

Trust and Cooperation among Economic Agents*

by

Partha Dasgupta**

University of Cambridge
and
University of Manchester

Revised: June 2009

* Text of lecture delivered at the Royal Society Discussion Meeting on 19/20 January 2009 on *The Evolution of Society*, in celebration of the 200th birth anniversary of Charles Darwin, contributions to which are forthcoming in the *Philosophical Transactions of the Royal Society, B*. I am grateful to the referees for their suggestions on an earlier draft of the paper, and to Kenneth Arrow, Scott Barrett, Patrick Bateson, Paul Ehrlich, and Robert Hinde for discussions over the years on the matter of trust.

** Frank Ramsey Professor of Economics, University of Cambridge; Fellow of St John's College, Cambridge; and Professor of Environmental and Development Economics, University of Manchester.

Abstract

The units that are subject to selection pressure in evolutionary biology are "strategies", which are conditional actions ("Do P if Q occurs"). In contrast, the units in economics select strategies from available menus so as to further their projects and purposes. As economic agents don't live in isolation, each agent's optimum choice in general depends on the choices made by others. Because their projects and purposes involve the future, not just the present, each agent reasons about the likely present and future consequences of their respective choices. That is why *beliefs*, about what others may do and what the consequences of those choices could be, are at the basis of strategy selection. In this article I construct a catalogue of social environments in which agents not only promise one another cooperation, but rationally believe that the promises will be kept. Unfortunately, non-cooperation arising from mistrust can be the outcome in those same environments: societies harbour multiple "equilibria" and can skid from cooperation to non-cooperation. Moreover, a pre-occupation among analysts with the Prisoners' Dilemma game has obscured the fact that cooperative arrangements can harbour not only inequality, but exploitation too. The analysis is used to discuss why international cooperation over the use of global public goods has proved to be so elusive.

Keywords: negotiation, trust, cooperation, culture, rational beliefs, Nash equilibrium, tipping points, inequality, exploitation, self-enforcing treaties, Kyoto Protocol, Montreal Protocol.

Contents

Introduction

1 Trust

2 Credible Promises

2.1 Mutual Affection

2.2 Pro-social Disposition

3 Incentives to Keep Promises

3.1 External Enforcement

3.2 Reputation as a Capital Asset

3.3 Long-term Relationships

4 Dark Matter: Breakdown of Cooperation

5 More Dark Matter: Exploitation in Long-Term Relationships

6 International Cooperation

7 Conclusions

Introduction

The units that are subject to selection pressure in evolutionary biology are "strategies" (Maynard Smith, 1982; Nowak, 2006), which are conditional actions, such as, "Do P if Q occurs". In contrast, the units in economics *select* strategies from available menus so as to further their projects and purposes. As agency assumes a central role in the social sciences, economic units are called "agents", or "parties". Sometimes, we economists even call them "people".

Robinson Crusoe aside, people don't live in isolation. So, an agent's optimum choice depends on the choices made by others. Moreover, as their projects and purposes involve not just the present but the future too, every agent reasons about the likely present and future consequences of their respective choices, while recognising that all others are engaged in similar reasoning. That is why *beliefs*, about what others may do and what the consequences of those choices could be, are at the basis of strategy selection. Economic environments are therefore inter-temporal games, in the sense of the theory of games (Binmore and Dasgupta, 1986).

Notice, we are not pre-judging that agents cooperate so as to improve their lot. Whether they do depends on the ease with which they have access to a cooperative "infrastructure" (e.g., commitment devices that can be used to make the promises agents make to one another credible, Section 3). In this paper I study the various social environments in which cooperation is possible. That the basis of cooperation is mutual trust is a banality, the deeper point is that trust in turn is based on beliefs. However, if the trust is not to be "blind", it has to be based on *rational* beliefs. In Sections 1-3 I develop these arguments in a sequence of increasing complexity. In Section 4 we study why cooperation is often very fragile. In Section 5 I show that cooperation among members of a group is not always benign, that it can harbour inequality, even exploitation. In Section 6 I apply the theoretical framework of Sections 3 and 4 to ask why, in contrast to the cooperation that is frequently observed among members of local communities over the use of geographically confined natural resources (Dasgupta and Heal, 1979; Ostrom, 1990; Baland and Platteau, 1996) and among people engaged in transactions in well-functioning markets, international cooperation in the management of global public goods (e.g., the atmosphere as a sink for pollutants, the oceans) has proved to be so elusive. Section 7 concludes.

1 Trust

Imagine that a group of people have discovered a mutually advantageous course of actions. At the grandest level, it could be that citizens see the benefits of adopting a Constitution for their country. At a more local level, the undertaking could be to share the costs and benefits of maintaining a communal resource (irrigation system, grazing field, coastal fishery); construct a jointly useable asset (drainage channel in a watershed); collaborate in political activity (civic engagement, lobbying); do business when the purchase and delivery of goods can't be synchronized (credit, insurance, wage labour); enter marriage; create a rotating saving and credit association (as in the institution of *iddir* in Ethiopia); initiate a reciprocal arrangement (I help you, now that you are in need, with the understanding that you will help me when I am in need); adopt a convention (send one another Christmas cards); create a partnership to produce goods for the market; conduct an instantaneous transaction (purchase something across the counter); and so on. Then there are mutually advantageous courses of action that involve being civil to one another. They range from such forms of civic behaviour as not disfiguring public spaces and obeying the law more generally, to respecting the rights of others.

Imagine next that the parties have agreed to share the benefits and costs in a certain way. The agreement could involve some members making side-payments to others. Again, at the grandest level the agreement could be a social contract among citizens to observe their Constitution. Or it could be a tacit agreement to be civil to one another, such as respecting the rights of others to be heard, to get on with their lives, and so forth. Here we will be thinking of agreements over transactions in goods and services. There would be situations where the agreement was based on a take-it-or-leave-it offer one party makes another (as when a purchaser accepts the terms and conditions in a supermarket). In other contexts, bargaining may have been involved (as in a Middle-Eastern bazaar). Here we will not ask how agreements have been reached, nor look for principles of equity that might have been invoked during negotiation (but see Section 5). We ask instead: *Under what circumstances would the parties who have reached agreement trust one another to keep their word?*

Because one's word must be credible if it is to be believed, mere promises wouldn't be enough. (Witness that we caution others, and ourselves

too, not to trust people "blindly".) If the parties are to trust one another to keep their promise, matters must be so arranged that: (1) at every stage of the agreed course of actions, it would be in the interest of each party to plan to keep his or her word if all others were to plan to keep their word; and (2) at every stage of the agreed course of actions, each party would believe that all others would keep their word. If the two conditions are met, a system of beliefs that the agreement will be kept would be self-confirming.

Notice that condition (2) on its own wouldn't do. Beliefs need to be justified. Condition (1) provides the justification. It offers the basis on which everyone could in principle believe that the agreement will be kept. A course of actions, one per party, satisfying condition (1) is called a *Nash equilibrium*, in honour of the mathematician John Nash (he of *the beautiful mind*) who proved that it is not a vacuous concept (Nash, 1950). By their very definition, Nash equilibria (there can be more than one equilibrium; see below) are *self-enforcing*, which is why the parties in question would seek to identify them.

Notice that condition (1) on its own wouldn't do either. It could be that it is in each agent's interest to behave opportunistically if everyone believed that everyone else would behave opportunistically. In that case non-cooperation is also a Nash equilibrium, meaning that a set of mutual beliefs that the agreement will not be kept would also be self-confirming, and so, non-cooperation would be self-enforcing. Stated formally, a Nash equilibrium is a set of strategies, one per agent, such that no agent would have any reason to deviate from his or her course of actions if all other agents were to pursue their courses of actions. As we have just seen, generally speaking societies harbour multiple Nash equilibria (see Maynard Smith, 1982; Nowak, 2006; and Osborne, 2004, for specific examples). Some yield desirable outcomes, others do not. The famous Prisoners' Dilemma is a game that has a unique Nash equilibrium in which all parties are worse off than they could have been if a suitable cooperative infrastructure had been in place (Section 6). The fundamental problem facing a society is to create institutions where conditions (1) and (2) apply to engagements that protect and promote its members' interests.

Conditions (1) and (2), taken together, require an awful lot of coordination among the parties. In order to probe the question of which Nash equilibrium can be expected to be reached, if a Nash equilibrium is expected to be reached at all, economists study human behaviour that are *not* Nash

equilibria. The idea is to model the way people form *beliefs* about the way the world works, the way people behave in response to those beliefs, and the way they revise their beliefs on the basis of what they observe. The idea is to track the consequences of those patterns of belief formation so as to check whether the model economy moves toward a Nash equilibrium over time, or whether it moves about in some fashion or other but not toward an equilibrium. Because beliefs (and their revisions) play a strong role in the evolution of cooperation among humans, evolutionary dynamics in economic environments involves a somewhat different set of drivers from the ones that are studied in evolutionary biology.¹

Theoretical research on the evolution of beliefs and the concomitant evolution of strategies has yielded one general conclusion: Suppose the economic environment in a certain place harbours multiple Nash equilibria. Which equilibrium should be expected to be approached, if the economy approaches an equilibrium at all, will depend on the beliefs that people held at some point in the past. It also depends on the way people have revised their beliefs on the basis of observations since that past date. This is another way of saying that history matters. Model building, statistical tests on data relating to the models, and historical narratives have to work together synergistically if we are to make progress in understanding our social world. Unfortunately, the study of disequilibrium behaviour would lengthen this paper greatly. We shall see though that, fortunately, a study of equilibrium behaviour takes us a long way.

2 Credible Promises

We began by observing that mutual trust is the basis of cooperation. In view of the multiplicity of Nash equilibria and the possible awfulness of equilibria in those social environments where a cooperative infrastructure is absent, we look for environments in which cooperation is possible. To do that it proves useful to classify the social environments in which the promises people make to one another are *credible*. Five come to mind (Dasgupta, 2000,

¹ Fudenberg and Levine (1998) and Evans and Honkapohja (2001) are key theoretical treatises on the evolution of beliefs in social systems. Axelrod and Hamilton (1981) and Nowak (2006) offer the evolutionary biologist's account of the emergence of cooperation in animal populations. Beliefs play no role there.

2005, 2007).²

2.1 Mutual Affection

Promises would be credible if the parties care about one another sufficiently. Innumerable transactions take place only because the people involved care about one another and rationally believe that they care about one another (each knows that the others know that they care about one another, each knows that the others know that each knows that they care about one another, and so on) and thus trust one another to carry out their obligations. The household best exemplifies institutions based on care and affection. (The corresponding notion in evolutionary biology is "kin selection"; Hamilton, 1964. Humans of course don't confine their affection to their kin.) Because people who cohabit are able to observe and know one another, they can be sanguine that members will not be unduly opportunistic. The problem is that, being few in number, members of a household, as a group, are unable to engage in those enterprises that require large numbers of people of varied talents and locations. That is why mutual affection is not the basis of cooperation in most other contexts.

2.2 Pro-social Disposition

Promises would be credible if it was common knowledge that those making the promises were trustworthy, or that they reciprocated by keeping their promise if others displayed trust in them. The new behavioural economics emphasises this aspect of human character (see, e.g., Rabin, 1993; Fehr and Fischbacher, 2002). Nature and nurture play a still little-understood combined role in developing in us a general disposition to reciprocate (Hinde and Groebel, 1991; Ehrlich, 2000). Our capacity to have such feelings as shame, affection, anger, elation, obligation, benevolence, and jealousy would appear to have emerged under selection pressure. No doubt culture helps to shape preferences, expectations, and thus, behaviour, which are known to differ widely across societies. But cultural coordinates enable us to identify the locus of points *upon* which shame, affection, anger, elation, obligation,

² The five-way classification that follows does not presume that the social environments in question are distinct. For example, mutual affection, pro-social disposition, reputation and mutual enforcement (see below in the text) overlap in many contexts. I offer the classification nonetheless because, for conceptual purposes, it is useful to regard them as distinct. Lehmann and Keller (2006) have offered a related classification in evolutionary biology. I remark on their work below in the text.

benevolence, and jealousy are put to work; they don't displace the centrality of those capacities in the human make-up. The thought I am pursuing here is that as adults we not only have a disposition for such behaviour as paying our dues, helping others at some cost to ourselves, and returning a favour, we also practise such norms as those which prescribe that we punish those who have hurt us intentionally; and even such higher-order-norms as shunning those who break agreements, on occasion frowning on those who socialise with people who have broken agreements; and so forth. Often enough, the disposition to be honest would be toward members of some particular group, not others. This amounts to group loyalty. The underlying group could be one's neighbours, or clan, or nation. The glue that binds could also be religion or ethnicity (Ehrlich, 2000, has an excellent discussion on these matters).

By internalizing specific norms, a person enables the springs of his actions to include them. He therefore feels shame or guilt in violating the norm, and this prevents him from doing so, or at the very least it puts a break on him, unless other considerations are found by him to be overriding. In short, his upbringing ensures that he has a disposition to obey the norm, be it moral or social. When he does violate it, neither guilt nor shame would typically be absent, but frequently the act will have been rationalized by him. For such a person, making a promise is a commitment, and it is essential for him that others recognise it to be so (Arrow, 1974).

Although trustworthiness isn't alien to human nature, people don't have their inherent trustworthiness stamped on their forehead. So they can't be expected to know in advance whom to trust. In any event, if relative to the gravity of the misdemeanour the pecuniary benefits of opportunistic behaviour were high, transgression could be expected. The problem is that one wouldn't know in advance who would be likely to transgress. Punishment assumes its role as deterrence because of these agency problems. As someone's trustworthiness isn't publicly observable, punishment is usually tailored to the "crime". In the next section we study the remaining three contexts in which people are able to trust one another to keep their promises. We will confirm that, by looking into someone's personal history it becomes possible to tailor punishment not only to the "crime", but also their past behaviour and circumstances.

3 Incentives to Keep Promises

The promises the parties have made to one another to keep to their

agreement would be credible if they could devise an institution in which keeping promises would be in the interest of each party if everyone else were to keep them. The problem therefore is to devise an institution in which keeping to the agreement is a Nash equilibrium. Recall that a strategy is a sequence of conditional actions. Strategies assume the forms, "I shall choose X if you choose Y", or "I shall do P if Q occurs", and so on. If promises are to be credible, it must be in the interest of those making promises to carry them out if and when the relevant occasions arise.

Societies everywhere have constructed solutions to the credibility problem, but in different ways. What all solutions have in common, however, is the imposition of collective sanctions on those who intentionally do not comply with agreements. Of course, a credible threat of punishment for misdemeanours would be an effective deterrence only if future costs and benefits aren't discounted at too high a rate relative to other parameters of the social environment, a matter to which I return presently.

Broadly speaking, there are three types of situation where parties to an agreement could expect everyone to keep to their words. (Of course, none may be potent in a particular context, in which case people would find themselves in a hole they cannot easily get out of, and what could have been mutually beneficial agreements will not take place. (The behaviour reported in the *Mezzogiorno* by Banfield, 1958, is an illustration of this possibility.) Each gives rise to a set of institutions that capitalize on its particular features. In practice, of course, the types would be expected to shade into one another, but it pays to study them separately. So, in the next three sub-sections I assume that the discount rates agents apply to their future costs and benefits are low relative to other parameters of the social environment.

3.1 External Enforcement

It could be that the agreement is translated into an explicit contract and enforced by an established structure of power and authority; that is, an external enforcer.

By an external enforcer I imagine here, for simplicity, the State. (Depending on the social environment, the "external enforcer" could be the tribal chieftain, the warlord, the priest, or the village elders.) Consider that the rules governing transactions in the formal market-place are embodied in the law. So markets are supported by a legal structure. Firms, for example, are legal entities. Even when you go to a supermarket, your purchases (paid in

cash or by card) involve the law, which provides protection for both parties (the grocer, in case the cash is counterfeit or the card is void; the purchaser, in case the product turns out on inspection to be sub-standard). The law is enforced by the coercive power of the State. Transactions involve legal contracts backed by an external enforcer, namely, the State. It is because you and the supermarket owner are confident that the State has the ability and willingness to enforce contracts that you and the owner of the supermarket are willing to transact.

What is the basis of that confidence? After all, the State apparatus is run by people, which means a further agency problem. In any event, the contemporary world has shown that there are States and there are States. Simply to invoke an external enforcer for solving the credibility problem won't do. For why should the parties trust the State to carry out its tasks in an honest and effective manner? A possible answer is that the government worries about its reputation (Section 3.2). So, for example, a free and inquisitive press in a democracy helps to sober the government into believing that incompetence or malfeasance would mean an end to its rule when the time comes for the next election. Because voters know that the government worries, they trust their government to enforce agreements. Even if senior members of the ruling party are getting on in years and don't much care what happens in the future, younger members would worry that the party's reputation would suffer if the government were not to behave.

The above argument involves a system of interlocking beliefs about one another's abilities and intentions. Consider that millions of households in many parts of the world trust their government (more or less!) to enforce contracts, because they know that government leaders know that not to enforce contracts efficiently would mean being thrown out of office. In their turn, each side of a contract trusts the other not to renege (again, more or less!), because each knows that the other knows that the government can be trusted to enforce contracts. And so on. Trust is maintained by the threat of punishment (a fine, a jail term, dismissal, or whatever) for anyone who breaks a contract. We are in the realm of equilibrium beliefs, held together by their own bootstraps.

Unfortunately, cooperation isn't the only possible outcome. Non-cooperation can also be held together by its own bootstrap. At a non-cooperative equilibrium the parties don't trust one another to keep their

promises, because the external enforcer cannot be trusted to enforce agreements. To ask whether cooperation or non-cooperation would prevail is to ask which system of beliefs is adopted by the parties about one another's intentions. Social systems harbour multiple equilibria.

3.2 Reputation as Capital Asset

Political parties are not the only entities that view reputation as a capital asset. Individuals and firms view it that way too. Consider someone who doesn't care what his reputation will be after death. Even he would care to build a reputation for honest dealing if by so doing he could cash in that reputation at the time of retirement. Brand names are an instance of such cases. The person owning the brand name no doubt changes over time, but the name itself remains. Consider a firm whose dishonest behaviour has been exposed. Suppose too that customers deal only with firms that have an unsullied reputation. On retirement, the owner would find no buyer for the firm. If the owner knew that in advance, she may well wish to maintain the firm's reputation for honesty. If the owner cared sufficiently about her quality of life after retirement, honesty would be an equilibrium strategy, just as boycotting ill-reputed firms would be a corresponding equilibrium strategy for customers (Kreps, 1990).

Of course, even in situations where reputation can be accumulated as a capital asset, it may be that agents don't accumulate reputations for honesty. It cannot be repeated often enough that social systems possess multiple equilibria.

The formal analysis of reputation as capital asset is similar to one where the parties expect to face transaction opportunities repeatedly in the future. Let us study those situations.

3.3 Long-term Relationships

Suppose the agents expect to face similar transaction opportunities in each period over an indefinite future. Imagine too that the parties can't depend on the law of contracts because the nearest courts are far from their residence. There may even be no lawyers in sight. In rural parts of sub-Saharan Africa, for example, much economic life is shaped outside a formal legal system. But even though no external enforcer may be available, people there do transact. Credit involves saying, "I lend to you now with your promise that you will repay me"; and so on. But why should the parties be sanguine that the agreements won't turn sour on account of opportunistic behaviour?

They would be sanguine if agreements were *mutually* enforced. The basic idea is this: a credible threat by members of a community that stiff sanctions would be imposed on anyone who broke an agreement could deter everyone from breaking it. (The corresponding mechanism in evolutionary biology is called "reciprocal altruism"; Trivers, 1971.) The problem then is to make the threat credible. The solution to the credibility problem in this case is achieved by recourse to social norms of behaviour.

By a *social norm* we mean a rule of behaviour, or a strategy, that is followed by members of a community. For a rule of behaviour to *be* a social norm, it must be in the interest of everyone to act in accordance with the rule if all others were to act in accordance with it. Social norms are (Nash) equilibrium rules of behaviour.

To see how social norms work, imagine that the gain to a party from breaking the agreement unilaterally during a period is less than the discounted value of the losses she would suffer if all other parties were to punish her subsequently. The punishment could involve all others refusing to engage in any transactions with the erring party in the following period, shunning her for suitable numbers of periods, and so on. Call a party "conformist" if she cooperates with parties who are conformists but punishes those who are non-conformists. That sounds circular, but it isn't, because the social norm we are studying here requires all parties to start the process by keeping their agreement. It would then be possible for any party in any period to determine which party is conformist and which party is not. For example, if ever someone were to break the original agreement, he would be judged to be non-conformist; so, the norm would require all parties to punish the non-conformist. Moreover, the norm would require that punishment be inflicted not only upon those in violation of the original agreement (first-order violation); but also upon those who fail to punish those in violation of the agreement (second-order violation); upon those who fail to punish those who fail to punish those in violation of the agreement (third-order violation); and so on, indefinitely. This infinite chain makes the threat of punishment for errant behaviour credible because, if all others were to conform to the norm, it would not be worth any party's while to violate the norm. Keeping one's agreement would then be self-enforcing.³

³ The literature on repeated games is huge. Mailath and Samuelson (2006) is the definitive treatise on the subject and contains a comprehensive list of references to original papers.

All traditional societies appear to have sanctions in place for first-order violations. Anthropologists and novelists have noted the use of sanctions for second-order violations. The fact that sanctions against higher-order violations haven't been documented much may be because they aren't needed to be built into social norms if it is commonly recognised that people feel a strong emotional urge to punish those who have broken agreements. Anger facilitates cooperation by making the threat of retaliation credible.⁴

Social norms that are enshrined in the culture of a community depend not only on the character of the agreements themselves, but also on the relative ease with which prospects are expected to arise for opportunistic behaviour. Sanctions can range from the punitive and unforgiving ("one strike and you are out!" - known in the literature as the "grim strategy"), which have been observed in places where tempting short-term outside economic opportunities appear from time to time. However, many rural communities (e.g. in the mountains of Nepal) are like enclaves: they live far from established markets. Adopting "grim" would prove counter-productive there. That is why sanctions there have been found to be graduated: the first misdemeanour is met by a small punishment, subsequent ones by stiffer punishments, persistent ones by punishments that are stiffer still (Ostrom, 1992). Where information is imperfect, a small penalty for the first misdemeanour would be warning that others were watching, or it could be that others signal their acknowledgement that the misdemeanour could have been an error on the part of the offender and that he should try harder next time. And so on.

It can be shown that the scope for cooperation can be increased by *tying* several agreements (e.g., agreements over the mutual provision of credit, insurance, and labour, respectively), so that the norm has it that violation of any one agreement is met by withdrawal of cooperation in all other engagements (Dasgupta, 2007). When separate agreements (whether among

⁴ On a riverboat ride in Australia's Kakadu National Park some years ago, my wife and I were informed by the guide, a young aborigine, that his tribe traditionally practiced a form of punishment that involved spearing the thigh muscle of the errant party. When I asked him what would happen if the party obliged to spear an errant party were to balk at doing so, the young man's reply was that he in turn would have been speared. When I asked him what would happen if the person obliged to spear the latter miscreant were to balk, he replied that he too would have been speared! I asked him if the chain he was describing would go on indefinitely. Our guide said he didn't know what I meant by "indefinitely", but as far as he knew, there was no end to the chain.

the same set of individuals or among different groups of individuals) are tied, the long-run benefits of cooperation become larger than the (short-run) gains from defection even at larger values of the rates at which individuals discount their future benefits than they would be if agreements were not tied. So, tying agreements makes cooperation robust against defection. Interestingly, tied relationships are a common feature of traditional societies in the contemporary world (Baland and Platteau, 1996). Greif (2006) has argued that tied relationships among Maghreb merchants who were engaged in long-distance trade fuelled economic growth in medieval southern Europe.

Unfortunately, even when cooperation is a possible equilibrium, non-cooperation is an equilibrium too. To see why, imagine that each party believes that all others will renege on the agreement. It would then be in each one's interest to renege at once, meaning that there would be no cooperation. Failure to cooperate could be due simply to an unfortunate pair of self-confirming beliefs, nothing else. No doubt it is mutual suspicion that ruins their chance to cooperate, but the suspicions are internally self-consistent. In short, even when people don't discount future costs and benefits at a high rate and appropriate institutions are in place to enable people to cooperate, it can be that they do not cooperate. Whether they cooperate depends on mutual beliefs, nothing more. I have known this result for many years, but still find it a surprising and disturbing fact about social life

Remark: In their review of the theoretical literature on the emergence of cooperation and altruism in behavioural ecology and game theory, Lehmann and Keller (2006) report that the models assume that one or more of the following four conditions need to be fulfilled (provided of course that the benefits of cooperation exceed the short run gains from defection): (i) direct benefits to the individual to the focal individual performing a cooperative act; (ii) direct or indirect information allowing a better than random guess about whether a given individual will behave cooperatively in repeated reciprocal interactions; (iii) preferential interactions between related individuals; and (iv) genetic correlation between genes coding for altruism and phenotypic traits that can be identified. Notice that (i) is pre-supposed in the latter four social environments in our five-way classification; (ii) is involved in all five of our social environments (but, obviously, in different manners); while (iii) and (iv) are implicit in, respectively, the first and second social environments in our five-way classification.

4 Dark Matters: Breakdown of Cooperation

We have so far assumed that the discount rates people apply to their future gains and losses are small. It is, of course, obvious that if the rates were large, cooperation wouldn't be possible (the present discounted value of the flow of future private benefits of cooperation would fall short of the short term gains from defection). So we now have in hand a tool to explain how a community where members have been cooperating can skid to a state of affairs where they cease to cooperate. Ecological stress (caused, for example, by high population growth and prolonged droughts) often leads people to fight over land and natural resources (Homer-Dixon, 1999; Diamond, 2005). More generally, political instability (in the extreme, civil war) would be a reason why people discount the future benefits of cooperation at a high rate, if for no other reason than a heightened fear that their community will not survive in its present shape. For whatever reason, if discount rates were to increase sufficiently relative to the parameters characterising the social environment, cooperation would cease. Mathematicians call the points at which those switches occur, "bifurcations", sociologists call them "tipping points". Social norms work only when people have reasons to value the future benefits of cooperation.

Contemporary examples illustrate this. Local institutions have been observed to deteriorate in the unsettled regions of sub-Saharan Africa. Communal management systems that once protected Sahelian forests from unsustainable use were destroyed by governments keen to establish their authority over rural people. But Sahelian officials had no expertise at forestry, nor did they have the resources to observe who took what from the forests. Many were corrupt. Rural communities were unable to switch from communal governance to governance based on the law: the former was destroyed and the latter didn't really get going. The collective vacuum has had a terrible impact on people whose lives had been built round their forests and woodlands (Dasgupta, 2008a).

Ominously, there are subtler pathways by which societies can tip from a state of mutual trust to one of mutual distrust. We have seen that when discount rates are low, both cooperation and non-cooperation are equilibrium outcomes. So, a society could tip over from cooperation to non-cooperation simply because of a change in beliefs. The tipping may have nothing to do with any discernable change in circumstances; the entire shift in behaviour could

be triggered in people's minds. The switch could occur quickly and unexpectedly, which is why it would be impossible to predict and why it would cause surprise and dismay. People who woke up in the morning as friends would discover at noon that they are at war with one another. Of course, in practice there are usually cues to be found. False rumours and propaganda create pathways by which people's beliefs can so alter that they tip a society where people trust one another to one where they don't.

The reverse can happen too, but it takes a lot longer. Rebuilding a community that was previously racked by civil strife involves building trust. Non-cooperation doesn't require as much coordination as cooperation does. Not to cooperate usually means to withdraw. To cooperate, people must not only trust one another to do so, they must also coordinate on a social norm that everyone understands. That is why it's a lot easier to destroy a society than to build it.

How does an increase or decrease in cooperation translate into macroeconomic performance? Consider two communities that are identical in all respects, excepting that in one people have coordinated at an equilibrium state of affairs where they trust one another, while people in the other have coordinated at an equilibrium where they don't trust one another. The difference between the two economies would be reflected in the productivity of their assets, which would be higher in the community where people trust one another than in the one where they don't. Enjoying greater incomes, individuals in the former economy are able to put aside more of their income to accumulate capital assets, other things being equal. So it would become relatively wealthier. Mutual trust would be interpreted from the statistics as a driver of economic growth, but the statistics wouldn't reveal how that trust was created and maintained.

5 More Dark Matter: Exploitation in Long-Term Relationships

Both theory and empirics tell us that cooperation can harbour inequality (see Dasgupta, 2008a, for a review). Unhappily, it can also harbour exploitation, a far worse state of affairs. We began by considering a group of people who have not only discovered a *mutually* beneficial course of actions, but have also agreed to follow that course. We identified circumstances in which people would be able to enter long-term relationships in which they would trust one another to do what they are required to under the terms of the agreement. In studying long-term relationships, we assumed that all who enter

them benefit (although not perhaps equally). I now want to show that long-term relationships can be *bad* for some members of a cooperative; in that there are circumstances where some people are worse off being part of a long-term relationship than they would have been if a long-term relationship had not been entered into.

If that sounds implausible, it may be because in studying cooperation and the benefits that accrue from it, we are used to drawing on the Prisoners' Dilemma (PD) game. Indeed, the PD game has been used almost universally to illustrate the problem of collective action people face in producing public goods (e.g. flood barriers) or managing common property resources (e.g. local woodlands). However, societal problems involving the production of public goods and the management of common property resources (CPRs) are *not* reflected in the PD game (Dasgupta and Heal, 1979; Dasgupta, 2008b).

To see why, recall that the PD game, when played once, has two distinguishing features: (1) Every agent has a dominant strategy, that is, a strategy that is best for her *no matter* what strategies are chosen by the others; and (2) there is a set of strategies, one for each agent, which, if it were chosen, would lead to an outcome that is better for all than the outcome that obtains when each agent plays her dominant strategy. In the absence of a cooperative infrastructure, agents would choose their dominant strategies. So, the dominant strategies constitute the unique Nash equilibrium. Moreover the equilibrium is collectively sub-optimal. That's the dilemma.

Recall from the theory of games that an agent's *min-max* payoff is the payoff she can guarantee for herself even if all others were bent on making her life as miserable as possible. The reason we are interested in the concept of min-max payoffs in the present context is that in the PD game agents receive their min-max payoffs when they play their dominant strategies. So, at the unique Nash equilibrium of the PD game, agents' payoffs are their min-max payoffs.

Dasgupta and Heal (1979: Ch. 3) showed that in the absence of a cooperative infrastructure the production of public goods and the management of CPRs involve games in which (the unique) Nash equilibrium is indeed collectively sub-optimal; but the authors also showed that agents' equilibrium payoffs are *not* their min-max payoffs. This means that in public-goods games and CPR games there is a gap between the equilibrium payoff of an agent and her min-max payoff: *the former exceeds the latter*. Now consider a long-term

relationship among people engaged in the management of a CPR. Suppose the social norm they use in order to maintain cooperation instructs members to punish non-conformists by pushing them down to their min-max payoffs for a suitable number of periods. Using results from the theory of repeated games (e.g., Mailath and Samuelson, 2006), Dasgupta (2000, 2008b) showed that if the agents discount their future payoffs at a low enough rate, such a norm would support outcomes in which the payoffs over time to some agents are less than their respective Nash equilibrium payoffs, but in excess of their min-max payoffs. Those unfortunate agents accept the conditions of the long-term relationship only because not to do so would mean that they are driven down to their min-max payoffs for an extended period of time. Plainly, those agents would have been better off if there had been no long-term relationship. This is the sense in which cooperation can involve exploitation.

Thus far, theory. Unearthing exploitation from data will prove to be fiendishly difficult, because they would involve answering a counterfactual question: what would life have been like for those who are suspected of being exploited had long-term relationships not been entered into? Nevertheless, there are informal grounds for thinking that long-term relationships can give rise to exploitation. Under the caste system in India, for example, “untouchables” in rural areas are frequently barred from drawing water from the village well, whose use is restricted to caste Hindus. And there are many other similar restrictions on “untouchables” besides. To be sure, “untouchables” are members of the village community, but each group has its assigned role. Can one prove that “untouchables” are exploited in village India, in the precise sense in which I am using the term? Probably not, but the theory I am appealing to is suggestive. And that’s a virtue of the theory.

6 International Cooperation

Several of the pre-conditions for cooperation would be found to be missing if we consider the prospects of international cooperation in the management of global public goods (e.g., the global climate). In Sections 2-4 we assumed that the parties have discovered a mutually advantageous course of actions and have reached an agreement over the way the costs and benefits are to be shared. Sadly, that can't be assumed in the international context.

Consider international negotiations over climate change. Nations (by which, of course, I mean national leaders) differ greatly in their assessment of the costs and benefits to them from continuing increases in carbon

concentration: some nations are small, while others are large; some are rich, while others are poor; some are in the tropics, others in temperate zones; some are governed by leaders who take science seriously, others are less fortunate; and so on. Side payments would be needed if all nations were to sign a treaty, but the promise of such payments may not be credible. For a treaty to be believable, it must be self-enforcing.

Among the possible outcomes of international negotiations over climate change is the "null-treaty", meaning global non-cooperation, commonly referred to as "business as usual". Moreover, it can be that the negotiations harbour more than one self-enforcing treaty. Treaties would differ in their efficiency and in the distribution of benefits and burdens among nations. Carraro (2002), Barrett (2003), and Dutta and Radner (2004), among others, have shown that not all countries should be expected to sign a potential treaty on climate change. Some (among them many small countries) would free ride. Among the choices to be made in designing a treaty are adaptation and mitigation measures. The costs and benefits involving the two kinds of investment would be expected to differ among countries. So, economists who study the political economy of climate change face the problem of having to explain which equilibrium would be selected. Factors outside theoretical models would be particularly relevant here. The power of rich countries could be expected to tilt the selection toward their favour.

As the Kyoto Protocol didn't lay the groundwork for a self-enforcing treaty on climate change (Barrett, 2003; Dutta and Radner, 2004), it has been a failure. On the other hand, the Montreal Protocol on the emission of chlorofluorocarbons (CFCs) has been a success. Why? Barrett (2003) has argued that if, relative to the costs of curbing emissions, the perceived benefits are large, it is possible for large numbers of nations to reach agreement. In effect, very little in the way of side-payments needs to be made in order for signatories to enjoy the benefits. This was the case with curbing CFCs. Carbon emissions are a problem of a different order of magnitude. The costs of controlling emissions to any significant degree are huge, while the benefits of controlling them are likely to be diffuse. Unlike radiation arriving through holes in the Ozone layer, global climate change doesn't kill people in a direct, identifiable, and immediate way. It is easy to go into denial over climate change.

Barrett (2008) observes that although in discussions on global climate

change it is frequently claimed that adaptation and mitigation are complementary activities, they are more like substitutes. As countries invest more in the former, they suffer less from climate change and find mitigation less attractive. But mitigation is a global public good ("windmills"), whereas adaptation is a national public good ("dikes"). One can imagine a situation where the globally optimal investment policy would have every country invest in windmills, but where under non-cooperation each nation constructs only dikes. Imagine that the ideal international treaty (with appropriate, credible side payments) sustains a high level of participation and requires so many windmills to be built that no one needs to construct dikes. Barrett constructs examples where, nevertheless, the treaties that are signed are ones under which rich countries construct dikes and pollute the atmosphere, leaving poor countries not so much high and dry, as "low and wet". Such an ominous possibility cannot yet be ruled out.

7 Conclusions

In this article I have identified five social environments where cooperation is possible. They range from environments where people care about one another, to those where people are to a greater or lesser extent self-seeking but laws and/or social norms are in place to make cooperation self-enforcing. The bad news is that in all but the social environment where the fact that people care about one another is common knowledge, non-cooperation is also self-enforcing. Societies harbour multiple equilibria. The *beliefs* people hold about one another and about the way behaviour translates into social consequences would appear to be central to the possibilities of cooperation. Alarmingly, societies can tip from cooperation to conflict because of a mere change in beliefs. Our analysis showed why it is a lot easier for a society to destroy itself than to re-build. Creating trust is no easy matter. I have also shown that long-term relationships, which can sustain cooperation, have a dark side to them. They can not only sustain inequality among people engaged in cooperation; they can involve exploitation too.

We used the our findings on the possibilities of cooperation to explain why international cooperation over the use of global public services, such as the ecological services that are provided by the atmosphere and the stratosphere, has proved to be so uneven: the Montreal Protocol over the emission of CFCs was a success, but the Kyoto Protocol over carbon emissions was a failure. The answer would seem to be that the social infrastructures that

are necessary for cooperation are all too fragile in the international sphere. Unlike the Montreal Protocol, the Kyoto Protocol did not form the basis of a self-enforcing treaty. The prospects that humanity will be able to contain carbon concentrations in the atmosphere within reasonable limits are not large.

References

- Arrow, K.J. (1974), *The Limits of Organization* (New York: W.W. Norton).
- Axelrod, R. and W.D. Hamilton (1981), "The Evolution of Cooperation", *Science*, 211 (27 March), 1390-1396.
- Baland, J.-M. and J.-P. Platteau (1996), *Halting Degradation of Natural Resources: Is There a Role for Rural Communities?* (Oxford: Clarendon Press).
- Banfield, E. (1958), *The Moral Basis of a Backward Society* (Chicago: Free Press).
- Barrett, S. (2003), *Environment & Statecraft: The Strategy of Environmental Treaty-Making* (New York: Oxford University Press).
- Barrett, S. (2008), "Dikes vs. Windmills: Climate Treatise and Adaptation", Discussion Paper, The Johns Hopkins University.
- Binmore, K. and P. Dasgupta (1986), "Game Theory: A Survey", in K. Binmore and P. Dasgupta, eds., *Economic Environments as Games* (Oxford: Basil Blackwell).
- Carraro, C. (2002), "Climate Change Policy: Models, Controversies, and Strategies", in T. Tietenberg and H. Folmer, eds., *The International Yearbook of Environmental and Resource Economics 2002/2003* (Cheltenham: Edward Elgar).
- Dasgupta, P. (2000), "Economic Progress and the Idea of Social Capital", in P. Dasgupta and I. Serageldin, eds., *Social Capital: A Multifaceted Perspective* (Washington DC: World Bank).
- Dasgupta, P. (2005), "The Economics of Social Capital", *The Economic Record*, 2005, 81(255: Supplement), S2-S21.
- Dasgupta, P. (2007), *Economics: A Very Short Introduction* (Oxford: Oxford University Press).
- Dasgupta, P. (2008a), "The Role of Nature in Economic Development", in D. Rodrik and M. Rosenzweig, eds., *Handbook of Development Economics, Vol. 5* (Amsterdam: North Holland), forthcoming 2009.
- Dasgupta, P. (2008b), "Common Property Resources: Economic Analytics", in R. Ghate, N.S. Jodha, and P. Mukhopadhyay, eds., *Promise, Trust, and Evolution: Managing the Commons of South Asia* (Oxford: Oxford University Press).

Dasgupta, P. and G. Heal (1979), *Economic Theory and Exhaustible Resources* (Cambridge: Cambridge University Press).

Diamond, J. (2005), *Collapse: How Societies Choose to Fail or Survive* (London: Allen Lane).

Dutta, P.K. and R. Radner (2004), "Self-enforcing Climate Change Treatise", *Proceedings of the National Academy of Sciences*, 101(14), 5174-5179.

Ehrlich, P.R. (2000), *Human Natures: Genes, Culture, and the Human Prospect* (Washington, DC: Island Press).

Evans, G.W. and S. Honkapohja (2001), *Learning and Expectations in Macroeconomics* (Princeton: Princeton University Press).

Fehr, E. and U. Fischbacher (2002), "Why Social Preferences Matter: The Impact of Non-selfish Motives on Competition, Cooperation and Incentives", *Economic Journal*, 112(478), C1-33.

Fudenberg, D. and D.K. Levine (1998), *The Theory of Learning in Games* (Cambridge, MA: MIT Press).

Greif, A. (2006), *Institutions and the Path to the Modern Economy: Lessons from Medieval Trade* (New York: Cambridge University Press).

Hamilton, W.D. (1964), "The General Evolution of Social Behavior", *Journal of Theoretical Biology*, 7(1), 1-55.

Hinde, R.A. and J. Groebel, eds. (1991), *Cooperation and Prosocial Behaviour* (Cambridge: Cambridge University Press).

Homer-Dixon, T.E. (1999), *Environment, Scarcity, and Violence* (Princeton, NJ: Princeton University Press).

Kreps, D. (1990), "Corporate Culture and Economic Theory", in J.E. Alt and K.A. Shepsle, eds., *Perspectives on Positive Political Economy* (New York: Cambridge University Press).

Lehmann, L. and L. Keller (2006), "The Evolution of Cooperation and Altruism: a general framework and classification of models", *Journal of Evolutionary Biology*, 19, 1365-1378.

Mailath, G. and L. Samuelson (2006), *Repeated Games and Reputation: Long-Run Relationships* (New York: Oxford University Press).

Maynard Smith, J. (1982), *Evolution and the Theory of Games* (Cambridge: Cambridge University Press).

Nash, J.F. (1950), "Equilibrium points in N -person games", *Proceedings of the National Academy of Sciences*, 36(1), 48-49.

Nowak, M.A. (2006), *Evolutionary Dynamics: Exploring the Equations of Life* (Cambridge, MA: The Belkamps Press).

Osborne, M.J. (2004), *An Introduction to Game Theory* (New York: Oxford University Press).

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

Ostrom, E. (1992), *Crafting Institutions for Self-Governing Irrigation Systems* (San Francisco, CA: International Center for Self-Governance Press).

Rabin, M. (1993), "Incorporating Fairness into Game Theory and Economics", *American Economic Review*, 83(5), 1281-1302.

Trivers, R.L. (1971), "The Evolution of Reciprocal Altruism", *Quarterly Review of Biology*, 46(1), 35-57.