low-α (or high-selfishness) person can ensure that society will have a *unique* equilibrium, where nobody cooperates.

The algebra of this kind of result is rooted in the idea of 'global games' and Bayes–Nash equilibria; in different contexts a similar reasoning has been used – see, for instance, Baliga and Sjostrom (2004). The intuition is straightforward. Assume that the first t persons (i.e. persons 1 to t) prefer D over C. Now consider the $(t+1)^{\text{th}}$ person's decision problem. We know from (1), he will prefer D if

$$\lambda < \frac{3 - 5\alpha_{t+1}}{1 + \alpha_{t+1}}$$

Now, since the first t persons prefer D, the probability that a randomly chosen person will play D must be greater than or equal to t/n. Hence, the λ (probability that the other player will play C) that player $t+1$ faces is less than or equal to $(1 - \frac{t}{n})$.

https://doi.org/10.1017/S0266267110000192 Published online by Cambridge University Press

Hence, (continuing with the assumption that players 1 to t play D), player $t+1$ will certainly play D if

$$1 - \frac{t}{n} < \frac{3 - 5\alpha_{t+1}}{1 + \alpha_{t+1}}.$$

This may be rewritten as

$$\frac{2n + t}{6n - t} > \alpha_{t+1} \tag{3}$$

This is the crucial equation that can be used to show how a small injection of selfish individuals into society can cause a total breakdown in cooperation.

Here is an example. Let us start with a society of 9 individuals ranging from person 2 to person 10. So, as of now, mysteriously, there is no one called person 1. For person t in this society let α_t be equal to $(t + 19)/60$. Hence,

$$\alpha_2 = \frac{21}{60}, \alpha_3 = \frac{22}{60}, \ldots, \alpha_{10} = \frac{29}{60}.$$

As we have already seen, since in this society all α's lie between $1/3$ and $3/5$, this society can be in an equilibrium where everybody cooperates at all times.

Let another person now join this society whose altruism parameter is $19/60$. Call him person 1. That is, $\alpha_1 = 19/60$.

So now we have a 10-person society. It is easy to verify that, for every t, going from 1 to 9, (3) holds. Let us, for instance, check this for $t = 5$. Since $n = 10$, the left-hand side of (3) is $25/55$. Clearly this exceeds $\alpha_6 = 25/60$.

Next note that $\alpha_1 < 1/3$. Hence, player 1 will certainly choose D. Now, since all players, 2 to 10, (that is, $t + 1 = 2, \ldots, 10$), satisfy (3), we know that every player will strictly prefer D. Hence, this society of 10 persons has a unique equilibrium, where nobody cooperates. Though everybody's altruism parameter is unchanged, the injection of one habitual non-cooperator results in a total breakdown of cooperation. In words, the addition of a new person who is innately non-cooperative, vitiates the atmosphere for those individuals in society who were close to the borderline – that is, they needed a lot of assurance that they will not be let down before they decided to play C. These individuals now switch to playing D. This means that for the rest of society the probability of encountering a D player rises. Hence, those on the next borderline switch their choice; and so on.

It should be obvious and can be demonstrated formally with a little additional algebra that if, instead of the addition of a non-cooperative person, one existing citizen had a change of preference whereby he or she became habitually non-cooperative, this could have the same

https://doi.org/10.1017/S0266267110000192 Published online by Cambridge University Press

cascading effect. One person's change of preference can cause a change in the behaviour of all other persons in society, despite their preferences remaining unaltered.

This result is akin to what I have in a different context described (Basu 2005) as the 'malignancy of identity' whereby what may be a dormant marker of identity with no consequence on behaviour can, with a little egging on, acquire malignancy, leading to conflict between the races and different religious groups. This alerts us to the very real risk of how the injection of a small dose of new social norms or individuals carrying those different norms can create a cascading effect of change and breakdown. This must be happening nowadays with the global movement of people. And this must have happened in the heyday of colonialism, when the colonial masters arrived in new lands prepared to cooperate among themselves but not with the indigenous people. Radical writing in developing countries often talks about how the harmony of these *economically* backward societies, which nevertheless may have had a high moral code of behaviour among themselves, got disrupted by the colonial invasions. There may be an element of exaggeration and false nostalgia, and a tendency to glorify the distant past in this, but that huge disruptions in behaviour codes and social norms *can* happen is clear enough, as the above theoretical construction illustrates. Just as we now recognize that the injection of new viruses in a society can spell havoc, so can the injection of new norms. It is also conceivable that 'good norms' carried into a society by newcomers can spread through the entire society. These are subjects that will need to be studied much more fully in the future. What the above model does is to provide a few basic building blocks for such a venture.

3.4 Alcoves of altruism

Thus far, it was assumed that the altruism person i feels, she feels for everybody in her society. But, as the last discussion in the above subsection alerts us and the examples in section 2 highlight, this need not be so. People do have different ethics and altruism for in-groups and out-groups. There are many societies fractured along lines of race, gender, religion, country of origin, language identities and caste and people often show extra trust and have an altruism premium for those with whom they share some common identity (see Glaeser *et al*. 2000; Luttmer 2001).[21]

[21] The importance of identity in determining behavior has long been recognized in sociology but is relatively new in economics (Akerlof and Kranton 2000; Fryer and Jackson 2003; Hoff and Pandey 2003; Darity *et al.* 2006; Basu 2005; Iversen 2005; Sen 2006). Our identities are, however, not set in stone. The boundaries of our identities can be fuzzy and we often choose our identity and, equally, on occasions opt to overlook some of our existing identities. The reader should be warned that I take a very simplistic view of identity here

With this recognition comes the possibility of many complexities. The simplest case is where in-group trust partitions the society into different alcoves, within each of which there is trust and altruism, but these do not extend to across groups. But there can occur situations where i treats j as belonging to i's in-group, unaware that this feeling is not reciprocated. Cooperation in a nation or a group can break down when there are these cross-cutting allegiances. If a nation tries to create fellow feeling and a sense of commitment among its citizens, but a subset of citizens have allegiance to a group identity which is different from that of common citizenship, then cooperation can break down.

Moreover, in the previous sub-sections altruism was always good. But in a society that is fractured, with altruism confined to in-groups, these traits can become instruments of group oppression – where one group oppresses another, building up greater power in the oppressing group than it would have managed if the members of the group tried to carry out the oppression atomistically.

These are directions that will take a lot of time and research effort to pursue. What I will do here is to take some short, tentative steps to illustrate the scope of research that opens up once we allow altruism to be limited to those with whom a player shares a common group identity. Where this sense of identity comes from, whether it is malleable or permanent and whether it can be contained from malignancy are large topics on which much has been written[22] and much more remains to be written. I shall here treat these as primitives by simply assuming that, when playing such games, people make use of some pre-existing sense of in-group allegiance to decide how they will classify their opponents and how they will play against them.

Let me return to the assumption where α is a constant and work with the more interesting case, where $\frac{1}{3} < \alpha < \frac{3}{5}$. It is not as if I am assuming that everybody feels altruism vis-à-vis everybody, but simply that, when i feels altruism towards j, it is always at a constant altruism parameter of α. We could, in principle, allow the $\alpha's$ to differ but that would complicate the algebra unnecessarily.

To fix the idea of non-symmetric identities, let $N = \{1, \ldots, n\}$ be the set of all people and for each $i \varepsilon N$, let $G(i)$ be the set of people with whom i believes that she shares identity. The presumption is that i's altruism extends only to members of $G(i)$. In the above section, we assumed $G(i) = N$, for all $i \varepsilon N$. That is, everyone shared the same identity, which in other words means that there was no sense of group identity of any consequence. What is now being claimed is that need not always be true.

because the aim is to solely illustrate the complications this brings into our analysis of altruistic behaviour.

[22] See Tajfel 1974; Macy 1997; Turner 1999; Akerlof and Kranton 2000; Basu 2005; Sen 2006.

https://doi.org/10.1017/S0266267110000192 Published online by Cambridge University Press

Define $C \equiv \{X \subset N |$ there exists $i \varepsilon N$, such that $X = G(i)\}$. In words, C is the collection of all subsets of N, which have the property that, for each subset, there is some person who considers the subset to be exactly equal to the set of people with whom he or she identifies.

If C happens to be a partition of N, then the analysis will be virtually the same as in the above sections. Within each element of the partition, the game is played exactly as described above. If we suppose people feel altruism only for their own group members, then we could do the same analysis as in sub-section 3.2, but simply think of each group as a society. The analysis then is trivial. When people play across groups they are selfish, that is, they choose D. But within each group there could be cooperation or defection as in Section 3.2. So we could, for instance, have an equilibrium, where group A cooperates and progresses economically, whereas group B is a fractious community living in poverty.

The interesting variations occur when C is not a partition. Suppose society consists of two groups. Let a fraction γ of the population belong to group A (A can be race, caste or the fact of belonging to the same fraternity) and $(1 - \gamma)$ belong to group B. Hence, γn is the population of A and $(1 - \gamma)n$ is the population of B. In the formal language developed above, $C = \{A, N\}$, where $\{A, B\}$ is a partition of N.

So the people of group B think of A *and* B as a common identity, that is, their identity is a general national identity, whereas those in group A share an in-group identity with members of A. It could be that members of group A recognize each other because, for instance, they belong to a secret society, whereas to members in B everybody looks the same. So members of B feel altruism for all individuals in this society and cannot tell who belongs to A and who belongs to B. But members of A can tell a member of A from a non-member, and they have cultivated altruism α only towards their own group members.

Now when a type-B meets another player, the probability that the other player will cooperate is, *at most*, $(1 - \gamma)$. Hence, using the same calculation that went behind equation (1) we can see that a type-B will cooperate only if

$$1 - \gamma \geq \frac{3 - 5\alpha}{1 + \alpha}$$

$$\text{or,} \qquad \frac{6\alpha - 2}{\alpha + 1} \geq \gamma \tag{4}$$

Assume, for instance, that $\alpha = 2/5$. Then (4) gives us the condition $\gamma \leq 2/7$. Let us suppose this is true and all type Bs cooperate. Type As, on the other hand, cooperate only with their own types.

Hence in this equilibrium type As earn an expected *dollar* income of $6\gamma + 8(1 - \gamma)$ every time they play the Prisoner's Dilemma. This is

https://doi.org/10.1017/S0266267110000192 Published online by Cambridge University Press

because whenever they meet a type A (probability γ) they earn \$6 and, when they meet type B (the trusting type whom they let down), they earn \$8.

On the other hand, the expected income of type B is $6(1-\gamma)$. Hence, type As earn more than Bs. But not just that, type As, by forming this in-group collusive block, earn more than they would have earned if they cooperated with all. The latter would give them a per-game income of \$6.

There is a Machiavellian lesson tucked away in this algebra. Consider the case where $\alpha = 2/5$ and $\gamma > 2/7$. We know from (4) that type Bs will now not cooperate. It is however in the interest of type As to get them to play cooperatively, because that way they can be better 'exploited'. One way of restoring the 'exploitative equilibrium' is for type As to decide, collusively, not to play D against type Bs always, but to occasionally play C. This will enable them to delude the masses into believing that they all share one common identity and play collusively at all times. It is in fact arguable that some of the most successful exploitations of the masses rely, wittingly or unwittingly, on strategies of this kind.

One question that may arise in the reader's mind is about the general applicability of these results, since all the derivations are being done here with the example of the Prisoner's Dilemma and that too for a certain class of payoffs. This would indeed have been cause for concern if I were trying to establish general results – about what will always be true in society. Instead, the aim here is to illustrate how society *can* exhibit certain kinds of behaviour that were treated as not possible in our textbook models. We have just shown how some groups can use their innate traits of (in-group) altruism to control or even exploit other groups. It is not being claimed that this will always happen but simply that it can happen under plausible conditions. Hence, the illustration of this argument with a game that is accepted as a good model for some social situations suffices for the present context. Of course, testing the frontiers of its generalization would be an interesting exercise for the future.

3.5 Focal identity

The discussion of in-group trust draws attention to another difficulty that could arise with identity-based collusive behaviour. As we have already seen, even if people want to trust others and cooperate, one problem could arise from there being no 'focal identity' in the society. In Subsections 3.2 and 3.3 we had assumed that an entire nation shares a common identity and they are bound by a common altruism towards all (though in 3.3 one person's extent of altruism could be different from another's). In 3.4 we saw cases where there could be conflicting identities and this could lead to a subset of society playing cooperatively.

https://doi.org/10.1017/S0266267110000192 Published online by Cambridge University Press

One variant of this problem can lead to a total failure of cooperation in society. It is of course well-recognized that we have multiple identities and this can often (in fact, I believe, more often than not) help hold societies together (Dahrendorf 1959; Sen 2005). But this can also lead to a failure of cooperation. To see this, suppose people in a country resolve to be cooperative among those with whom he or she share their primary identity. But if this society lacks a focal identity or has overlapping identities instead of partitioned identities, cooperation may fail to occur in equilibrium.

To see this, suppose in a nation there are two races, 1 and 2, two religions, 1 and 2, and two language groups, 1 and 2. Using notation in an obvious way, we can describe a person as (1,2,1) or (2,2,1) and so on, where (1,2,1) means a person of race 1, religion 2 and language 1. Let me use A to denote the set of all people of type (1,2,1), B to denote all of type (1,1,2) and C to denote (2,1,1). Assume $\frac{1}{3}$ of the population is of type A, $\frac{1}{3}$ of type B and $\frac{1}{3}$ of type C.

Let us now assume that all As think that race is the primary identity (that is, they try to be cooperative with all and only those who share their race), all Bs think that religion is the primary identity and all Cs think that one's mother-tongue is the primary identity. In this society, each person will find that at least $\frac{1}{3}$ of the times they will have the other player choose defect.

Hence, we can see that if α is less than $\frac{1}{2}$, the right-hand term in (1) is greater than $\frac{1}{3}$. Since in this society λ is below $\frac{1}{3}$, by (1) we know that no one will play cooperatively. Thus, even if every player has $\alpha = \frac{2}{5}$, no cooperation will occur in this society. The reason for this is the lack of a focal identity.

This has the policy implication that if a government or some collectivity wants to encourage cooperative behaviour in the country or among its members, it must try to create a focal identity among its citizens. Conversely, various repressed groups that fail to rise collectively against their oppressors probably do so for the reason that they do not have a focal identity among themselves. This is an equally useful result for a tyrant or malevolent government trying to prevent some group or nation from acting cooperatively within itself. The aim of the tyrant must be to destroy the group's ability to form a focal identity. Through a deliberate policy of splintering the group's identity into various overlapping and conflicting identities it can keep the group under control and keep at bay the possibility of group rebellion. If you can break up a large group into a partition of smaller groups, that can be useful in foiling rebellion. But if you can destroy the large group's focal identity by nurturing *overlapping* identities you can do even more damage to the large group. This is the reason why analyses of this kind can be both useful *and* dangerous.

4. REMARKS

The model above is best treated as an allegory of the real world. Nevertheless, it talks to us about policy and, like all science, does so whether our aims are noble or mean. It tells us how to prosper economically and gives hints and suggestions for people trying to cooperate among themselves and escape oppression, and also for people wanting to cooperate in order to oppress others, not belonging to their group. It shows, for instance, that one way to exploit a large mass of people is to form a collusive sub-group the members of which identify primarily with the sub-group but deludes the large mass into believing that it identifies totally with the large mass. Of course, and mercifully, the effort of the sub-group can be foiled by there being other sub-groups trying to do the same. If too many opportunistic groups come into existence, society could crumble into the low-output equilibrium of selfish anarchy.

A central lesson that comes out from this allegory and one that contrasts sharply with popular wisdom concerns the ubiquitous 'invisible hand'. The 'invisible hand theorem', which has come down to us from Adam Smith,[23] and was discussed in Section 1, has had enormous influence in shaping economic policy and has been prominent in the advice that various think tanks and organizations, not to mentions legions of economists, have given to developing country governments. One inadvertent implication of the theorem that many have taken away from it and that has had considerable influence on the organization of our economic and social life and also in the way we conduct ourselves is that it is fine to be selfish, since in the end that is good for society.[24] This selfishness axiom has in recent times spilled over into other disciplines, such as sociology and the new political science.

As a consequence, we are taught that not only are consumers and producers necessarily self-seeking but so are politicians, bureaucrats and judges; and, more significantly, that that is fine. This has some alarming consequences. It means that all we can expect of a judge is for verdicts that best serve his or her own interest. And so the only way to make judges and magistrates give just verdict is to design the institutional and incentive structure of the courts in such a way that it is in each judge's self-interest to be just.

[23] As a digression on attribution, note that, though modern social scientists treat the 'invisible hand' as the central message of Smith's *Wealth of Nations*, it is in reality a trivially small part of that book, and occurs when dealing with international trade. Smith had used the expression earlier, but in a different sense, in his *Theory of Moral Sentiments* (1759) and even earlier in *History of Astronomy*, which was however published posthumously.

[24] This is what makes the occasional dissenting voice refreshing: see Rubinstein (2006b).

This ubiquitous philosophy has been damaging not only socially and morally but even in terms of economic growth and development, because the truth about development is that it needs human beings to be other-regarding, fair, and trustworthy. And since these traits are innately available to most of us, what we need is not to have them muted through training and socialization. Take the problem of bureaucratic corruption, which has been eating into the fabric of so many societies, and blighting the possibility of development. The standard policy response to this, inspired by the popularity of the invisible hand theorem and the very visible global economists, is to argue that government ought to redesign the system of incentives and punishments for bureaucrats. What we do not say is that the ubiquity of corruption has a lot to do with the lack (or, more appropriately, suppression) of personal integrity and individual moral commitments. The design of incentives plays a role, but a bigger role is played by our own sense of values and morals. Governments which are non-corrupt are largely so not because of third-party monitoring of such corruption but because of the self-monitoring of bureaucrats. There is no scope for this in standard economics because it provides little space to *self*-monitoring.

Hence, there is no reason to believe that countries with rampant corruption are populated by citizens who are innately less moral; but simply that they *act* less morally in equilibrium. This is related to the findings from the celebrated experiments by Frank *et al*. (1993). They showed that in games where one can be selfish to different degrees, economists play the most selfishly. There are different ways of interpreting the result but I take the view that, since economists learn from their textbooks that everybody is selfish and it is fine to be selfish, they, like all human beings, try to conform to what they take to be the standard behaviour (see also Rubinstein 2006a).[25] In corrupt environments, people begin to treat corruption as the norm (moreover deviating from that norm also has larger costs than in more honest environments) and, like economists in the above-mentioned experiments, try to replicate what they take to be normal behaviour.[26]

[25] It is conceivable though that in experimental and examination-like situations people give the answers they feel are expected of them, and so these findings merely reflect the disciplinary training of economists; and that, in reality, the behaviour of economists would be no different from that of others.

[26] This brings us back to the methodological point made earlier. What is being claimed is that it is not possible to fully understand human behaviour without explicit recognition of the collectivity in which the human being happens to be situated. Pettit (2002) calls this the method of 'social holism'. According to this, the situatedness is an integral part of each human being. As he puts it (Pettit 2002: 117), 'As no one can be a sibling without having or having had a brother or a sister, so no one can be a proper human being, according to this claim, without enjoying or having enjoyed the presence of others in his or her life'.

The starkest examples of this one sees in the streets of developing countries. With drivers willing to break every rule and showing a relentless commitment to serving their own interests and with very little presence of the traffic warden, the streets of the developing nations should be textbook models of neoclassical efficiency. The fact that they are not should alert us to the possibility that the central message of many of our textbooks may be wrong.

The truth is that human beings are not relentlessly selfishness – though they can learn to be so if it is drilled into them that that is normal or they grow up in societies caught in an ethos of selfish behaviour. If we want society to progress and economic development to occur, we need to nurture our innate sense of social values – such as altruism, integrity, and fairness.

REFERENCES

Akerlof, G. and R. Kranton 2000. Economics and identity. *Quarterly Journal of Economics* 115: 715–753.

Arrow, K. J. 1994. Methodological individualism and social knowledge. *American Economic Review* 84: 1–10.

Arrow, K. J. 1998. The place of institutions in the economy: a theoretical perspective. In *The Institutional Foundations of East Asian Economic Development*, ed. M. Aoki and Y. Hayami. Basingstoke: Macmillan.

Baliga, S. and T. Sjostrom 2004. Arms races and negotiations. *Review of Economic Studies* 71: 351–369.

Basu, K. 2000. *Prelude to Political Economy: A Study of the Social and Political Foundations of Economics*. Oxford and New York: Oxford University Press.

Basu, K. 2005. Racial conflict and the malignancy of identity. *Journal of Economic Inequality* 3.

Basu, K. 2006. Gender and say: A model of household decision-making with endogenous balance of power. *Economic Journal* 116.

Basu, K. 2007. Coercion, contract and the limits of the market. *Social Choice and Welfare* 29: 559–579.

Basu, K. 2008. Methodological individualism. In *The New Palgrave Dictionary of Economic*, L. Blume and S. Durlauf. New York: Palgrave.

Battigalli, P. and M. Dufwenberg 2005. *Dynamic Psychological Games*. mimeo: Bocconi University and University of Arizona.

Benabou, R. and J. Tirole 2006. Incentives and prosocial behavior. *American Economic Review* 96: 1652–1679.

Bernard, J. 1954. The theory of games of strategy as a modern sociology of conflict. *American Sociological Review* 59: 411–424.

Blumberg, R. and M. Coleman 1989. A theoretical look at the gender balance of power in the American couple. *Journal of Family Issues* 10: 225–250.

Bowles, S. 2004. *Microeconomics: Behavior, Institutions and Evolution*. Princeton, NJ: Princeton University Press.

Burns, J. 2004. *Race and Trust in Post Apartheid South Africa*. mimeo.

Dahrendorf, R. 1959. *Class and Class Conflict in Industrial Society*. Stanford, CA: Stanford University Press.

Darity, W. A. Jr., P. L. Mason and J. B. Stewart 2006. The economics of identity: The origin and persistence of racial identity norms. *Journal of Economic Behavior and Organization* 60: 283–305.

Dietrich, F. 2006. *Welfarism, Preferencism, Judgementism*. mimeo: University of Maastricht.
Eckel, C. C. and R. K. Wilson 2002. *Conditional Trust: Sex, Race and Facial Expressions in a Trust Game*. mimeo: Virginia Tech.
Ellingsen, T. and M. Johannesson 2008. Pride and prejudice: The human side of incentive theory. *American Economic Review* 98: 990–1008.
Elster, J. 1989. *The Cement of Society*. Cambridge: Cambridge University Press.
Ensminger, J. 2000. Experimental economics in the bush: How institutions matter. In *Institutions and Organizations*, ed. C. Menard. London: Edward Elgar.
Fehr, E. and A. Falk 2002. Psychological foundations of incentives. *European Economic Review* 46: 687–724.
Fehr, E. and S. Gachter 2000. Cooperation and punishment in public goods experiments. *American Economic Review* 90: 980–994.
Fershtman, C. and U. Gneezy 2001. Discrimination in a segmented society: An experimental approach. *Quarterly Journal of Economics* 116: 351–377.
Francois, P. 2002. *Social Capital and Economic Development*. New York: Routledge.
Frank, R. H., T. Gilovich and D. T. Regan 1993. Does studying economics inhibit cooperation? *Journal of Economic Perspectives* 7: 159–171.
Fryer, R. and M. Jackson 2003. *Categorical Cognition: A Psychological Model of Categories and Identification in Decision Making*. mimeo: Harvard University.
Fukuyama, F. 1996. *Trust: The Social Virtues and the Creation of Prosperity.* New York: Free Press.
GambettaD. (ed.) 1990. *Trust: The Making and Breaking of Cooperative Relations*. Oxford: Blackwell.
Gans, H. 1972. The positive functions of poverty. *American Journal of Sociology* 78: 275–288.
Ghosh, D. 2005. *Terrorism in Bengal: Political Violence in the Interwar Years*. mimeo: Cornell University, Department of History.
Gintis, H., S. Bowles, R. Boyd and E. Fehr 2003. Explaining altruistic behavior in humans. *Evolution and Human Behavior* 4: 153–172.
Glaeser, E., D. Laibson, J. Scheinkman and C. Soutter 2000. Measuring trust. *Quarterly Journal of Economics* 115: 811–846.
Granovetter, M. 1985. Economic action and social structure: the problem of embeddedness. *American Journal of Sociology* 91: 481–510.
Granovetter, M. and R. Soong 1983. Threshold models of diffusion and collective behavior. *Journal of Mathematical Sociology* 9: 165–179.
Hauser, M. D. 2006. *Moral Minds*. New York: Harper Collins.
Heinrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr and H. Gintis 2004. *Foundations of Human Sociality: Economic and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press.
Hoff, K. and P. Pandey 2003. *Why are Social Inequalities so Durable? An Experimental Test of the Effects of Indian Caste on Performance*. mimeo: The World Bank, Washington.
Hoff, K., M. Kshetramade and E. Fehr 2006. *Norm Enforcement under Social Discrimination*. mimeo: World Bank.
Iversen, V. 2005. *Segmentation, Network Multipliers and Spillovers: A Theory of Rural Urban Migration for a Traditional Economy*. mimeo: University of East Anglia.
Knack, S. and P. Keefer 1997. Does social capital have an economy payoff? A cross-country investigation. *Quarterly Journal of Economics* 112: 1251–1288.
Levine, D. K. 1998. Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics* 1: 593–622.
Loewenstein, G. and T. O'Donoghue 2005. *Animal Spirits: Affective and Deliberative Processes in Economic Behavior*. mimeo: Carnegie Mellon University.
Luhman, N. 1979. *Trust and Power*. Chichester: Wiley.
Luttmer, E. 2001. Group loyalty and the taste for redistribution. *Journal of Political Economy* 109: 500–528.

Macy, M. W. 1997. Identity, interest and emergent rationality: An evolutionary synthesis. *Rationality and Society* 9: 427–448.
Medema, S. 2009. *The Hesitant Hand: Taming Self-Interest in the History of Ideas*. Princeton, NJ: Princeton University Press.
Minkler, L. 2004. Shirking and motivations in firms: Survey evidence on worker attitudes. *International Journal of Industrial Organization* 22: 863–884.
Myerson, R. 2004. Justice, institutions, and multiple equilibria. *Chicago Journal of International Law* 5: 91–107.
Nee, V. and Ingram, P. 1998. Embeddedness and beyond: Institutions, exchange, and social structure. In *The New Institutionalism in Sociology*, ed. M. Brinton and V. Nee. New York: Russell Sage Foundation.
Pettit, P. 1993. *The Common Mind: The Essay on Psychology, Society, and Politics*. New York: Oxford University Press.
Pettit, P. 2002. *Rules, Reasons and Norms*. Oxford: Clarendon Press.
Platteau, J.-P. 2000. *Institutions, Social Norms, and Economic Development*. Amsterdam: Harwood Academic Publishers.
Qizilbash, M. 2002. Rationality, comparability and maximization. *Economics and Philosophy* 18: 141–156.
Rabin, M. 1993. Incorporating fairness into Game Theory and economics. *American Economic Review* 83: 1281–1302.
Rothschild, E. 2001. *Economic Sentiments: Adam Smith, Condorcet, and the Enlightenment*. Cambridge, MA: Harvard University Press.
Rubinstein, A. 2006a. A skeptic's comment on the study of economics. *Economic Journal* 116: C1–C9.
Rubinstein, A. 2006b. Dilemmas of an economic theorist. *Econometrica* 4: 865–883.
Schelling, T. 1972. A process of residential segregation: Neighborhood tipping. In *Racial Discrimination in Economic Life*, ed. A. H. Pascal. Lexington, MA: D.C. Heath.
Sen, A. 1974. Choice, orderings and morality. In *Practical Reasoning*, ed. S. Korner. Oxford: Blackwell.
Sen, A. 1983. Liberty and social choice. *Journal of Philosophy* 80: 5–28.
Sen, A. 2005. *The Argumentative Indian: Writings on Indian History, Culture and Identity*. London: Penguin Books.
Sen, A. 2006. *Identity and Violence: The Illusion of Destiny*. New York: Norton & Co.
Sethi, R. and E. Somanathan 2001. Preference evolution and reciprocity. *Journal of Economic Theory* 97: 273–297.
Smith, A. 1759 (1976). *The Theory of Moral Sentiments*. Indianapolis: Liberty Classics.
Smith, A. 1776 (1976). *An Inquiry into the Nature and Causes of the Wealth of Nations*. Oxford: Clarendon Press.
Swedberg, R. 2001. Sociology and Game Theory: Contemporary and historical perspectives. *Theory and Society* 30: 301–335.
Tajfel, H. 1974. Social identity and intergroup behavior. *Social Science Information* 13: 65–93.
Turner, J. C. 1999. Some current issues in research on social identity and self-categorization theories. In *Social Identity*, ed. N. Ellemers, R. Spears and B. Doosje. Oxford: Blackwell.
Watkins, J. W. N. 1952. The principle of methodological individualism. *British Journal for the Philosophy of Science* 3: 186–189.
Weibull, J. 2004. *Testing Game Theory.* In *Advances in Understanding Strategic Behaviour: Game Theory, Experiments and Bounded Rationality*, ed. S. Huck. London: Palgrave MacMillan.
Zelizer, V. 2005. *The Purchase of Intimacy*. Princeton, NJ: Princeton University Press.