THE SOCIALLY OPTIMAL LEVEL OF ALTRUISM

Richard Povey, St Edmund Hall

Thesis submitted in partial fulfilment of the requirements for the degree of D.Phil in Economics in the Department of Economics at the University of Oxford, Trinity Term 2011.

March 12, 2012

Word count: 50812

The Socially Optimal Level of Altruism

Richard Povey, St Edmund Hall, University of Oxford

Trinity Term, 2011

Abstract

It is already recognized by some specific models in the existing literature that altruism may have socially counterproductive effects. Economic theory also shows that self-interest often produces efficient outcomes. This thesis explores the relationship between altruistic preferences, punishment systems and the cultural evolution of morality. The central argument is that altruism has detrimental effects on the efficacy of punishment and the resultant incentives of agents to co-operate with socially efficient equilibria, and that the use of punishment can have a negative effect on the evolution of altruism.

The sequential punishment model is presented – akin to an infinitely-repeated stage game, but sufficiently simple to allow determinate optimal punishment paths to be derived – and the impact of different levels of altruism fully analysed. It is shown that high levels of altruistic motivation – close to but slightly less than full altruism – cause the socially efficient equilibrium to break down.

Although the model is only a highly stylized representation of social interaction, the key effects that drive these results should appear in many more specific examples. In summary, these are the temptation effect (more altruistic individuals are less tempted to do harm to others), the willingness effect (more altruistic individuals are less willing to inflict punishment), and the severity effect (punishments, such as a fine where the revenue is redistributed, are less severe for more altruistic individuals, because they place a higher value on the contribution of the revenue to the welfare of others).

By embedding a simplified version of the sequential punishment model in a simulated indirect evolution framework, it is also established that the use of punishment can weaken the group selection mechanism, resulting in a lower level of altruism evolving. The normative consequences of this are shown to be ambiguous.

Acknowledgements

I would like to express my deep gratitude to the United Kingdom Economic and Social Research Council and to the Royal Economic Society for funding this research.

This thesis would never have been completed without the help and support of a number of people. My supervisor, Kevin Roberts, has offered me patient, careful and insightful advice on numerous matters throughout my time as a graduate student at Oxford, for which I am extremely grateful. I was initially encouraged to embark upon postgraduate study by my parents, Valerie Brind and Robert Povey, and inspired by my tutors as an undergraduate at St Edmund Hall, particularly Outi Aarnio and Stephen Blamey.

I am also indebted to Godfrey Keller, who gave me extensive and helpful comments on an earlier draft of this thesis. Alan Beggs and Michael Griebe are others who have been most generous in reading and offering their views on my work on multiple occasions. My examiners, Chris Wallace and Peter Hammond, also gave me invaluable and detailed feedback on earlier submissions. I would like to thank as well my graduate advisers at St Edmund Hall - John Knight, Martin Slater and Linda Yueh.

The teaching positions to which I was fortunate enough to be appointed during my time at Oxford were also crucial in enabling me to complete this research. My colleagues at St Edmund Hall mentioned above offered me my first college lectureship. I am also particularly grateful to Howard Smith at Keble College, Lucia Nixon, Petra Schleiter and Anita Avramides at St Hilda's College, Alan Beggs, Scott Sturgeon and Paul Martin at Wadham College, and Ian Crawford, Richard Mash, Richard Whittington and Martin Ceadel at New College, for being such supportive teaching colleagues.

Last, but most certainly not least, I want to thank Scot Peterson for his unfailing love and encouragement throughout.

Contents

List of Figures

The Limits to Altruism - A Survey

> There is much confusion of the ideal that a person ought to be allowed to pursue his own aims with the belief that, if left free, he will or ought to pursue solely his selfish aims.
>
> (Hayek, 1960)

> In my view the ideal society would be one in which each citizen developed a real split personality, acting selfishly in the market place and altruistically at the ballot box.
>
> (Meade, 1973)

1.1 *Overview*

This survey is concerned with altruism, and its relationship to economic theory. The key tasks are to summarise and marshal the evidence that altruism exists as a significant empirical phenomenon, and to make a contribution to the normative analysis of altruism by bringing together a number of theoretical strands from economics and other social sciences. The central questions to be considered are why particular patterns of altruistic motivation and behaviour exist, and which conceivable patterns would be most beneficial. The answers to these two questions are closely connected if there is a theoretical presumption that cultural norms emerge at least partly due to their social functionality. We clearly live in a world of imperfect altruism. The central question is *why* this should be the case. Are individuals imperfectly altruistic because we live in an imperfect world, or could it be that these imperfections are somehow functional for society?

The appropriate role of human altruism in social scientific theory and methodology is subtle and controversial. Firstly, altruism can be defined in numerous incompatible ways, particularly when we compare the approaches of economists, anthropologists, sociologists, psychologists and biologists. A central factor which commonly causes confusion here is the distinction between altruism as a concept for classifying observed behaviour, and altruism as a theory of motivation. There are also important distinctions between economic and psychological theories of altruistic motivation, due to their differing criteria for what constitutes an explanation for a particular phenomenon.

A second key area of debate is the normative analysis of altruism. The temptation that some approaches have fallen into is to assume that altruism on the part of individuals is always good for the group, and thus that minimising individual selfishness is the most important role for the social structure. Economics has, at least since Adam Smith, had a fairly clear understanding that a key litmus test for the effectiveness of a social and economic system is its ability to correct for, and harness, the limitations upon individual altruism (Smith, 1976). This is just as important as ensuring that individual selfishness is kept constrained within certain boundaries. For example, the effective rule of law is required for a well-functioning market economy, and this requires that individuals refrain from harming others indiscriminately whenever they can gain selfishly by doing so. In other words, selfishness must, in order to act as a force for social good, be constrained to operate within a broader moral framework. However, the relationship between the positive assumption of rational self-interest as a parsimonious explanatory framework, and the normative argument that self-interest is a force that can be harnessed for the common good, has remained inadequate and blurred.

1.2 *Altruism: A Road Map*

If we wish to achieve clarity in an exploration of the past and future potential role of altruism in economic theory specifically, we must have a clear conception of the distinctions and connections between a normative concept of social efficiency (the common good), models of individual altruistic motivation (which are explanatory rather than normative), and a way of classifying the degree of altruism exhibited in various behaviours based upon observable criteria. The fact that apparently altruistic behaviour does not necessarily imply altruistic motivation and that altruistic motivation does not necessarily imply social efficiency suggests eight logical possibilities for the interpretation of a particular behaviour:

(1). Apparently altruistic behaviour which is altruistically motivated and is socially beneficial.

(2). Apparently altruistic behaviour which is not altruistically motivated and is socially beneficial.

(3). Apparently non-altruistic behaviour which is altruistically motivated and is socially beneficial.

(4). Apparently non-altruistic behaviour which is not altruistically motivated and is socially beneficial.

(5). Apparently altruistic behaviour which is altruistically motivated and is socially detrimental.

(6). Apparently altruistic behaviour which is not altruistically motivated and is socially detrimental.

(7). Apparently non-altruistic behaviour which is altruistically motivated and is socially detrimental.

(8). Apparently non-altruistic behaviour which is not altruistically motivated and is socially detrimental.

This system of categorization is useful in maintaining a separation between the behavioural, motivational and normative dimensions to altruistic phenomena. The different cases are illustrated graphically in figure 1.1. The models examined later in the survey can be fitted into it. The way in which the categories are applied, however, will vary depending on one's theoretical perspective.

Welfare economic theory already has a well-defined set of tools for judging when changes in behaviour are socially beneficial and socially detrimental, provided one is philosophically willing to make inter-personal comparisons of utility from consuming different goods (Sen, 1974) (d'Aspremont & Gevers, 1977) (Gevers, 1979) (Roberts, 1980). The concepts of a behaviour being "apparently altruistic" and "altruistically motivated" are, however, more problematic. In asking whether a behaviour appears to be altruistic, we must ask who or what appears to be paying the cost, and who is benefiting. Dawkins' selfish gene theory (Dawkins, 1976), for example, seeks to categorise all cases of apparent altruism by individual organisms as the manifestation of the entirely selfish adaptive "behaviour" of genes. In this case, there is strictly speaking no apparent or actual altruism at the gene level. So, if we are taking the "gene's eye view", then cases (1), (2), (3), (5), (6) and (7) drop out. The only question that remains to be answered in the analysis of a particular behavioural phenotype is therefore whether the interaction of competition of selfish genes leads to a "socially beneficial" or "socially detrimental" outcome.

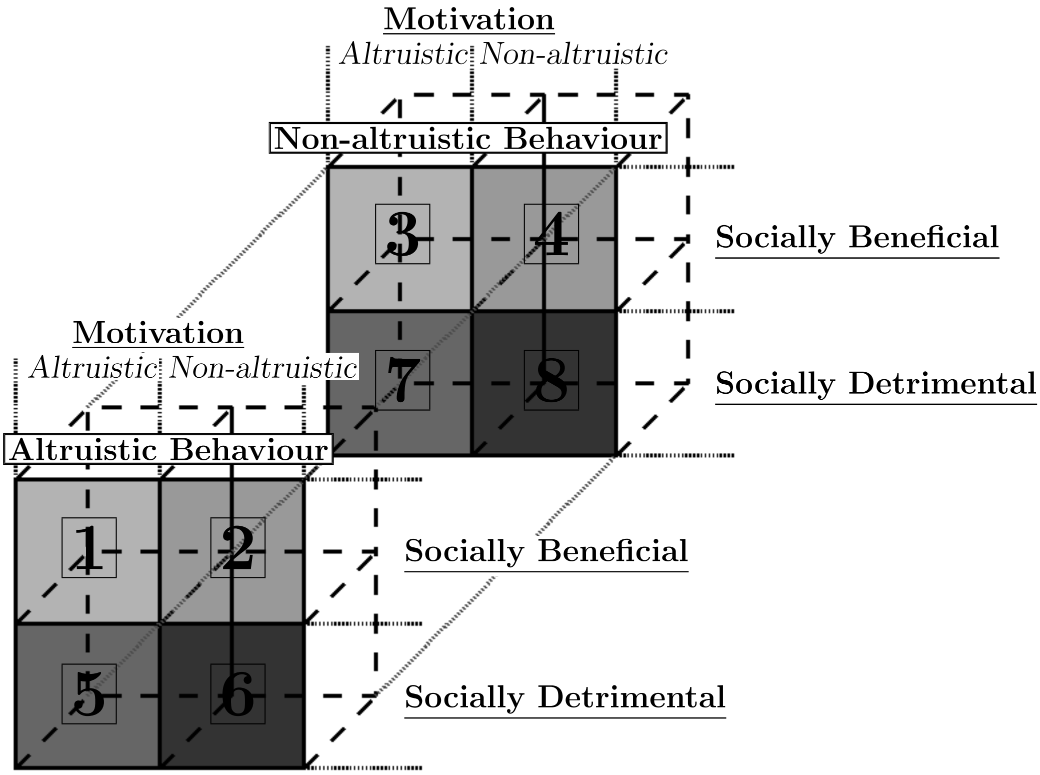

Figure 1.1: Road-map: cases of altruism

To give just one example from evolutionary biology, when animals of the same species fight for territory and status they do not usually fight to the death. In fact, fights usually appear to the human eye as something of a spectacle rather than a cut-throat struggle for survival. It was fashionable among biologists in the earlier part of the twentieth century to interpret this as a social convention which evolved for the good of group stability. However, evolutionary game theory shows that the evolution of such "conventions" can be more cleanly explained as the result of the strategic interaction of selfish animal organisms which are entirely concerned with their own survival. In a duel, the weaker animal will usually give up long before the viciousness of the conflict escalates very much, once it becomes reasonably clear what the outcome will be. It does this in its own self-interest, since there is no point "throwing good money after bad" once it has enough evidence about the strength of its opponent to "know" that it will lose. (Obviously, we are not necessarily talking about a conscious process but about evolved "rules of thumb" hardwired into the animal brain.) Nonetheless, it remains the case that these animal "social conventions" do contribute to the stability of hierarchical animal groups, and therefore indirectly to the safety of all of their members.

Gene selfishness is a distinct concept from that of the biological selfishness of the individual organism. Looked at from this perspective, there is a clear definition that an altruistic action increases the likelihood of other organisms to reproduce at the expense of the altruistic individual. This can be compatible with the selfish gene theory, because it can be brought about via the selfish genes seeking to aid copies of themselves in other animal bodies by damaging the narrower biological interests of the particular body they currently inhabit. Note also that this biological definition is distinct from the modelling of the psychological processes which lead to a particular behaviour. Evolutionarily altruistic acts committed in order to avoid a feeling of guilt, for instance, are psychologically egoistic.

None of the above definitions of apparent altruism are adequate for the social sciences in general, or economics in particular, because the connection between behaviours which economically benefit others and fitness in the biological evolutionary sense is no longer present in contemporary societies. Economically poor people on average breed more and thus have a higher evolutionary fitness in narrowly biological terms (Simon, 1993). To give another example, economists would wish to model charitable giving as altruistic, even though it may be performed for psychologically egoistic reasons.

A theoretical framework for distinguishing between economically altruistic and non-altruistic acts must specify a number of different goods which enter into each individual's utility function and distinguish those which are directly beneficial to the individual's economic well-being. This part of the individual's utility function is usually referred to as **felicity**. This is not the same as the biologist's definition of the fitness of either individual organisms or genes, because economic well-being occurs within a human culture that evolves with a degree of autonomy from the biological sphere. Cultural evolution, popularised in the idea of memes (Blackmore, 1999), is made possible because of the level of complexity to which biological evolution has developed the human body, and particularly the brain (Boyd & Richerson, 1982) (Soltis et al., 1995). However, memetic evolution occurs according to its own set of rules. Memes such as suicide bombing can spread culturally even though they reduce the biological fitness of the individuals and groups who adopt them.

The individual in economic theory is a socially constructed cultural and theoretical entity, and thus distinct from the individual organism of biology. This is sometimes obvious such as when economists treat firms or families as if they were rational individuals. However, even if the fundamental explanatory unit in a particular theory is the modelling of individual people's decisions, there is still not a simple overlap between "economic man" and "biological man". To use the terminology of Thomas Kuhn (Kuhn, 1970), the concepts used in the two disciplines are incommensurable. Both biologists and economists talk of "individuals" but the meaning of these concepts depends holistically upon the web of other concepts in the theoretical scheme, such as rationality or utility. To return to our practical concern over which goods to consider "economically fundamental", the ability of an individual to prosper in a market society depends upon more goods than those which determine narrow biological fitness. Therefore, although food and shelter should clearly enter into an individual's economic felicity, so also should biologically non-beneficial goods such as education.

Not all goods which an economically altruistic individual gains utility from can be counted in the felicity function, however. For example, if we define "lack of guilty feeling" as a good entering the individual's felicity function, then we will define away acts that we would wish to consider economically altruistic. Therefore, we instead model individuals as maximising weighted additive combinations of their own felicity and the felicity of other individuals. We are not concerned with the psychological motivations that lead individuals to behave in this way.[1]

An appealing analogy may exist between the distinction between economically fundamental goods and those which are not included in individuals' felicity functions and the Marxist concept of the economic base and the superstructure.[2] Ultimately the relative success of individuals is determined by their consumption of economically fundamental resources. However, the production and distribution of these resources within a particular group is shaped by individual preferences which can include a concern for the felicity of other individuals. Whether the psychological mechanism for this is guilt, concern for others, duty to God or whatever, these can be modelled as economically altruistic preferences. The economic base is served to a greater or lesser extent by particular types of preferences within the moral superstructure. The success of these morals in helping individuals and groups to "get on" determines which will spread and propagate. Here, the Marxist view of social change has been woefully inadequate, in that it fails to take proper account of the perennial tension between the individual and the group. Modern cultural evolution theory, however, offers a way to overcome this limitation. Ultimately, the "base" shapes the "superstructure". However, the superstructure influences the evolution of the base and has a degree of autonomy from it at any one time (Cohen, 1978).

[1] Henceforth, this particular concept of altruism shall be referred to simply as "altruism", but the caveats expressed so far about the differences compared to other definitions should be kept in mind.

[2] Note that this parallel with Marxist theory does not imply a commitment to Marxist political conclusions.

1.2.1 Classifications

We will now consider examples which fit into the different cases enumerated earlier. These range from true altruism, where there is no ambiguity about the social desirability or divergence between appearance and motivation of an altruistic phenomenon, to true selfishness. In between, however, are many cases where the water is muddier.

(1) Apparently altruistic behaviour which is altruistically motivated and is socially beneficial

The classic case of an altruistic act is one which benefits others at a cost to the altruistic individual, and where the total benefit conferred is greater than the cost. An individual with truly altruistic preferences would carry out such an action because it would increase their utility (since this includes the felicity of other individuals).

(2) Apparently altruistic behaviour which is not altruistically motivated and is socially beneficial

This is the case of enlightened self-interest; individuals may commit an act in a situation which at first glance appears identical to the pure altruism case. However, the act may in fact be in the individual's rational self-interest once the environment in which the decision is made is taken into account. Peter Hammond has discussed the subtleties of distinguishing between enlightened self-interest and true altruism in a game theoretic context, and suggests that the two may not be as empirically distinct as it first appears, when a refinement to Nash equilibrium is considered which enables efficient outcomes to be identified as unique non-cooperative equilibria (Hammond, 1975).

(3) Apparently non-altruistic behaviour which is altruistically motivated and is socially beneficial.

This is not a situation which is normally explored by economists, but it is important to realise, as argued by Hammond (Hammond, 1975), that observed behaviour will often be explicable both in terms of rational self-interest and rational altruism. The evidence under-determines the theory. The argument is usually made that self-interest is more parsimonious, and that therefore altruism should not be introduced unless it adds explanatory power. However, this indeterminacy can "cut both ways", and one of the central arguments of this survey will be that the abstract modelling of bone fide altruism should occupy a more comprehensive and secure position within economic theory.

(4) Apparently non-altruistic behaviour which is not altruistically motivated and is socially beneficial.

Here we have the classic case of some kind of “invisible hand” mechanism leading rational self-interested individuals to act in a socially beneficial manner. The most famous original proponents of this idea were Bernard Mandeville (Mandeville, 1732) and Adam Smith (Smith, 1976). Neoclassical economics has enabled this concept to be fleshed out and formalised in the framework of general equilibrium theory (Arrow & Debreu, 1954).

(5) Apparently altruistic behaviour which is altruistically motivated and is socially detrimental.

This is an important case which will be fleshed out in this survey, based on new research into the consequences of altruistic preferences using an abstract model of sequential punishment. More genuinely altruistic behaviour does not necessarily lead to a more socially efficient outcome. In fact, too high a level of altruism will be socially detrimental for generally applicable reasons.

(6) Apparently altruistic behaviour which is not altruistically motivated and is socially detrimental.

Perhaps the most common example of this category of phenomenon is imperfect altruism, such as localised altruism within a nation, trade union or business lobby, enabling them to overcome a collective action problem and increase group welfare at the expense of broader social efficiency. This can appear altruistic if ones’ theoretical perspective fails to take into account all of the costs imposed.

(7) Apparently non-altruistic behaviour which is altruistically motivated and is socially detrimental.

This is another important case which will be explored in this survey. Altruistic motivations may lead to social outcomes which are indistinguishable from those generated in a less altruistic world, and therefore no more socially efficient.

(8) Apparently non-altruistic behaviour which is not altruistically motivated and is socially detrimental.

Into this category would fit standard cases of market failure where rational self-interested individuals are, for whatever reason, not led to behave socially efficiently by the economic environment in which they make their decisions.

1.3 *Altruism and Social Structure*

Section 1.10 surveys the empirical research which has been carried out regarding altruistic motivation in order to establish a series of “stylised facts”. These can be summarised as: (A) Altruism towards other individuals is highly prevalent as a motivation for human social behaviour. (B) Real world altruism suffers from **imperfections**; meaning that real world individuals do not appear to act in a way which is consistent with putting an equal weight upon the welfare of all individuals.

Before we examine the empirical research, it is helpful to begin with some anecdotal evidence for the nature of altruism described above. Statement (A) will be denied by some, and it is undoubtedly true that methodological individualism as an approach in the social and biological sciences has greatly enhanced our understanding of Adam Smith’s proposition that behaviour which serves the common good can emerge from the interactions of entirely selfish individuals. It is therefore illegitimate to assume that any example of socially functional individual behaviour must be motivated by altruism. Nonetheless, modern industrial market societies exhibit an enormous level of complex functional integration. Although economists’ models of the market show why we need only enforce certain abstract rules (e.g. property rights, prevention of fraud, minimum welfare safety net) in order to achieve an efficient social order, they do not explain how the set of rules itself can be created, sustained and adapted to new situations. In other words, the provision of the legal and moral framework for a market economy is itself a public good, and subject to the free rider problem (Heckathorn, 1990).

The analysis of the repeated prisoners’ dilemma has shown us that although co-operation can emerge and be sustained in small groups, it is much more difficult with a large number of individuals playing the game because of the problems of monitoring and the less credible threat of Nash-reversion or other strategies to punish the individual defector. This is why competition works, but it is also why some externalities cannot be negotiated away via Coasian bargaining. It is commonly accepted that public goods games involving a large number of interacting individuals (the global reduction of carbon dioxide emissions being a perfect case) can only be solved by compulsion or altruism. The moral and legal framework which is a prerequisite for a functioning market would seem to be a similar case. It is in each of our self-interest that everyone else respect property rights and refrain from fraud, but we each can break the rules at great gain to ourselves. Reputation formation may be able to explain some of the spontaneous formation of a market order, but the problems of underdeveloped countries developing proper functioning market economies surely underlines the fact that some kind of effectively-run co-ordinated unit with power of compulsion (i.e. a benevolent state) is vital.

If we accept theoretically that public goods exist which can only be provided by compulsion or altruism, and that benevolently-run compulsive institutions fall into that category, the attempt to use only compulsion to explain their provision leads to an infinite regress. This is because the provision of the compulsive institution is itself a public goods problem, which can either be solved by compulsion or altruism. Sooner or later, we must postulate altruistic motivation on the part of some individuals. This does not of course settle how far back the regress goes. For example, it could be that there is only one single powerful altruistic individual in the world, who is manipulating the environment of all the others, who are entirely selfish. This, however, is not very plausible, and it seems more likely that individuals differ in their degree of altruism, with those more altruistic ones contributing to sustaining a social structure which can harness the limited altruism of the less altruistic ones (the market economy being an example of an institutional structure which achieves this).

We now come to statement (B). From the above discussion, the fact that society requires private markets, a legal system and prisons is itself powerful anecdotal evidence that individuals have greater or lesser imperfections in their altruistic perceptions and motivations. If this were not the case, we could have a society more along the lines of a democratic anarchistic society, where all individuals, due to their perfect altruism, were able to discuss, agree upon the common good and then act in concert. (Some people might see this as a real political possibility, but the key point is to be clear about the rationale concerning the type and degree of altruism that would be required of the individuals in society in order for such alternative political structures to operate efficiently.)

The institutions of a liberal democratic market society, such as markets, legislatures (and the corresponding rules to facilitate organised discussion), legal systems and law enforcement agencies can all be seen as **altruism amplification devices**; they attempt to get the best possible social outcome from the imperfect altruists who form the "human material" of the polity (Sober & Wilson, 1999). In "The Socially Optimal Level of Altruism" (Chapter 2), a simplified abstract model of this process is presented, and it is shown that altruistic preferences are ambiguous in their effect on social welfare, once they are permitted to interact with a system of social incentives, implemented via schemes of punishment.

1.4 *The Self-Interest Assumption*

The assumption of rational self-interested behaviour has been a highly fruitful one in economic theory. It has a simplicity and elegance that has allowed economics to live up to its founders' project of showing how a socially desirable spontaneous order can arise in market society which encompasses the knowledge and aims of many individuals (Hayek, 1960). Despite this, in recent years there has been a growth in the desire to move beyond the self-interest assumption. As we will see in section 1.10, some of this has taken the form of empirical work demonstrating convincingly that many specific cases of individual behaviour can be explained more adequately once the self-interest assumption is dropped.

What remains unsatisfactory is the degree to which the standard paradigm of welfare economic analysis continues to assume that whilst social institutions may ideally be designed by a benevolent utilitarian social planner, the ordinary individuals who act within them usually remain highly rational but at the same time highly selfish. The issue of whether this approach is adequate for normative analysis in economics takes us right to the heart of key issues in the philosophy of the social sciences.

Much of the initial attraction and success of the self-interest assumption, as well as the apparent explicit basis for the more recent work which has questioned or modified it, is based on a narrowly "positivistic" view of economics. This label is intended not to refer to a specific detailed position on the ontology and epistemology of social sciences, but rather to the view, most famously espoused by Karl Popper, that the value of the social sciences over and above mere superstition is their ability to make testable predictions (Popper, 1959).

In this light, the attractiveness of the self-interest assumption is clear; it provides a good basis point for a research programme in the social sciences. This concept was introduced by one of Popper's followers Lakatos. He distinguished (Lakatos, 1970) between the core of a research programme, a set of firm theoretical principles which are not questioned, and the peripheral ideas that are brought into a theory as it is tested against reality and modified. As time goes on and the possibilities of adaptation of the core principles become exhausted, it will become clear which areas of reality are difficult to explain by a particular research programme. Eventually, "progressive" research programmes become "degenerative" ones, with the peripheral assumptions becoming increasingly incoherent. With this approach, Lakatos sought to explain the discontinuities in scientific progress which Popper's theory found more difficult to explain.

The core of the standard research programme in economics could therefore be thought of as individual rationality and self interest, whereas asymmetric information or different equilibrium concepts would be peripheral components introduced and altered more freely in order to improve the fit between theory and reality. The recent increase in interest in dropping the self-interest assumption in empirical applications could be seen as a sign that this research programme has begun to exhaust the possibilities, so that in certain empirical areas it is time to consider a different research programme.

There are problems, however, with making this the philosophical justification for taking economics into the arena of altruistic motivation. Firstly, as a predictive framework, the self-interest framework is still highly productive. It would seem that if altruism is to be introduced into the body of the economic theory on this pretext, it will remain fairly ad hoc, in the sense that it will only be used when the self-interest assumption evidently fails empirically. Secondly, one senses that the attraction of altruism is not merely that it provides an alternative predictive research programme. It can be argued that economics has neglected a large area of human potential, and left the consideration of human morality and socialization largely in the hands of sociologists, when there is nothing inherent in the economist's arsenal of rigorous tools and techniques which prevents them from being applied in this area. This kind of thinking, however, requires a more subtle view of the role and potential of the social sciences in the body of human intellectual endeavour.

Kuhn took a more radical view of progress in the natural sciences than Popper and Lakatos. He argued (Kuhn, 1977) that the decision between what Lakatos would have called "progressive" and "degenerative" research programmes (the process which Kuhn referred to as "paradigm shift") could never be nailed down to evidence in such a simple way that all scientists could agree which programme provided the best potential for future development. Whenever paradigm shifts occur, there are competing considerations or scientific values which must be balanced. Ultimately, only the conscience of the individual scientist can decide.

A good example of this is Kuhn's account of the paradigm shift from eighteenth century Phlogiston theory to the modern paradigm in chemistry. When Phlogiston theory was abandoned in favour of the theory based on elements in the early nineteenth century, many older members of the chemistry profession resisted because they remained attracted by the wide degree of physical properties which Phlogiston theory was apparently able to explain. In contrast, the main attraction of the new theory was that despite its limited explanatory scope, it was able to make predictions of mathematically precise proportions of ingredients in chemical reactions. Eventually, the attraction of the value of mathematical precision won the day, but it took more than another 150 years for developments in quantum mechanics to allow the new mathematical chemistry to explain the same breadth of physical features as the old Phlogiston theory.

The relevance of this analogy for the philosophy of economics is that the decision between different research programmes or paradigms in economics must also ultimately rest on values such as aesthetic appeal as well as sheer "number of facts explained" (which is of course a concept incapable of precise and rigorous operationalization anyway). Another factor which must be considered is the ethical dimension to the social sciences. Although there is an attractiveness to the view, espoused by both Kuhn and Popper, that there is nothing inherent in the social sciences which precludes them from having similar aims and status to those of the natural sciences, it should be admitted that if this view on the ultimately value-based nature of scientific discovery is correct, then the ethical image that the social sciences uphold for humanity must also be part of this value judgement. This links in with the indictment that interpretivist sociologists have made against economics that, to put it crudely, by developing its intellectual appeal, it helps to create a society of selfish egoists in its own image.

One does not need to accept the radical thesis that there exists no social reality independent of our theoretical constructs (whether through language, customs or social science) to accept that there is a great deal of validity in the idea that there is a feedback process between social reality and the concepts that social scientists use to explain and describe it. Economics has played an ethical role in promoting market societies, because of the view of many economists that individualistic societies produce greater economic efficiency and thus a better way of life than societies where people's economic behaviour is more closely controlled. From this perspective, the "core" assumptions of economics are not merely useful components of a predictive framework. They are, rather, central to the ethical vision of human nature, and its potential when permitted to develop freely, at the heart of economics. The view that economics should aim to be value free is thus self-defeating and self-deceiving. Pattanaik, for instance, has argued that value judgements have a necessary place in economic theory, and that, by elaborating the consequences of normative assumptions, progress can be made in normative theory in a rational manner, in a similar way to positive economics (Pattanaik, 1971).

Consider, for example, the assumption of rationality. In the normative sense that the preferences of the individual should be sovereign, this is not the kind of proposition that can be proved or disproved. The fact that people act as if they know what they want does not imply anything about the moral status of these "desires". The assumption of rational behaviour at the heart of many models in welfare economics is therefore much more than just a predictive modelling technique; it is the embodiment in economic theory of the moral value of a society based on respect for individual autonomy.

Despite its empirical usefulness, the assumption of selfish preferences does not, on the face of it, share the same positive ethical basis. This, arguably, provides a strong reason to bring the modelling of altruism into the heart of welfare economic theory. It also changes the emphasis of the level of analysis from that of empirical explanation of specific phenomena to that of assessing the economic efficiency, and therefore the social desirability, of different levels and forms of altruism in human societies. If we live in societies in which people exhibit partial altruism, or altruism in some contexts but not in others, then this is something that welfare economics must seek to explain, and not be content merely to assume. Indeed, it is probably the undesirability of making ad hoc assumptions about partial altruism that has so far led to the cleaner solution of simple self-interest remaining the main workhorse of abstract welfare economic analysis.

A model in welfare economics will require a number of properties if it is to satisfy the general prescription laid out above. One of the most important methodological questions which arises concerns the relationship between the social welfare function and the utility functions which the agents in the model seek to maximise. If the level of altruism is to be treated as an endogenous variable which can be altered (e.g. in different societies or via differential socialisation processes in the context of an existing society) then the possibility must be left open that each individual could be directly acting so as to maximise the social welfare function. Harsanyi has laid out the requirements for interpersonal comparability of utility if this benchmark case of perfect utilitarian, or "impartial", altruism is to be coherent (Harsanyi, 1986). (See section 1.5.) If we are to justify any kind of arrangement as being superior to this perfect utilitarian society, we need to introduce additional structure to the model (corresponding to Lakatos' peripheral assumptions). The opening quotation by Meade, for instance, suggests that there might be differences between the political and economic "marketplace" which justify a different level or kind of altruism as being socially optimal in different "scenarios" (Meade, 1973).

An interesting analogy to the role of the altruism assumption being suggested here is that of rational expectations in macroeconomics (Lucas, 1976). Just as there seems to be a kind of logical inconsistency between the assumption of rational agents with perfect understanding of an accurate model of the macro-economy and the use by these agents of adaptive expectations, there appears to be a parallel inconsistency between the assumption of moral human beings who design their society along the lines of utilitarianism but then act so as to selfishly maximise their own utility. Just as there are "hidden costs" to processing information which can explain why adaptive expectations are often a more plausible modelling technique in macroeconomics than rational expectations, there are "hidden costs" to individual altruism which may explain why it is socially optimal, despite the possibility of a society of utilitarian altruists within the structure of the model, for people to exhibit imperfections.

1.5 *Modelling Altruism*

Although altruism has become a somewhat peripheral issue in modern normative welfare economic theory, it has been an integral and well-recognised part of neo-classical economics for almost as long as the discipline has existed. Edgeworth and those who further extended his work quickly developed the implications of altruistic preferences for competitive equilibria in an Edgeworth box (Collard, 1978). The most important result from this body of work is what is known as the **non-twisting theorem**. This states that provided altruistic preferences are non-paternalistic, the contract curve will be the same as if individuals were fully selfish, except that it will be shrunken (the ends will be cut off) because extremely unequal distributions will be undesirable for both rich and poor. What is essentially required for this result to hold is that, in a precisely definable way, altruistic preferences should respect the relative valuations placed on different goods by the other individuals towards whom the altruism is directed. A formulation of social utility in terms of weighted sums of felicities will guarantee this. In the limit where individuals value each other as much as themselves, the contract curve shrinks to a single point. This is the bliss point for both individuals, since both individuals have the same utility function. This corresponds to the society of perfect utilitarian altruists introduced above.

The normative analysis of altruism raises some interesting additional issues. Suppose we have a society containing two individuals, each of whom cares about the other. As we have argued, it is necessary to distinguish between **felicity** (represented by $V_1(X_1)$ and $V_2(X_2)$), which is a measure of the satisfaction that each individual gets from consuming goods, and **utility** (denoted by $U_1(X_1, X_2)$ and $U_2(X_1, X_2)$), which is a representation of the overall preferences of the individual as they determine his behaviour. Having made this distinction, we could take a number of directions in representing the two individuals' utility functions. If people care directly about each other's *utility*, then it is possible to get multiplier effects which can have perverse results. These will be discussed shortly. We could, on the other hand, have each individual's utility depend on a weighted sum of his felicity and the felicity of the other. Letting θ be the **coefficient of altruism**, and X_1 and X_2 be the consumption bundles of person 1 and person 2, this situation would be represented as:

$$U_1(X_1, X_2) = V_1(X_1) + \theta V_2(X_2)$$

$$U_2(X_1, X_2) = V_2(X_2) + \theta V_1(X_1)$$

The potential problem even with this formulation is that increasing the level of altruism automatically increases the utilities of both individuals, regardless of any effect upon their behaviour. The sequential punishment model presented in “The Socially Optimal Level of Altruism” (Chapter 2) nonetheless uses this formulation, because the felicities of the individuals in the model, rather than their utilities (which we shall call social utility, in order to be clear), form the basis of the normative assessment of the outcome of the model via a utilitarian social welfare function. This approach, however, does raise its own set of questions about whether it is legitimate to make a distinction between felicity and the “moral preferences” embodied in the social utility function. The justification is that we are seeking to assess moral preferences in terms of their contribution to economic efficiency. It could, however, be objected that this is an overly narrow view about the benefits of altruism, since compassion and empathy should be valued in and of themselves. The counter-argument, of course, is that compassion and empathy can lead to misguided actions, in which case it is hard to see why they should be automatically desirable.

Figure 1.2 shows a two person pure-exchange economy in an Edgeworth box where individuals are partially altruistic, with utility functions as specified below, and with $\theta_1 = \theta_2 = \frac{3}{4}$. Points c and d are the bliss points for the two individuals. Due to their partial altruism, and declining marginal felicity from goods X_1 and X_2, both would choose not to consume the entire endowment even if they could dictate the allocation (although, since they are only partially rather than fully altruistic, they would each like to take more than half).

$$V_1 = \sqrt{X_1} + \sqrt{Y_1}$$

$$V_2 = \sqrt{X_2} + \sqrt{Y_2}$$

$$U_1 = V_1 + \theta_1 V_2 = \sqrt{X_1} + \sqrt{Y_1} + \theta_1 \left(\sqrt{X_2} + \sqrt{Y_2}\right)$$

$$U_2 = V_2 + \theta_2 V_1 = \sqrt{X_2} + \sqrt{Y_2} + \theta_2 \left(\sqrt{X_1} + \sqrt{Y_1}\right)$$

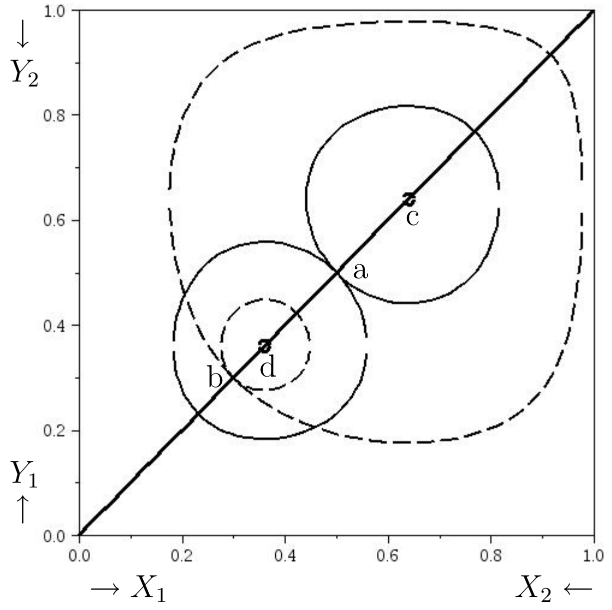

Figure 1.2: The shrunken contract curve

Point *a* shows a possible Pareto-efficient allocation, where there is a tangency between the two individuals' indifference curves. Point *b* also shows a tangency point, but it is *not* Pareto efficient, because moving closer to point *d* will make *both* players better off. The solid diagonal line represents all the possible points of tangency between the two individuals' indifference curves. However, the contract curve only consists of the part of this line lying between points *c* and *d*, because the two individuals would always bargain away from an allocation involving a level of inequality more extreme than *c* and *d*. Figure 1.3 shows the derivative of the individuals' marginal rates of substitution at each X_1 value along the tangency line. Only when these are both positive do we have Pareto-efficient allocations, since only then will the indifference curves "curve away" from each other in the standard textbook manner, where there is no altruism present.

This framework allows the introduction of one of the most commonly recognised problems with the translation from altruistic preferences to social outcomes. If the two individuals are indeed single

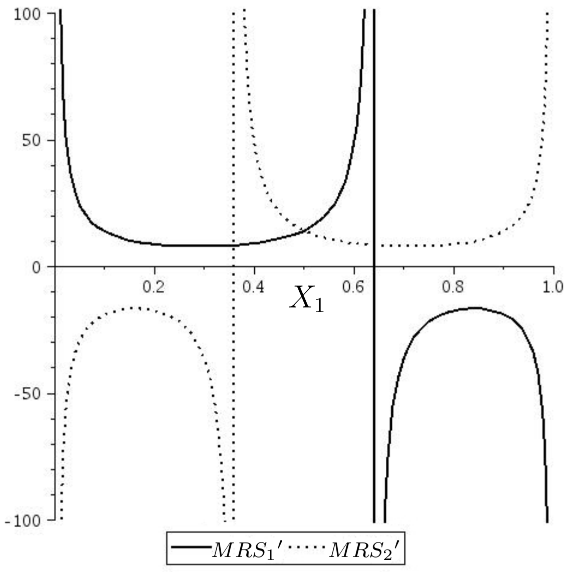

Figure 1.3: Derivative of the MRS of individuals 1 and 2

individuals, then it is reasonable to suppose that they would bargain to a point on the shrunken contract curve (i.e. to their unique shared bliss point if they are fully altruistic). However, if the two individuals in the model are taken to represent *groups* of individuals then there is a "public goods" problem in the sense that an individual in group A cannot redistribute to group B as a whole because a distribution of X to group B only increases the average income of group B by $\frac{X}{N_B}$ (where N_B is the number of members of group B). This implies that redistribution will not occur voluntarily. Hence a point like b can be a competitive general equilibrium, even though it is not Pareto-efficient.

A degree of state enforced wealth redistribution can, in theory, make both rich and poor better off in utility terms (essentially, greater equality is a state-provided public good). It should be noted, however, that this interpretation of group redistribution has already moved away from the assumption of perfect utilitarian altruism, since it has defined altruistic preferences as depending on the *average* utility of other individuals, whereas in a society of fully utilitarian altruists each individual cares for the marginal benefit of a gift to any individual *just as much* as that individual cares about himself. The public goods problem discussed above therefore does not arise in the case of full utilitarian altruism.

Although the alternative modelling approaches mentioned above have not been taken in "The Socially Optimal Level of Altruism" (Chapter 2), it is worth considering the consequences and issues that arise with them in order to achieve greater clarity. One possibility is to have each individual's utility be a weighted average of the felicities of all individuals (including himself). This approach solves the problem which arose above of greater altruism automatically increasing the amount of utility in the economy. However, from the perspective of the sequential punishment model, it is an unnecessary complication. It would also be problematic because the model includes an infinite number of individuals. Although empathy with an infinite number of individuals in the sequential punishment model may seem unintuitive, it must be borne in mind that each individual need empathise with only one individual at any one time.

Another, more interesting, alternative approach is to have each individual's utility depend on his or her own felicity and the *utility* of the other individual. This could be represented as follows:

$$U_1(X_1, X_2) = (1-\theta)V_1(X_1) + \theta U_2(X_1, X_2)$$

$$U_2(X_1, X_2) = (1-\theta)V_2(X_2) + \theta U_1(X_1, X_2)$$

When this system of simultaneous equations is solved, we get:

$$U_1(X_1, X_2) = \frac{1-\theta}{1-\theta^2}\Big(V_1(X_1) + \theta V_2(X_2)\Big)$$

$$U_2(X_1, X_2) = \frac{1-\theta}{1-\theta^2}\Big(V_2(X_2) + \theta V_1(X_1)\Big)$$

Even though the utilities of both individuals have been normalised so that they are an average of their own felicity and the utility of the other individual, we still get a fairly complex multiplier term at the beginning of each individual's solved out utility function. This multiplier effect can produce interesting results in some models. Note however, that as long as certain regularity conditions are fulfilled (e.g. that $\theta \neq 1$) we can still express each individual's utility as being ultimately dependent only on the felicities of all individuals (and, of course, the coefficients in the model). This suggests that, if we want to abstract away from these multiplier effects caused by altruism, it would seem to be sensible simply to base individuals' social utilities directly upon felicities, as outlined above.

Bernheim and Stark have presented a model of marriage partnership which is a good example of the multiplier phenomenon in action (Bernheim & Stark, 1988). They assume that "females" are all equally altruistic, but that "males" vary in the degree of altruism towards their partner. Females choose male partners so as to maximise their own utility. It turns out that under certain conditions, females prefer a more selfish male partner because then there is less of a cost from the relatively low utility of the female entering into the utility of the male and then feeding back into the utility of the female via the altruism of each partner for the other. Basically, women with difficulty getting utility from their own felicity (i.e. a low coefficient on their own felicity in their utility function) will prefer to be with a selfish male partner, who will not care that she is unhappy and whose happiness will perk her up at least a little bit. This provides a possible justification in economic theory for why "nice guys finish last" (although the coefficients in the model can sometimes lead the other way, with more altruistic males being preferred by females who are better able to get utility from their own felicity, due to the positive "spillover effect").

The definition of social utility functions in terms of weighted averages of felicities, and the assessment of outcomes using a social welfare function defined as the simple sum of felicities, requires ratio-scale interpersonal comparability of felicities (Roberts, 1980). As argued by Harsanyi (Harsanyi, 1986), only once we permit such interpersonal comparisons does it make sense to view the utilitarian social welfare function as requiring perfect, which he calls "impartial", altruism (all individuals weighted equally), and individuals' social utility functions as falling short of this by exhibiting imperfect altruism (lower weighting on some or all other individuals than upon onesself).

1.6 *Altruistic Punishment*

A major contemporary area of concern regarding the desirability of altruistic preferences is the role of punishment of defectors as a mechanism to achieve socially desirable outcomes. Empirical research by Fehr and his collaborators (Fehr & Gächter, 2000a) has shown that the ability of agents to punish, even at a cost to themselves, is of vital importance to prevent more selfishly inclined individuals from free riding. Fehr et al. have used small scale experimental prisoners' dilemma style games with real payoffs to demonstrate that people's anger against defectors leads them to be willing to altruistically punish, even when this comes at a cost to them so that it is individually irrational to do so.

The sequential punishment model presented in "The Socially Optimal Level of Altruism" (Chapter 2) examines the issue of punishment and the role it plays as an altruism amplification device from an abstract perspective. It is similar in nature to an infinitely repeated non-zero-sum stage game, with the added feature that the level of altruism is endogenous to the normative analysis of the model. The model demonstrates that there are offsetting benefits and costs to increased altruism. Although more altruistic individuals are less tempted to commit socially damaging acts, they are also less afraid of being punished, because they care about their own utility less relative to that of others. Less altruistic individuals are also willing to punish the transgressions of others more frequently. It turns out that for a wide range of parameters, the level of altruism is irrelevant to the conditions required to achieve the socially optimal co-operative outcome. Under fairly general conditions, too high a level of altruism will even be counterproductive, and make a socially efficient outcome *harder* to achieve.

1.7 The Evolution of Altruistic Preferences

Treating the coefficient of altruism as endogenous to the optimization procedure encapsulated in the model obviously raises the question of how it is that a certain level or type of altruism comes about, and by what dynamic process this can be changed. This is not an orthodox area for economic modelling, which usually treats people's preferences as an a priori assumption rather than something which can be explained. There are, however, some precedents in the economics literature. There is also a vast literature in biology and cultural theory on the use of evolutionary theory to explain human altruistic behaviour. Usually, however, preferences are not explicitly modelled and instead highly simplified behavioural phenotypes are used.

It has been argued by Oded Stark that family demographics can be better explained through the passing of altruistic preferences from parents to children by example than by other forms of enlightened self-interest (Stark, 1995). One of the most interesting of the models discussed is one which seeks to explain why children's altruism towards their parents increases once the children have children of their own. Empirical evidence is cited to show that the presence of children increases the amount of altruistic behaviour by parents towards grandparents. Stark argues that this occurs so that parents can set the example to their children to look after them in a similar way when they become old. The microeconomic basis of such a framework is that there is a certain probability that children will simply copy the behaviour of their parents rather than do what is optimal for their self-interest.

The dynamic fragility of altruistic preferences has been reflected in models of the psychological processes which lead to norms of co-operative altruistic behaviour and the potential for government intervention to negatively impact upon them, resulting in possible welfare loss (Hollander, 1990). Work has also been conducted on models of mutually reinforcing altruism among individuals in situations of strategic complementarity such as the workplace (Rotemberg, 1994). Individuals whose preferences become genuinely more altruistic may do better if this leads others to be more altruistic towards them. Other approaches to the dynamic survival of altruistic behaviour have emphasised the connection between altruism and "docility" due to limited human cognitive abilities (Simon, 1993). The idea here is that because individuals cannot distinguish between cultural norms which improve their own fitness and those which require altruistic behaviour, altruism can "piggy-back" on the back of other norms provided that the overall impact is to improve individual fitness.

1.8 *The Multilevel Selection Paradigm*

The sociobiological approach would perceive differing levels of altruism as the result of differing levels of genetic relatedness. However, as we have seen, this does not seem to be adequate in explaining the role of cultural factors in determining the altruistic interactions between genetically unrelated individuals which is the vital glue that holds together a complex society, regardless of the degree to which market forces are relied upon to achieve economic coordination. The functionalist approach in sociology, most famously associated with the work of Robert Merton, would seek to explain the presence of a particular level and pattern of altruism in terms of its functionality for the overall social system (Merton, 1968). This approach, however, can be dangerous if it fails to use a modelling approach based upon methodological individualism and if it neglects to provide a causal explanation to underpin the functionalist one.

More recently, the social sciences have seen sophisticated and innovative attempts to use Darwinian ideas to integrate the functionalist and individualist approaches to the study of human society (Sober & Wilson, 1999). It has been argued that evolutionary progress has resulted in the functional integration of smaller units into larger ones via the process of "group selection". For example, competing strands of DNA "work together" within cells, as do cells within the body. There seems to be no good reason to assume that functional integration on the societal level should be impossible, and thus individual human animals can be seen as working together cooperatively within society in an analogous way.

The multilevel selection approach, however, demands that the explanation for the emergence of group-functional behaviour also takes into account the potential for conflict between the component members of groups. For example, sometimes "selfish genes" will take actions that lower the fitness of the other genes in order to benefit themselves. However, the structure of a DNA strand makes it fairly difficult for individual genes to do this. Dawkins uses the analogy that they are like "rowers in a boat" who are unable to escape their dependence upon each other and so are forced to work together in order to maximise their own survival chances (Dawkins, 1976). The situation is compatible with an individualist explanation[3] of the functional integration of individual animal bodies because genes are better off, overall, as part of DNA strands protected within cells and bodies, than as separate molecules in the outside world. Groups of genes whose "boat" does not tie them together sufficiently firmly will not succeed in replicating themselves.

The analogy between genes and DNA strands works also between individual bodies and social structures; social structures, like the organization of genes into DNA strands and cells, help individuals to overcome their individual selfishness and work together to improve their overall fitness. In this sense, human society as a totality acts as an altruism amplification device by providing a structure in which the temptation towards socially harmful acts of individual selfishness is reduced. The existing types of human social organization must be explained partly by the process of competition between groups leading to the expansion of more efficient forms of organization and partly by the need for any social system to contain destructive competition between members, which may require that certain features be present, the punishment of wrongdoers being a particularly important example.

The "invisible hand" of the market, as rigorously encapsulated in the First and Second Theorems of Welfare Economics (Arrow & Debreu, 1954), can also be understood within this context as showing how, provided a complete set of markets is created with well-defined property rights, entirely selfishly motivated economic activity *within* this system leads to a mutually beneficial outcome in which all individuals are made better off than their original endowment. This suggests that social control of all economic activity[4] is not necessary to achieve efficient functional integration at the level of human society. On the other hand, the temptation of individuals to break the law for their own gain will never be fully removed due to the evolutionary pressure against functional integration at the individual level, and so the idealised assumptions of the First and Second Theorems are unlikely to be fully achieved.

Economic theory has other important roles to play in this Darwinian framework because it provides a set of tools to assess the efficiency of particular institutional arrangements through use of the "ideal benchmark" of a social welfare function and paradigms for explaining theoretically the success or

[3] Individualist in the sense of starting from the selfish gene as the fundamental explanatory unit.

[4] This could be achieved through direct state control or by, for example, a very strict moral code.

failure of decentralised social systems to achieve efficient outcomes. Normally, the preferences of individuals are assumed to be fully selfish and other features of the model such as the presence of asymmetric information or externalities are used to explain the differences in the performance of market coordination of economic activities in various situations. No reason, however, seems to preclude treatment of the level of altruism present in individual preferences as an alterable variable that may have an impact upon social efficiency. The sequential punishment model presented in "The Socially Optimal Level of Altruism" (Chapter 2) takes this approach.

1.9 *The Limits to Altruism*

There are three logically distinct facets to the concept of limits to altruism. Firstly, there are limits to how much there is in society. We shall call these "positive limits". We shall explore these in sections 1.10 to 1.16. Secondly, there are limits to how socially beneficial altruism can be. As laid on in the road-map in section 1.2, there is a key distinction to be made here between particular kinds of altruistic action, and altruistic preferences themselves. We will focus on these "normative limits" in section 1.17. Thirdly, there is the question of the broader consequences to the presence of these limits, particularly for normative policy issues. We explore these in section 1.18

1.10 *"Stylised Facts" About Altruistic Behaviour*

The idea that human beings are motivated partly by altruistic and partly by selfish objectives, and that individuals differ in the degree to which they exhibit these types of preference, is fairly intuitive and non-controversial. Explaining why this pattern should emerge is more difficult. Thus far, we have considered the overall role that economic theory can play in helping to explain the altruistic behaviour we observe. We will now consider evidence from empirical economics for the "stylised facts" about the imperfections of real world altruism that the models presented in sections 1.17 and 1.19 are needed to help explain. These can be summarised as:

(1). Altruism is often present as a motivating factor in the economic realm.

(2). Altruism is frequently subject to imperfections.

(3). These imperfections differ between individuals.

(4). Malevolence is frequently evident.

Our aim in the following sections will be to substantiate the above statements. Before considering the empirical evidence in detail, it is helpful to quickly summarise the areas of empirical economics where evidence of altruism can be found. Firstly, there are results of experiments where individuals play games with small payoffs (usually, but not always, in cash form), some of which have already been mentioned. Secondly, there is the evidence of individuals making charitable donations and contributing voluntarily to the provision of public goods. Closely connected, there is thirdly the issue of parental bequests to their offspring. A fourth body of evidence comes from responses to surveys about willingness to pay for certain public goods, such as environmental protection. There is, finally, evidence of limits to altruism from issues of international political economy.

1.11 *Experimental Economics*

One of the most obvious areas of empirical research in economics that provides evidence of the failure of the assumption of rational self-interested behaviour involves the use of experimental games with small (usually monetary) payoffs. The classes of game most commonly played are the ultimatum game, the public goods game, the prisoners' dilemma, the centipede game and the bargaining game. In all these cases, it is well established that the observed behaviour conclusively violates the predictions of the standard theoretical framework based on rational self-interest. However, since self-interest must always be combined with other assumptions, such as full knowledge of common rationality, it is always open (but with varying degrees of plausibility) to consider these peripheral assumptions as having been falsified, rather than self-interest itself.

1.11.1 The ultimatum game

The "ultimatum game" is played between two individuals. The first individual proposes the division of £1 between the two individuals and the second individual can either accept the offer or refuse, in which case both get a payoff of 0. If both individuals are entirely rational and selfish, and there is full common knowledge of rationality, non-co-operative game theory would predict that the first individual will offer the smallest amount they can that is higher than 0 (i.e. 1p) and that the second individual will accept. However, when the game is played in experimental situations with real people, the predicted outcome occurs extremely rarely, and there is significant variation between cultures regarding the amount that the first individual offers to the second.[5] It is clear, therefore, that in most cases individuals are constrained from fully selfish behaviour by moral norms. Not only are people in the position of individual 1 observed to offer more than the minimum to the second person, if the offer is not generous enough, people in the position of individual 2 are observed to refuse the offer, even though this is economically irrational for them.[6]

The research into the ultimatum game and some variants has revealed a number of other important phenomena. Camerer and Thaler also cite experiments comparing the ultimatum game to the dictator game, where the second individual is simply forced to accept the allocation proposed by the first individual (Camerer & Thaler, 1995). In the dictator game, offers are usually lower than in the ultimatum game, but not as low as predicted by pure self-interest. This shows that the offers in the ultimatum game are being partly driven by a desire of the proposer to be fair and partly by fear that the second individual will refuse. Other experimental results referenced in the same paper show that changes in the framing of the question (which should be irrelevant to the strategies played if individuals are self-interested), such as whether the proposer is chosen randomly by a lottery or arbitrarily by the researcher, and whether or not the income is described to have been "earned" by the proposer, can make a significant difference to the altruism exhibited by the proposer towards the second individual.[7]

Andreoni and his collaborators (Andreoni et al., 2003) have extended the ultimatum game to convexify the strategy space of the second individual by allowing them to continuously shrink the "pie" after the allocation is chosen by the proposer (the standard ultimatum game allows a discontinuous choice only between 100% and 0% in terms of the possible moves by the second individual in this convexified ultimatum game). This has the advantage that the experimental data becomes sufficiently rich to allow the axioms of revealed preference to be tested. They find that the behaviour of experimental subjects is compatible with the hypothesis of rational altruism.

Individuals have preferences not just over their own money payoff, but over the money payoff of the other individual. These preferences can be both "benevolent" and "malevolent" in the sense that if an individual has a high monetary payoff relative to others, then they gain utility when others get a greater monetary payoff, but if the other individual is already better off in monetary terms then

[5] The empirical evidence has been summarised (Camerer & Thaler, 1995) as showing that offers are usually between 30% and 40%, with the mode often being 50%.

[6] Very few offers are below 20%, and those which are this low are often rejected (Camerer & Thaler, 1995).

[7] Clark finds similar evidence for framing effects in a game where individuals can make costly votes to reduce outcome inequality (Clark, 1998).

the individual's utility is *decreasing* in their opponent's monetary payoff. These types of rational inequality-aversion explain why it is rational for individuals to make more equal offers at their own material expense and to "refuse" (in the sense of partially or fully shrinking the pie) unequal offers, again at their own material expense. Andreoni finds a very similar result with experiments using the dictator game (Andreoni & Miller, 2002).

1.11.2 The public goods game

The public goods game involves a situation where a number of individuals must choose whether or not to contribute to a public good. Each unit of contribution produces more than 1 unit of the public good. This is split over the N individuals, however, and so it does not pay each individual to contribute if they are self-interested, because they are still able to free ride on the contributions of other individuals. The evidence has been summarised (Dawes & Thaler, 1988) as showing that contributions are usually in the region of 40%-60%.[8] When the game is repeated with the same individuals playing, the average level of contributions tends to drop over time. Fehr and Fischbacher argue that this can be explained by a model in which the desired level of contributions depends upon the amount others contribute[9], but that desired contributions climb at less than a slope of one, either because there are selfish individuals present or because altruistic individuals are not completely willing to match the contributions of others (Fehr & Fischbacher, 2003). This causes the levels of contribution to the public good to "unravel" over time.

Framing effects have also been found to be significant in public goods games. It has been found (Andreoni, 1995b) that the level of contributions is significantly affected by whether the public good is seen as the provision of a positive benefit or the avoidance of a bad externality, even though the payoff structure is identical in both cases. Such results could only be explained if individuals have utility functions which value the size and structure of changes to outcomes, as well as final outcomes. Other research which shows that allowing group discussion before contributions are made greatly increases the contribution level (Dawes & Thaler, 1988). This may be because people are induced to make promises to contribute, which they then feel are binding, or because they get to know the members of their group, and thus feel more altruistic towards them. Interestingly, when individuals are split into

[8] Andreoni considers whether this is due to imperfect rationality or altruism, and concludes that both play a part (Andreoni, 1995a).

[9] This would require some kind of altruistic preference for reciprocity ("if others contribute then so should I").

groups but the beneficial externality goes to another group, the result is reversed and free riding is reinforced because the group agree together to act in their group self interest and refuse to contribute to the intergroup public good. This is strong evidence for the narrow-group-based nature of human co-operation which, as we shall see later, relates closely to explanations of the evolution of altruism based on group selection.

1.11.3 The prisoners' dilemma

The main empirical anomaly for the self-interest assumption that concerns the prisoners' dilemma is the observation that co-operation occurs in a finitely-repeated prisoners' dilemma game when the well-known backwards-induction argument should lead co-operation to unravel and defection to occur from the first period onwards. Andreoni and Miller conclude on the basis of experimental evidence that rational reputation-building on the part of most agents plus true altruistic behaviour on the part of a minority offers the best explanation for this phenomenon (Andreoni & Miller, 1993).

1.11.4 The centipede game

The centipede game is a sequential-move game with finite length in which two players get chances in sequence either to take the larger of two payoffs on the table, or to pass, in which case both payoffs are multiplied by a factor. After N alternating moves ($\frac{N}{2}$ for each player), the game ends. By backwards induction, the last player to move will take the larger pile, since otherwise they get nothing. Since the payoffs are set up so that the larger pile at the $N-1^{\text{th}}$ move is bigger than the smaller pile at the N^{th} move, the penultimate player to move will take the larger pile. Therefore the player at the $N-2^{\text{th}}$ move will also take the larger pile. By backwards induction, this reasoning can continue to be applied to show that the first player to move will take the larger pile, even though if the two players could co-operate they could end up with massively higher payoffs.

McKelvey and Palfrey conduct experimental centipede games and find that typically players pass for a number of periods before somebody takes the larger pile (McKelvey & Palfrey, 1992). They explain this using the idea that a proportion of the population is altruistic, and that selfish individuals can pretend to be altruistic in order to get their opponent to co-operate. By calibrating the model to their data, they estimate that 5% of the population is believed to be altruistic.

1.11.5 Bargaining games

Unlike the other games, bargaining games do not have sufficient structure to make unambiguous equilibrium predictions. Assuming individuals are selfish, however, co-operative game theory does unambiguously predict that no player will bargain to an outcome that is worse than their status quo payoff. Hoffman and Spitzer have used experimental bargaining games to show that this prediction is violated (Hoffman & Spitzer, 1985). Given a chance to work together in a co-operative endeavour, individuals seem willing to share payoffs equally even when their outside options are unequal.

1.11.6 The nature of human altruism

Ernst Fehr and a number of collaborators have recently worked on a synthesis of the results of experimental economics with evolutionary theory (Fehr & Fischbacher, 2003) (Fehr & Gächter, 2002a). They argue that empirical results such as the ones outlined above can only be explained by the presence of what they call strong reciprocity in human motivation (Fehr & Gächter, 2000b). Weak reciprocity is the form of altruism that can be seen as enlightened self-interest[10] – individuals do each other a good turn because they rationally expect to be "paid back". However, this limited form of altruism is not sufficient to explain the complex functional integration of human societies consisting of insufficiently genetically related members, and is also insufficient to explain the observed experimental regularities.

Strongly reciprocal altruism can take a positive and negative form, where individuals either help or harm others at true material cost to themselves. This acts as the "glue" that holds social institutions together, because the willingness on strong reciprocators to punish cheats even at cost to themselves forces selfish individuals to also behave themselves. Fehr and Gachter find that the ability of players to make costly punishments significantly increases the level of co-operation in the public goods game (Fehr & Gächter, 2000a).

The importance of the willingness of individuals to engage in altruistic punishment has also been reflected in recent work on cultural selection theory. Altruistic punishment is the main mechanism by which social organization can act as an altruism amplification device, because it is usually less costly to punish another individual (e.g. by ostracising them) than it is to make an altruistic sacrifice for their benefit (Sober & Wilson, 1999).

[10] Category (2) in the road map.

1.12 *Voluntary Giving*

The empirical result that the rational self-interest assumption does not correctly predict subjects' behaviour in experimental public goods games, along with the casual observation that people do in fact contribute en masse to public goods using both their time and their money, has led to attempts to develop microeconomic models of charitable giving that fit real human behaviour.

Andreoni constructs a model of contributions to a public good, where individuals must choose how much of their wealth to donate to the public good, with the rest being privately consumed (Andreoni, 1990). He shows that a pure altruism model, where individuals care only about their consumption of the private good and the overall amount of the public good provided, fails to explain important empirical phenomena. A pure altruism model predicts that forced donation to the public good (e.g. through taxation) should be completely crowded out by reduced contributions, because the individual's budget constraint, given expected contributions by everyone else, will remain unchanged, and so the optimally chosen combination of private and public goods consumption will also stay the same. The evidence, however, shows that crowding out is much smaller than this.

A pure altruism model also predicts that, assuming everyone has identical wealth and preferences, subsidising voluntary donations by a certain amount will be identical to taxation because the subsidy per unit of donation will be removed from the representative individual's wealth level, leaving the opportunity cost of donating unchanged. This renders the policy of subsidising donations unjustified.

By replacing the pure altruism model with a model of impure altruism where individuals also care about their specific donation, the empirical regularities can be more adequately explained, and the subsidization of private giving can be shown to result in a greater increase in contributions than the equivalent direct expenditure on the public good from taxation. Sugden similarly argues that pure altruism cannot explain the observed behaviour of contribution to public goods (Sugden, 1982).

Frank et al. find evidence for the theory of warm glow as opposed to pure altruism in data on charity care by private nonprofit hospitals in the US (Frank et al., 1996). McGranahan finds evidence from 17th century wills which supports the existence of a combination of altruism and warm glow motives for charitable bequests (McGranahan, 2000). Freeman studies volunteer labour supply and finds results compatible with the warm glow hypothesis; individuals are discerning about the causes they support and make rational decisions whether to donate time or money (Freeman, 1997).

Sugden explicitly models reciprocal preferences by considering individuals who follow the moral norm that "if everyone else provides at least their fair share in my eyes then so should I" (Sugden, 1984). This turns the public goods provision game from a prisoners' dilemma into an assurance game, where there will be multiple equilibria but only one of which will be Pareto efficient. This is an impure form of altruism because people's perception of a fair contribution depends upon their valuation of the public good, whereas a pure altruist would value it at the sum of the valuations of all individuals.

1.13 *The Economics of the Family*

1.13.1 The "rotten kid" theorem

The application of economic methodology to the study of the family was pioneered by Gary Becker. The most famous result to emerge from his work is the "Rotten Kid Theorem" (Becker, 1974) (Becker, 1981). This could be thought of as a theory of how the family can act as an altruism amplification device. The theorem states that all members of a family will behave efficiently, even if they are completely selfish (or imperfectly altruistic) provided that the head of the family is sufficiently altruistic to make an operative transfer. (The head of the family is defined as the member who has sufficient private income to make positive transfers to all other family members due to altruism.) The head of the family thus provides an "altruistic linkage" between all other members.

Suppose we examine the decision of a "rotten kid" (selfish family member) over whether to take an action that increases or decreases their pre-transfer income at the expense or gain of another family member. The head of household will take into account the decision made by the rotten kid when deciding how big a transfer of resources to give him. In other words, the head of the family ensures, through the alteration of the size of the transfer, that the rotten kid gets a share of the total family wealth as determined by their need relative to other family members, as perceived by the family head. Any action which increases the overall collective family wealth therefore makes the rotten kid better off. The rotten kid thus behaves fully efficiently; the operative altruistic transfer from the family head ensures that they fully internalise the costs and benefits their actions cause to other family members.

The technical conditions required for the Rotten Kid Theorem to hold were clarified by Bergstrom, who concluded that, although not universally applicable, it is still valid in many situations (Bergstrom, 1989). The requirement is that the actions taken by the rotten kids must be of such a nature that they cause a shift rather than a change in the slope of the family's utility possibilities frontier. This ensures that the family head, provided she is non-paternalistic and benevolent towards all family members, will definitely reallocate income so that the rotten kid is better off individually when he takes an action which expands the family utility possibilities frontier.

When rotten kids can take actions that distort the slope of the utility possibilities frontier, the rotten kid theorem no longer applies and they may take inefficient actions from the perspective of the family. It has been pointed out (Bruce & Waldman, 1990) that one of the main situations in which this condition does not apply is when there is a moral hazard problem between the head of the family and the rotten kid due to certain goods in the rotten kid's utility function not being under the direct control of the head of household via the transfer process. This is also known as the "Samaritan's Dilemma". Andreoni has described the problem as that when parents can be manipulated by children into "spoiling" them then they do become "spoilt rotten" (Andreoni, 1989).

1.13.2 Evidence on transfers and "altruistic linkage"

The Rotten Kid Theorem emphasises the importance of transfers between family members. One of its main empirical predictions is that redistribution of income between family members should not affect the distribution of consumption. This prediction has been tested on data on the US extended family (Altonji et al., 1992) and the Japanese extended family (Hayashi, 1995). Both studies found conclusively against the hypothesis that extended families are "altruistically linked" by a single head. Although this does not necessarily show imperfect altruism on the part of individuals within an extended family, since there are other factors such as asymmetric information and non-operative transfers in play, it is at least suggestive.

Empirical evidence (Altonji et al., 1997) also shows that the theory of pure altruistic linkage cannot explain other phenomena such as the responsiveness of inter-vivos transfers to changes in parental and child income. Altonji et al. argue that introduction of a "warm glow" factor into the model could allow this to be better explained. Bernheim and Bagwell have argued that the central flaw with the traditional model of the altruistically-linked dynastic family is the single-linkage assumption, because, once multiple linkages are recognized as part of "dynastic networks", such a framework has absurd implications which are clearly empirically false (Bernheim & Bagwell, 1988).

The altruism-based model of transfers within the family has also been challenged by the “exchange” approach, which sees the family as an institution enabling insurance contracts to be made between self-interested members. It is difficult to settle the issue empirically due to the similarities in the predictions of both frameworks (Kotlikoff & Spivak, 1981). Some studies have found that exchange is a more important explanatory factor than altruism (Cox, 1987) (Cox & Rank, 1992). Altruism has still often been found to improve the explanatory power of models which explicitly take into account such possibilities, however (Tcha, 1996) (Sloan et al., 2002). It has been argued that the altruism, exchange and warm glow models should not be seen as mutually exclusive and should be integrated together in a pragmatic manner (Stark & Falk, 1998).

1.13.3 Bargaining models of the family

The empirical failure of models based upon the notion that decisions within the family can essentially be modelled as being made by a single head has led to a body of work applying bargaining theory to decisions within the family. The most commonly used framework is that in which a husband and wife with differing interests bargain over the allocation of resources within the family. It has been found that although the empirical evidence rejects the hypothesis of decisions being made in accordance with a single household utility function, it does fit the hypothesis of efficient bargaining between members with differing, even if partially altruistic, utility functions (Chiappori, 1992) (Browning & Chiappori, 1998). This therefore constitutes strong evidence for imperfections to altruism even within the nuclear family. Research on divorce has found evidence that such efficient bargaining breaks down, resulting in a reduction in child welfare (Weiss & Willis, 1993) and the idea of bargaining between men and women has also been used to explain out-of-wedlock childbearing (Willis, 1999).

1.13.4 Biased altruism

Some of the most interesting evidence on the imperfections to altruistic motivation within the family comes from the study of unequal treatment of children on the basis of gender, birth order and biological relatedness. Researchers into differentials in wages and human capital investment between males and females in the US (Behrman et al., 1986) have attempted to determine whether or not this is driven by greater weight upon the success of male children in parents’ altruistic utility functions. The conclusion of this study was that existing wage differentials reinforce inequalities in human capital investment, but that parents do not, at root favour boys (in fact, if anything, the raw weighting on the welfare of girls is slightly higher). A similar study of the Phillipines (Davies & Zhang, 1995), however, found evidence for pure gender bias in parental preferences, underlining the fact that altruistic imperfections are culturally relative.

Research has also been conducted into why first-born children tend to do better in terms of measures of life success than children born subsequently to the same parents (Behrman & Taubman, 1986). The conclusion is that this is due to natural endowment effects rather than parental favouritism (as with gender bias, it may again be the case that parents compensate by weighting later-born children slightly *more*). Evidence on the treatment of step children in the US (Case et al., 1999), on the other hand, suggests that they do receive a smaller proportion of family income on food if they live with a stepmother, after controlling for income.

1.13.5 Discount rates

Research into the discount rates parents are revealed to apply to their children's welfare in making costly decisions related to children's health, as measured by lead contamination, has shown them to be similar to market interest rates for wealthier parents and higher for poorer parents, but nowhere near as high as the discount rates of 20% - 50% found to be applied to consumer durables (Agee & Crocker, 1996). This suggests a strong degree of parental altruism towards children.

1.14 *Bequests*

One of the most important topics concerning altruism in economic theory is that of bequests, most often from parents to children. There are two dimensions to this issue. The first is the microeconomic question of what motivates people to leave bequests. The second is the macroeconomic question of the implications of bequests for the role of government debt, income inequality, savings and investment and economic growth. Evidence in both these areas is potentially relevant to the existence of altruism, and its imperfections, because the answers to questions in the dimension of motivation have macroeconomic implications, and vice versa. There is a massive body of empirical research in this area, so we will only be able to take a cursory look at the main issues and findings.

Bequests have been estimated to account for about 80% of net private wealth in the US (Kotlikoff, 1988). A number of potential explanations have been offered for why parents leave bequests.

It may be due to a genuine altruistic concern for children's well-being. Such altruism may be of the "pure" or "warm glow" kind, as we will discuss in further detail below. Bequests may, on the other hand, be accidental, as a result of the parents dying earlier than they expected, and thus being unable to consume their entire wealth. A more sophisticated variation on this theme, which can potentially explain a greater role for bequests, is the theory that children and parents enter into a contract that parents will leave positive net wealth in return for children supporting them if they live longer than expected.

The microeconomic evidence for the importance of altruism in determining bequests is mixed. However, a reasonable conclusion would seem to be that there is genuine but imperfect and heterogeneous parental altruism towards children. Studies have found that inter-generational altruism is a plausible explanation for observed bequest behaviour (Adams, 1980), that individuals value bequests as highly as their own consumption[11] (Kuehlwein, 1993), that bequest behaviour reflects a high degree of heterogeneity, arguably driven by differences in preferences, even among a high income US sample (Laitner & Juster, 1996), and that there are statistically significant differences in bequest behaviour due to differences in income level, race and head of household gender (Kurz, 1984).

A pure altruism model of parental bequests is problematic. If parents are directly concerned with child utility, they should give different bequests to different children based on the children's income. Empirical evidence, however, does not support this prediction (Wilhelm, 1996). Parents tend to give equal bequests regardless of their children's differing situations. This suggests that some kind of impure "joy-of-giving" model where it is the size of bequest that enters directly into the parental utility function would be more realistic. It has been shown (Abel & Warshawsky, 1988), however, that in terms of the overall macroeconomic implications, the two theories are interchangeable in that empirically observed joy-of-giving parameters imply imputed coefficients of altruistic preference for children's welfare that are not implausibly high.

The resistance to accepting the primacy of parental altruism as the cause of bequests comes primarily from its macroeconomic implications rather than from the microeconomic sphere. In a seminal paper (Barro, 1974), Robert Barro showed that if parent and children are altruistically linked into a "dynasty" by a planned positive transfer of wealth (an operative transfer) either from parent

[11] Interestingly, this was even found to apply to individuals without direct heirs, providing evidence for more generalised altruism also.

to child or from child to parent, then cutting taxes and issuing government debt will have no effect on consumption expenditure because it does not alter the dynastic budget constraint out of which the altruistic parent (or child in the less frequent case of a child making the operative transfer to a parent) distributes income. This is the famous Ricardian Equivalence Theorem, which implies that government debt does not either crowd out private investment or cause a trade deficit. The traditional Keynesian view, however, is that government debt does crowd out investment and net exports. Although more recent attempts to weight the evidence have questioned the empirical rejection of Ricardian equivalence (Seater, 1993), other strong neutrality results not borne out by empirical evidence can also be shown to hold if the altruistically linked dynastic model is true, such as the invariance of inter-temporal consumption to inter-generational redistributions of income. (Kotlikoff, 1988).

It is thus probable that not all families are dynastically linked, because they do not have sufficiently strong altruism towards their children to make the transfer operative. (Note that, as argued by Bernheim and Bagwell, the observation that some people do not have children is not a good basis for an argument against Ricardian Equivalence because, going back far enough, all human individuals are related (Bernheim & Bagwell, 1988).) These are, of course, only a few among many reasons why these neutrality results are unlikely to hold fully in the real world. The use of the altruistic bequest motive in overlapping generations policy simulation models has now become commonplace (see, for example, (Altig, David et al., 2001)) and the consensus view would seem to be that it provides the best fit for building realistic models (Rangazas, 1996) compared to alternative theories.

1.15 *Environmental Goods*

Environmental goods are one of the most commonly examined types of public good for which there is evidence that individual valuations depend to some degree upon altruistic preferences. Much of this evidence comes from the use of the contingent valuation methodology, which asks survey respondents to give a valuation of a particular environmental good in a particular context. This method can be justified (Hanemann, 1994) on the grounds that its results have been found to match up reasonably well in specific cases with the findings of the traditional revealed preference approach. Hanemann concludes that although contingent valuation is subject to potential biases and inaccuracies, so also is revealed preference.

A study of voting behaviour in a Californian referendum on tightening water quality regulation (Holmes, 1990) found evidence for altruistic as well as selfish motivation for voting behaviour. A proxy was constructed for an environmental altruism variable using the residuals from regressing the results of an earlier Senate race, fought between a Democrat with a strong environmental voting record and a business Republican, on variables to proxy partisanship and a vector of self-interest variables. It was found that this proxy variable had a statistically significant impact upon voting in the referendum. This suggests an independent role for altruistic preference in contingent valuation as evidenced in voting behaviour, although the effect would be reduced and possibly eliminated if some variables were missing from the vector of self-interest variables.

Research on the relationship between environmental valuations reported in survey data and life expectancy (Popp, 2001) has also found evidence of a role for partial altruism. If people are fully altruistic, their life expectancy should not affect their contingent valuation of environmental goods. On the other hand, if they are fully selfish, the valuation should be, on average, zero as life expectancy goes to zero. The evidence, however, rejects both these hypotheses, suggesting the presence of partial altruism. The central estimate is of an equal weighting between individual welfare and the average welfare of future generations, but this estimate is very sensitive to the assumed discount rate.

Weak evidence has also been found (Weaver, 1996) of a role for altruistic valuation of environmental goods in *production* as well as consumption decisions. The decisions of Pennsylvania farmers to invest in particular technologies were found to depend upon the valuation of environmental goods as well as narrow profit motives.

The precise manner in which additional factors over and above self-interest drive the contingent valuation of environmental goods remains unclear. Some evidence indicates that the altruism here is connected with the desire of individuals to make a “fair” contribution rather than being driven by pure altruism (Stevens et al., 1994).

1.16 *International Comparisons*

Some of the most striking evidence for imperfections to altruism comes from the realm of international political economy. A 1998 study which sought to estimate the marginal cost of additional life expectancy in different countries found that the implicit valuation of a life year in the richest countries was 300 times that in the poorest countries. Once difference in average life expectancy are taken into account, the cost of saving an entire lifespan in the richest countries came to 1000 times that of the poorest countries (Dowrick et al., 1998).

One of the most documented phenomena in comparative political sociology is the difference in the extent of the welfare state between the US and the EU countries.[12] It has been argued (Alesina et al., 2001) that the most convincing explanation of these differences lies in differing attitudes to the poor, with Americans being more likely to view the less well off as lazy and responsible for their own situation.[13] This was demonstrated by regressing responses to an attitude survey on various possible causal variables. It was found that the dummy variable on the US remained statistically significant, demonstrating "American exceptionalism". The authors of this study tentatively suggested that racial fragmentation may be the causal factor, since the inclusion of a racial fragmentation index resulted in the dummy variable on the US becoming statistically insignificant.

1.17 *Normative Limits*

The volume of mainstream economics literature provides a number of powerfully suggestive reasons why altruism (of the fully utilitarian rather than "average utility" form) will have costs as well as benefits to society. These come about through the abandonment of the stringent assumptions of the non-twisting theorem (essentially also those of the perfect competition general equilibrium model).

One area that has been given extensive consideration is the compatibility of profit-making behaviour with altruistic motivation. Is an altruistic society necessarily a non-market one, and is it necessary for individuals to be selfish for a market society to operate efficiently? Becker has made the point that if owners of firms are altruistic towards other individuals, it is more efficient to give money away to the poorest than to subsidise the price of goods below marginal cost (Becker, 1981). This of course assumes that the market in which the altruistic firm operates does not in the first place suffer from market failures such as externalities, which would require self-imposed "subsidies" or "taxes" of a fully altruistic firm in order to internalise the externality.

[12] General government spending makes up 36% of GDP in the US, but an average of 48% in the European Union. Subsidies and transfers make up 11% of GDP in the US and 20% in the EU, thus accounting for three-quarters of this difference. The primary difference would seem, therefore, to be in the redistributive role of the state.

[13] A similar study (Freeman, 1984) found that differences in "personality type" in different countries helped to explain the size of the welfare state.

Even in the presence of market failure, the desirability of acts of “corporate social responsibility” also depends heavily upon informational considerations. Altruistically motivated behaviour requires that agents know the preferences of others, so that they are able to act so at to make them happier. A number of economists (Baumol, 1975) (McKean, 1975) have argued that whilst moral systems based on altruistic motivations work well in the case of maxims like “Do not be rude to people”, they are unlikely to work in the case of rules such as “Do not produce too much pollution” because the issue of how much pollution is socially optimal depends on information which simply is not available in such a form as to elicit a social consensus, even among a society of perfect altruists. Thus some issues should be left to regulation through the democratic system rather than unilateral attempts to act in an “ethical” manner by firms. The democratic system will be better able to predict, and thus regulate, firm behaviour if it can expect that they will act in a clear profit-maximising manner.

Limitations to the potential for altruistic socially responsible behaviour also apply to individual consumers as well as firms. For example, take the issue of the choice to buy organic foods. Assume firstly that these are more expensive than the non-organic versions and secondly that the felicity from consuming the two alternative products is the same. Assume also that both markets are competitive and operate in the same regulatory system so that there are no issues of monopoly or exploitation of labour. Compare two individual policy options. (1) Buy the more expensive organic product. (2) Buy the cheaper normal version and donate the relative saving to the poor. If the organic product is more expensive because it reduces a harmful externality with value in excess of the price difference, then there is a trade-off between the more efficient option (1) and the more equitable option (2). If, on the other hand, the organic option simply uses more resources (e.g. land, labour time) then option (2) is superior on both efficiency and equity grounds. It may be very difficult for a consumer to have the necessary information to make the optimal decision.

There are normative limitations not just to particular manifestations of altruistic behaviour, but to altruistic preferences themselves. A major possible drawback from altruistic motivation is that it may serve an antisocial purpose by helping specific groups to achieve rents via collective action (e.g. business lobbies, labour unions, cartels). As the application of the economic theory of public goods to political collective action shows (Olson, 1965), groups whose members feel more altruistic towards other members of the same group will be more successful at overcoming the free rider problem in their political collective action, and will be able to extract rents from the rest of the population in a socially inefficient manner. If it is very difficult to achieve generalised other-regarding preferences (as seems likely) then it may be better for people to be fully selfish rather than partially and selectively altruistic.

A specific application of the nonmonotonicity of altruism due to collective action effects is in the structure of collective wage bargaining institutions (Calmfors & Driffill, 1988). If unions are able to overcome the collective action problem at the industry level, this can result in higher wages and higher unemployment than a competitive labour market, or where collective bargaining occurs at the firm level. This is because the labour demand curve faced by a union is less elastic at the industry level, because consumers cannot go elsewhere for substitutes. If bargaining occurs at an economy-wide level through an incomes policy, on the other hand, the central union knows that it will only push up prices if it pushes up all wages, and so will exercise more restraint in nominal wage demands.

Moral hazard between altruist and beneficiary is another cause of normative limitation to altruism. This kind of effect relies upon some existing asymmetry in the altruism level of individuals, since if all individuals were fully altruistic they would not be tempted to do anything harmful to others. Stark and Bernheim present a version of the Samaritan's Dilemma (Bernheim & Stark, 1988). Take a three-stage game. Agents A and B each start off with a certain level of resources. First agent A decides how much to consume in period 1. Agent B then chooses his period 1 consumption. Agent A then chooses her period 2 consumption, transferring her residual resources to agent B. Suppose also that agent A is partially altruistic towards B whereas agent B is fully selfish. Agent B will have an incentive to consume an inefficiently high amount of his resources at stage 2, since he knows that agent A will compensate him for this at stage 3. In order to reduce agent B's incentive to do this, agent A will then consume an inefficiently large amount of her wealth at stage 1. This can lead to an outcome where higher altruism on the part of agent A results in a less efficient outcome.

1.18 *Consequences*

A central argument of this survey is that it would be desirable for variations in the level of altruism exhibited by agents to be a standard feature in analytic modelling, in a similar manner to imperfect information. This may potentially have far-reaching implications. Firstly, the impact of the presence of altruism has been analysed in the context of the theory of cost-benefit analysis. It has been shown that in the presence of non-paternalistic altruism, household willingness to pay for a public good exceeds the sum of individuals' willingness to pay (Quiggin, 1998). An attempt to estimate the results of the presence of paternalistic altruism on the value of statistical life by calibrating to 2.5 individuals per altruistic family predicted that this value is 10%-40% higher than the individual value (Jones-Lee, 1992).

A number of overlapping generations models of environmental degradation have shown that the presence of partial altruism does not guarantee an efficient internalization of these externalities (Jouvet et al., 2000) (Turner, 1997). It has also been found that co-operation between nations to internalise current environmental externalities may lead to a deterioration of future environment relative to non-co-operation because improved environmental technology frees up more resources for consumption (John & Pecchenino, 1997). The first-best solution requires internalization of the inter- as well as intra-generational externalities via a specific optimal form of altruism.

The final area where the normative impact of altruism has been considered is in the literature on optimal taxation. It has been shown that altruism can either increase or decrease optimal Pigovian taxation depending on its precise form (Johansson, 1997). The optimal subsidy for voluntary giving has also been found to depend in a dramatic way upon the motivation for giving, particularly whether it is of the warm-glow or pure altruism kind (Kaplow, 1998).

Altruism has interesting consequences for many areas of economic theory. It should have a regular status alongside the self-interest assumption. However, the prevalence of altruism in human behaviour, along with its manifest imperfections, raises the much more fundamental question of whether the level of altruism in human society as a whole is suboptimal or superoptimal.

1.19 *The Sequential Punishment Model*

> [I]n comparison with a situation wherein altruism is absent altogether, the prevalence of just some altruism could result in Pareto *inferior* outcomes. Hence, if the formation of altruism may not only fail to do any good but may actually make things worse whereas the formation of sufficiently high levels of altruism is almost always beneficial,...a troubling discontinuity arises: to the extent that the formation of altruism is like the rising of bread dough (i.e. it *has* to be gradual) groups yearning to build up their social stock of altruism may have to endure Paretial *deterioration* before experiencing Paretial gains. Perhaps one reason why a great many societies consist of self-interested economic men and women rather than altruistic economic men and women has to do with this nonmonotonicity.
>
> (Stark, 1989)

These are stimulating conjectures from Stark and Bernheim. “The Socially Optimal Level of Altruism” (Chapter 2) corroborates many of them. It makes a contribution to the existing literature on the “limits to altruism” by using an appealingly abstract and general model of social interaction to illustrate that altruistic preferences on the part of individuals can often be unnecessary, and even counterproductive, from a social welfare perspective. More altruism is not always better for society. The most valuable additional insights offered are, on the one hand, that even the very high levels of altruism exhibited by individuals who care almost as much about others as themselves can sometimes be insufficient to “get the dough to rise” but that, on the other, it can be the case that even malevolent individuals who wish to harm one another could be induced to behave socially beneficially with the appropriate system of social control, and that this need not be more difficult than with altruistic individuals. Central to this result is the fact that altruism has multiple opposing effects upon incentives.

In the sequential punishment model, individuals receive opportunities to harm one another, one after another. If they inflict harm, they gain a random amount of felicity (whose value is known in advance), and the person they choose to hurt loses one unit of felicity. Individuals have free choice, and are indifferent, as to whom they punish. In a single-move game, individuals clearly must be sufficiently altruistic in order to prevent them from giving in to the temptation to harm others for their own benefit; altruism is always a good thing, because it reduces the temptation of individuals to misbehave (this is therefore referred to as the “temptation effect”). When there are a number of sequential moves, however, individuals can, contingent upon earlier observed misbehaviour, punish selfish miscreants. This creates the possibility of counterproductive altruism, both because more altruistic individuals are less willing to punish (the “willingness effect”) and also because they are harder to hurt by threatening them in place of others (the “severity effect”).

The sequential punishment model has relevance to issues which have arisen in recent empirical work on altruistic behaviour. The use of altruistic punishment has come to be seen by a number of empirical economists, influenced by a cultural evolution perspective, as the key to explaining how imperfect individual altruism can be “magnified” to achieve socially efficient outcomes (Fehr & Gächter, 2002a). Altruistic punishment is thus central to the operation of the social structure as an altruism amplification device. The presence of some individuals who are willing to punish others, even at harm to themselves, is at least as vital to achieving social efficiency as the presence of “traditional” altruists.

Players are assumed to be indifferent as to which other player they punish. This may seem an objectionably strong assumption, but it is a reasonable simplifying one since it is tantamount to assuming that preferences over who is punished are sufficiently small to enable punishment schemes to be incentivized. The idea that society may have punitive actions which can be cheaply directed onto miscreants is also an important one in experimental economics and cultural evolution theory. For example, Fehr et al. (Fehr & Gächter, 2002b) argue that altruistic punishment, the willingness to harm others at cost to oneself, is vital in making altruism work as the glue of human society, because it is usually less costly to harm others than to help them, and this enables cheaters to be credibly threatened with punishment. Sober and Wilson (Sober & Wilson, 1999) similarly argue that an important reason why human societies exhibit complex functional integration is the low cost of punishing transgressors against social norms (for example, by simply refusing to interact with them).

The complete solution to the model uses optimal punishment paths, as developed by Abreu for infinitely-repeated stage games (Abreu, 1986) (Abreu, 1988) (Lambson, 1987). The analysis of the sequential punishment model succeeds in showing that the severity and willingness effects sometimes outweigh the temptation effect; greater altruism can make it more difficult to achieve a socially efficient outcome. As the coefficient of altruism θ approaches 1 from below (i.e. as individuals become perfectly altruistic), the socially efficient equilibrium will always break down. Hence, to put the central point in a "nutshell", too much altruism is bad for society.

The introduction of some degree of preference by individuals over who they punish, via heterogeneous weightings applied to each other individual in the utility function, would greatly complicate the results of the model in that the conditions for co-operation with ongoing punishment would be different depending on the punisher and punishee. This would not, however, qualitatively change the property that punishment can be imposed to incentivize co-operation with a socially efficient initial path provided players are sufficiently patient. Equally importantly, although the precise result of the interaction between the temptation, willingness and severity effects would become more complex, their key role in driving the model would remain, and the result that too much altruism can sometimes be socially damaging would not be overturned.

Stark and Bernheim (Bernheim & Stark, 1988) have also considered this issue, and have established a theorem which shows that as altruism becomes perfect (by which we mean that equal weighting is placed on the welfare of others relative to oneself), the Nash equilibrium outcome in a repeated game becomes arbitrarily close to the socially efficient outcome. Although it is strictly speaking a sequential rather than a repeated game, the sequential punishment model exhibits this property. However, it also exhibits the property that intermediate levels of altruism enable full social efficiency to be achieved, and that high enough levels of altruism which are still less than perfect cause the socially efficient equilibrium to break down, with a non-negligible negative effect on the efficiency of the equilibrium.

The sequential punishment model treats altruism in a general and abstract way, and provides good reasons to conclude that intermediate rather than high levels are the most socially desirable. Stark and Bernheim show an awareness of these possibilities in the following passage:

> While high levels of altruism are generally beneficial, this should provide little comfort to those who subscribe to the traditional views...The effects of altruism on economic interaction are complex, and must be assessed on a case-by-case basis.
>
> (Bernheim & Stark, 1988)

Stark and Bernheim discuss one of the factors which is of central concern in the sequential punishment model, in the form of the willingness effect. This they describe as the possibility that a more altruistic individual will be perceived as a "softy". Stark and Bernheim show, using a simple infinitely-repeated co-operation game, that altruism can, for this reason, make an efficient co-operative outcome harder to achieve. They also hint that a full analysis of optimal punishment could provide richer results in this area:

> In analysis of repeated games, it is standard practice to enforce co-operative outcomes through "Nash reversion"...While more severe punishments are frequently available..., Nash reversion is by far the simplest and most widely discussed method of enforcing co-operation. We employ it here.
>
> (Bernheim & Stark, 1988)

The sequential punishment model, although not a true repeated game, turns out both to be simple enough, and to share enough features with repeated games, to enable a satisfyingly full characterization of the optimal punishment paths, and thus a complete analysis of the effect of altruism on the sustainability of socially efficient equilibria. The most important general result from the model is that high levels of altruism close to, but not quite reaching, perfect utilitarian altruism unambiguously break the supportability constraint on the socially efficient outcome.

The sequential punishment model is intended to capture a general feature of the economic world in a simple but abstract manner. Other models with a similar idiom include Samuelson's "pension game" (Samuelson, 1958) and Diamond's model of fiat money in a "coconut economy" (Diamond, 1984). In Samuelson's model, an infinite series of individuals, each of whom lives for two periods, must decide whether to make a gift to the individual who is old when they are young. Young individuals have an endowment of one "chocolate", but the old have no endowment. Multiple subgame-perfect Nash equilibria exist, some where chocolate pensions are voluntarily given (because it is believed that receiving a pension from the next generation will be conditioned on having donated to the previous generation) and some where the pension system never "gets off the ground". This model captures, in a simple and elegant manner, the key features that make pension systems (either state or private) fragile, since they depend on the belief that the next generation will provide.

Diamond's model captures an essential point of economic life relevant to the microfoundations of macroeconomics. Production is only profitable when trading partners can be found. This creates the potential for multiple equilibria, with high and low levels of economic activity. The economy becomes a giant co-ordination game. The "parable" for the Diamond model involves a society of individuals who live on an island consisting of palm trees of varying height, each with a coconut at the top. Climbing trees is costly. Individuals can only gain utility from consuming their coconut by finding another individual to "swap" with. Each individual can only carry one coconut at a time. Whether or not it is worth climbing a particular tree therefore depends upon how many potential trading partners there are "out there". This simple abstract framework, and neat intuitive back-story, captures perfectly the existence of multiple search equilibria in a macroeconomic system which uses fiat money.

1.19.1 The evolutionary sequential punishment model

In a second paper, "Punishment and the Potency of Group Selection" (Chapter 3), a different aspect of the relationship between altruistic preferences and punishment systems is explored, and a new twist to the story offered. A simplified three-player two-move finite version of the sequential punishment model is used, but the preferences of the individuals playing the game are now permitted to evolve over time. It is shown that the use of punishment equilibria weakens the potency of the group selection mechanism, making it harder for altruistic preferences to evolve.

The intuition for this result is that group selection depends upon more altruistic groups doing better on average than more selfish groups. However, by making selfish individuals behave better, the use of a punishment system weakens this performance differential, thus weakening group-level selection pressures. The normative consequences of this phenomenon are ambiguous because there are two counteracting effects. The static effect improves social welfare, because, given a certain population composition, the use of punishment makes people behave better. The dynamic effect, that fewer altruists are able to evolve, can sometimes, however, outweigh the static gain, so that society is actually worse off in the long run when using a punishment system.

1.20 *Conclusion*

We have examined disparate evidence from a number of areas of economics in order to establish the pervasive presence of altruism in human motivation, but also the prevalence of imperfections. This survey has explored a number of positive and normative theoretical reasons for this phenomenon. In particular, it has presented a broader context for the innovations in the fields of economic and cultural evolution theory offered respectively in "The Socially Optimal Level of Altruism" (Chapter 2) and "Punishment and the Potency of Group Selection" (Chapter 3). The first of these papers shows that too high a level of altruism has a detrimental impact on the effectiveness of punishment systems to act as an "altruism amplification device". The second paper explores the reverse phenomenon - the use of a punishment system may weaken the ability of altruistic preferences to survive in a dynamic environment of cultural evolution.

The Socially Optimal Level of Altruism

> [W]hen altruism improves static non-cooperative outcomes, it lessens the severity of credible punishments. An altruist may well be perceived as a "softy" and his threats may not be taken seriously.
>
> (Bernheim & Stark, 1988)

> [T]he most efficient way to provide low payoffs, in terms of incentives to cheat, is to combine a grim present with a credibly rosy future.
>
> (Abreu, 1986)

2.1 *Introduction - A Parable*

The central problem of economic and social policy, indeed the essential prerequisite of the social order itself, is that of bestowing upon the individual agent the incentive to act in a manner which is beneficial for society as a whole. Such incentives can be *intrinsic* to the individual (altruistic preferences) or *extrinsic* (threats of punishment if the individual agent does not comply with the socially prescribed action). This paper analyses the interaction between these two alternative "technologies". We see that the two methods of achieving social order cannot be freely mixed at will, and that, in order for extrinsic incentives to work most effectively, it is necessary to limit the operation of intrinsic incentives, so as to avoid counterproductive interference between the two. An important implication is that the heavy (perhaps predominant) reliance upon extrinsic incentives in sophisticated human societies may in fact be the result of a socially optimal "policy mix", rather than a mere second-best correction for the inadequate intrinsic motivation to "do the right thing".

Altruistic behaviour has been fruitfully modelled in economic theory using infinitely-repeated stage games (Fudenberg & Maskin, 1986) (Abreu, 1986), and infinite dynamic sequential games such as models with overlapping generations (Samuelson, 1958) (Hammond, 1975) (Cremer, 1986). The traditional view is that apparent altruistic co-operation can occur in such models with self-interested agents through the use of punishment equilibria which can deter actions which would be individually optimal but socially sub-optimal in a non-repeated or non-sequential game. This paper provides a general workhorse model in which the results of such models can be extended in order to accommodate *bona fide* altruistic motivation. In common with the infinitely-repeated stage game model, agents are infinitely-lived and discount the future. In common with overlapping generations models, on the other hand, players move sequentially, and each player only gets to move once during the entire game.[1]

The sequential punishment model presented in this paper is intended to capture an abstract essential feature of the social and economic world in a simple but general manner. Other models with a similar idiom include the "Robinson Crusoe" economy (Ruffin, 1972), Samuelson's "pension game" (Samuelson, 1958) and Diamond's model of fiat money in a "coconut economy" (Diamond, 1984). Each of these models can be illustrated intuitively with the help of a simple "parable", as can the sequential punishment model.

Consider a desert island where individuals are sufficiently settled to have established their own "back gardens". Each individual is sitting in their garden drinking a cold beer. One by one, at regular discrete intervals, one inhabitant finishes their drink, and must decide whether or not to walk to the bin, or to throw their bottle into one of their neighbour's gardens. All gardens are adjacent, and each person's bin is a variable distance away (imagine a giant pie-shaped island, each garden being a "wedge" - everyone is sitting at the middle of the island). It is possible for each inhabitant to be threatened that, if they throw their bottle, everyone who subsequently finishes *their* beer will throw a bottle into the malefactor's garden. Sometimes this threat will be enough to make every inhabitant walk to the bin every time, leading to a socially efficient litter-free island. Sometimes the threat will not be enough, and some or all of the bottles will be thrown, leading to a socially inefficient outcome.

The central feature encapsulated in this model is the fundamental vicariousness of human social interaction. In any society, individuals are able to impose negative externalities upon one another for

[1] This may seem an unrealistic assumption, but it can be argued that we only need split a finite number of players into an infinite number of "egos" (Hammond, 1975).

personal gain, in myriad ways. However, the very existence of this problem also offers a potential solution, in that it creates the possibility of punishing miscreants who take such opportunities, by threatening *them* with harm in the future.[2] Another potential solution to the problem, however, is altruism; if people care about others, they will refrain from harming them.

This paper aims to show that these two alternative incentive technologies can interact with one another in a perverse manner. The question we will set out to answer is whether greater altruism on the part of the inhabitants of the island will always make it easier to achieve a litter-free island. The answer is a resounding and conclusive "no". In particular, it is shown that too high a level of altruism will in general lead to a worsening of the societal outcome. Altruism "dents" social incentive systems based upon extrinsic rewards or punishments.

2.2 *Overview*

It is commonly recognised that the repetition of stage games such as the prisoners' dilemma, or the Cournot and Bertrand oligopoly games, allows apparently altruistic behaviour to be incentivized, resulting in a Pareto superior outcome for the players.[3] However, such apparent altruism only reflects "enlightened self-interest". This paper introduces genuine altruistic motivation, exploring its effect on the achievability of a socially efficient equilibrium.

The well-known "Folk Theorem" establishes that, if players are sufficiently patient, any equilibrium which Pareto dominates the min-max payoff in the stage game can be supported as a subgame perfect equilibrium in the infinitely repeated game (Aumann & Shapley, 1992) (Rubinstein, 1979) (Fudenberg & Maskin, 1986). The key implication is that, with imperfect altruism leading to a Pareto-inefficient equilibrium in the stage game, if there is sufficiently low discounting of the future then a Pareto efficient outcome in the infinitely-repeated game can be achieved. This paper asks the reverse question: Given a certain positive level of impatience, how high or low can the level of altruism be in order for a first-best socially efficient (and therefore Pareto-efficient) outcome to be attained? It is shown that, in general, there must be a "Goldilocks" level of altruism, which is neither too high or too low.

[2] Throughout the paper, we will use "harm" to refer to the infliction of a negative externality and "punish" to refer to the specific use of such harm opportunities to construct punishment equilibria.

[3] This framework has important implications for such varied theoretical areas as environmental economics (international co-operation in pollution abatement can be seen as a repeated game) (Barrett, 1994) and competition policy (firms are sometimes able to collude in a repeated oligopoly game, to the detriment of overall social welfare) (Rees, 1993).

In a game in which individuals make sequential moves (or in a repeated simultaneous-move game structure), they are able to punish one another based upon their previously observed behaviour. Individuals who are less altruistic are more willing to harm others because they place a lower value on the cost to the person being harmed. We shall therefore call this the **willingness effect**. Agents are also more afraid of being harmed because they value their own welfare more relative to that of others. For example, if a criminal is fined a certain amount and the revenue spent on other individuals, this is a more severe punishment for a less altruistic criminal because the fact that the fine revenue is spent on others mitigates the effect of the fine on the criminal's utility by a smaller amount than if the criminal were more altruistic.[4] We shall therefore refer to this as the **severity effect**. Together, these two effects create a potential social benefit from individuals not being too altruistic. However, this must be balanced against the greater temptation towards wrongdoing by a less altruistic individual. Hence there is also a **temptation effect** from greater altruism.

Stark and Bernheim have observed that altruism can reduce the credibility of punishment if the potential punisher is perceived as a "softy" (Bernheim & Stark, 1988). They observe that this can lead greater altruism to have a negative impact[5] but argue that this must be analysed on a case-by-case basis. We establish a canonical framework in which to study the interaction of this willingness effect with the severity and temptation effects. This enables us to establish the stronger result that a high level of altruism will *in general* be socially detrimental. The central result of this paper is that, under certain fairly non-restrictive assumptions, the three effects conspire to render a socially efficient outcome impossible if the level of altruism becomes high enough.[6]

2.3 *The Model*

Suppose there are an infinite number of players and that distinct players, referenced by the period in which they move, each get a chance in sequence to impose damage upon another player. In period t, player t receives a **harm opportunity**, and must decide whether to accept or reject it, and a "target" for the harm opportunity, player A_t. If player t **rejects** the opportunity, then there are no changes

[4] The revenue from the fine could, of course, be "thrown away" in order to avoid this adverse effect, but this would be wasteful in that it would create a deadweight loss to punishment.

[5] This is precisely the phenomenon we term the willingness effect.

[6] In terms of the island parable, the assumptions we must make are that the cost of having a bottle land is weakly greater than the cost of walking to a bin at the edge of the island, that players move in sequence, with perfect information about past play, and that all individuals share the same coefficient of altruism θ (the weighting they place on the welfare of others relative to their own welfare) and discount the future by factor δ.

in felicity.[7] If player t chooses to **accept** their harm opportunity, then player A_t suffers a **cost** in **felicity** of 1 unit. If player t accepts, then he gains felicity equal to the **benefit**, π_t,[8] which is drawn randomly and independently from a distribution defined by the density function $g(\pi)$, with support $[\hat{\pi}, 1]$, where $0 \leq \hat{\pi} < 1$.[9] Therefore, $\int_{\hat{\pi}}^{1} g(\pi)d\pi = 1$. (The expected value of π, $\int_{\hat{\pi}}^{1} \pi g(\pi)d\pi$, is denoted $\bar{\pi}$.) All players publicly observe the value of π_t before player t moves.

The move made by player t in period t can be conceived to consist of choosing a **trigger level**, T_t, for the realized value of the benefit above which they will accept the harm opportunity[10], and the individual whom they harm, A_t. A player's strategy maps the observed past history at period t (including the observed value of π_t) to the move they make.

We will derive results assuming a continuous distribution of the benefit with bounded support. We assume throughout that $g(\pi)$ is twice continuously differentiable. We will frequently use as an exemplar the case where the benefit has a continuous uniform distribution. Here the probability density function for the distribution of π_t will take the value $\frac{1}{1-\hat{\pi}}$ between $\hat{\pi}$ and 1. Most straightforwardly, with $\hat{\pi} = 0$, $g(\pi) = 1$ between 0 and 1.

2.4 *Players' Preferences*

Players do not act to maximise their "private utility", or **felicity**. Instead, players act to maximise their **social utility function**[11], which is a weighted sum of the felicities of all players.[12] We assume all players are risk-neutral and share the same discount factor δ. We let the **coefficient of altruism** θ represent the weighting placed upon the felicities of others in each player's social utility function. We assume that θ is identical for all players and is always strictly less than 1.[13]

[7] See section 2.4 for a formal definition of this concept.

[8] π is used indicate the vector of random benefit values.

[9] Intuitively, this assumption is sufficient to ensure that no partially altruistic individual ever *wants* to be harmed. In terms of the parable from section 2.1, the cost of having a bottle land in one's garden is equal to the cost of walking to a bin at the edge of the island. Although this might seem a restrictive assumption, we can always, by taking the limit as the expected value of π, $\bar{\pi} = \int_{\hat{\pi}}^{1} \pi g(\pi)d\pi$, goes either to $\hat{\pi}$ or 1, examine the cases where the cost of a bottle landing is almost identical to the cost of walking to the bin ($\bar{\pi} \longrightarrow 1$), or almost always greater than the cost of walking to the bin ($\bar{\pi} \longrightarrow \hat{\pi}$). (See subsection 2.8.1.)

[10] Note that they do so *after* they observe the benefit value, so this is tantamount to choosing whether or not to punish for any given benefit value. It nonetheless proves useful later on to think in terms of trigger levels.

[11] We will frequently use the term "social utility" in order to emphasise the inclusion of altruistic preferences. However, the modelling role of the social utility function is similar to any standard utility function.

[12] Effectively, every individual's social utility function is a social welfare functional which aggregates the orderings represented by all players' felicities, and which satisfies the Pareto principle, independence of irrelevant alternatives and unrestricted domain. Ratio scale comparability (Roberts, 1980) must be assumed, with all individuals gaining 0 felicity in the state of the world where no harm opportunities at all are taken.

[13] We exclude from the outset "super-altruism", where players are "matryrs" who care about others *more* than themselves. Note, however, that we do allow for potentially infinite "malevolence" throughout the paper.

Definition 2.1: Let $f_{i,t}$ be the **felicity** of player i in period t:

$$f_{i,t} = \left\{ \begin{array}{ccc} -1 & \text{if} & T_t < \pi_t \text{ and } t \neq i \text{ and } A_t = i \\ \pi_t & \text{if} & T_t < \pi_t \text{ and } t = i \text{ and } A_t \neq i \\ \pi_t - 1 & \text{if} & T_t < \pi_t \text{ and } t = i \text{ and } A_t = i \\ 0 & & \text{otherwise} \end{array} \right\} \tag{2.1}$$

Let δ be the **discount factor**[14] and let θ be the **coefficient of altruism**:

$$0 \leq \delta < 1 \tag{2.2}$$

$$\theta < 1 \tag{2.3}$$

In period t, player t moves so as to maximize his **expected social utility** u_t, discounted looking forward:[15]

$$E_\pi \left[u_t\right] |_{\pi_1 \ldots \pi_t} = E_\pi \left[\sum_{j=t+1}^{\infty} \left(\delta^{j-t} \left(f_{t,j} + \theta \sum_{k \neq t}^{\infty} f_{k,j} \right) \right) \right] \Bigg|_{\pi_1 \ldots \pi_t} \tag{2.4}$$

2.5 *The Single-Move Game*

Consider first a single-move sequential punishment game in which a single individual (individual 1) has an opportunity to harm another (this can be thought of as a special case of the infinite-move game in which $\delta = 0$ so that there is no future). The individual's coefficient of altruism must be sufficiently high in order to prevent him from yielding to the temptation to inflict harm socially inefficiently, and so here the deleterious willingness and severity effects of greater altruism do not apply. In this simple case there is therefore no sense in which too much altruism is bad for society. It is socially efficient for a harm opportunity to be taken if and only if $\pi_1 > 1$. The individual receiving the harm opportunity (individual 1), meanwhile, will choose to inflict harm if and only $\pi_1 > \theta$, since he values 1 unit of harm done to another individual at θ.[16] The outcome can therefore only be socially efficient, meaning socially efficient for *all* possible revealed benefit values, if $\theta = 1$.

[14] The role of the assumption of discrete time periods with discounting of the future can be justified as the simplest way of capturing the idea that the technology used to detect deviation is imperfect and thus takes time (Cremer, 1986).

[15] The assumption that players are infinitely lived may appear restrictive, but its primary role is to simplify the model. Versions of the Folk Theorem have been proved for games with finitely-lived players and overlapping generations (Kotlikoff et al., 1988) (Kandori, 1992) (Messner & Polborn, 2003), and the general result is that having finitely-lived agents reduces, but does not eliminate, the possibility of supporting mutually beneficial equilibria in an infinitely-repeated stage game framework. It therefore seems reasonable to focus on the role of altruism by assuming away the issue of finite lifespans.

[16] We can assume that the individual receiving the harm opportunity will definitely harm another individual rather than himself because $\theta < 1$. We assume that if he is indifferent, he will not inflict harm.

2.6 *The Infinite-Move Game*

Once we introduce an infinite series of sequential moves, where different individuals are given an opportunity to inflict harm one after the other, the effect that altruism has on the willingness to punish and on the severity of punishment becomes important. The socially efficient outcome can now be achieved when individuals are less altruistic than the level required in the single-move game, through the use of punishment equilibria.

A socially efficient outcome can be achieved in the infinite-move game by using the information available, and the fact that players are able to choose whom they harm, to enforce credible threats of future punishment upon players who are tempted to socially inefficiently inflict harm in the current period. The ability of a player to inflict harm then plays the dual role of a temptation to impose a deadweight loss upon society at benefit to oneself, but also the opportunity for society to credibly threaten to punish those who do so. This means that there can then be some advantages to players being less than fully altruistic, since the threat of punishment is more severe the less altruistic players are, both because less altruistic players are willing to inflict harm more often (the willingness effect), but also because the loss of social utility from being harmed in place of another is greater for a less altruistic player (the severity effect).

It turns out that, provided there is sufficiently low discounting, the severity effect dominates and the lower constraint on the required level of altruism drops away because decreasing the level of altruism beyond a certain point always increases the severity of punishment more than enough to outweigh the increased temptation to deviate from the socially efficient equilibrium by inflicting harm (see Theorem 2.III). Most significantly, however, it transpires that too high a level of altruism will, for all values of the discount factor, *prevent* the socially efficient outcome from being achieved (see Theorem 2.II).

The possibility of achieving this kind of socially efficient subgame perfect Nash equilibrium begs the question, however, of how it is that the players are to be co-ordinated upon playing it. We might think of a social planner declaring the equilibrium that will be played, and then each player behaving unilaterally with no ability to communicate with the others. On the other hand, we could imagine a kind of "original position" (Rawls, 1999), where the players together agree to the planned equilibrium that will maximise their collective expected social utility. This is relevant since generally speaking although the equilibrium outcome will be socially efficient, the off-equilibrium behaviour prescribed by the punishment strategies will not be, and so there would be the temptation for a social planner to intervene in order to achieve a socially efficient outcome in the subgame beginning once someone has actually deviated. The social planner could, if able to alter the expectations of all players about the strategies being played by all the others, improve social welfare once someone has actually deviated by "letting bygones be bygones" and re-coordinating all players upon a new socially efficient equilibrium in the subgame starting in the current period.

The most stringent requirement we could put on punishment equilibria is that they be renegotiation-proof. This equilibrium concept refinement has been applied to other repeated games (Farrell & Maskin, 1989) (Benoit & Krishna, 1993), and shown to reduce the number of supportable equilibria. Sometimes, depending on the context, efficient subgame-perfect equilibria can be rendered unsupportable. The reason is that often the most severe punishments are not renegotiation proof because everyone would prefer to "let bygones be bygones" and renegotiate to a Pareto-superior path. In this paper, however, we stick with the requirement that punishment equilibria be subgame-perfect rather than renegotiation-proof. Intuitively, we assume that the social planner (or the community's decision-making process) is able to avoid the temptation to let malefactors "off the hook".

Figure 2.1 provides a preliminary schematic for the possible subgame-perfect socially efficient equilibria which are supportable for different values of θ and δ in the sequential punishment model. The most lightly shaded area A shows values of (θ, δ) where, by using "Nash reversion" (which requires each individual along an equilibrium path where a previous deviator is being punished to take their harm opportunity whenever $\pi_t > \theta$, as they would in a single-move game) a socially efficient equilibrium can be constructed. Individuals are incentivized to co-operate with the equilibrium due to the threat of focusing future punishment onto any deviator from the initial path.

The darker grey area B shows those values of (θ, δ) for which social efficiency can only be supported by using a punishment path more severe than Nash reversion. This requires that the individuals doing the punishing be required to go "beyond their comfort zone" by inflicting harm for values of $\pi_t \leq \theta$ for which they would not do so in a single-move game, due to their partial altruism. The main analytic task of this paper is to characterize the nature of the socially efficient equilibria that can be supported using the most severe available path. The black region shows those values of (θ, δ) for which social efficiency is not supportable, even with the use of such an optimal path. The central results of this

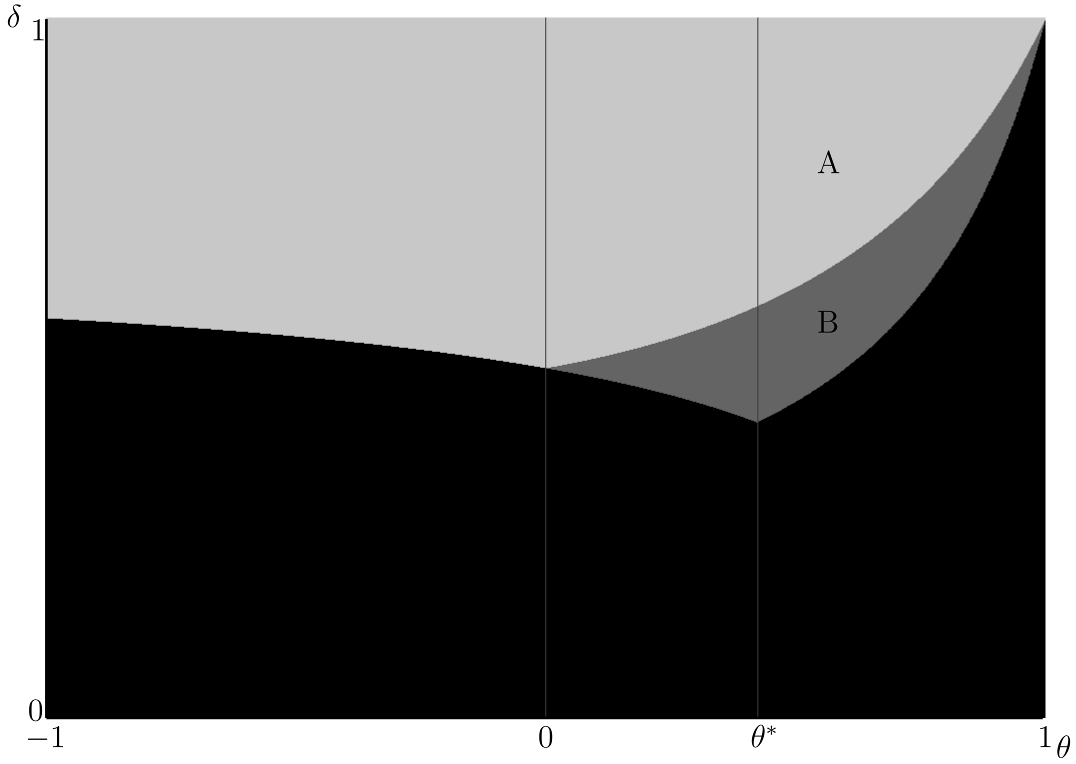

Figure 2.1: Socially efficient equilibria and the socially optimal level of altruism

paper are firstly that this region will be entered as $\theta \longrightarrow 1^-$ for all values of δ (see Theorem 2.II). Secondly, this region is "thinnest" at a **socially optimal level of altruism**, θ^*, and it will be shown that, under the fairly general assumptions made regarding the distribution of the benefit and the shared value of θ and δ, it is always the case that $0 < \theta^* < 1$ (see Theorem 2.IV).

2.7 *Punishment Paths*

The sequential punishment model has close parallels with the traditional framework of infinitely-repeated games with discounting. Seminal results for the nature of the optimal penal codes in these types of game were provided by Abreu[17] (Abreu, 1988), who showed that optimal punishment can be exhaustively described using **punishment paths**. These will in general have a **carrot-and-stick**

[17] Abreu also foresaw that his method would have far-reaching applications in other models:

> Analogues to the theorems established here ought to appear in any model with discounting and a "repeated" structure. Finally, the conceptualization of punishment in terms of paths and deviations from prescribed paths should prove useful in other contexts.
>
> (Abreu, 1988)

The sequential punishment model analysed here is one such context. Although the sequential punishment model is not strictly speaking a repeated stage game, the ability of individuals to condition their behaviour on the past, with deviations immediately observable next period, gives it an essentially analogous structure.

structure, with players incentivized to co-operate with the more unpleasant early stages of the path by the "carrot" offered by the return to more pleasant co-operative behaviour in the later part of the path. The introduction of non-stationary carrot-and-stick punishments is particularly interesting in the sequential punishment model because partially altruistic individuals must themselves be threatened with harm if they refuse to co-operate with the punishment of others. This feature of the model generates a rich interaction between the altruistic preferences of the players and the structure of optimal punishment paths.

Strategy profiles and the corresponding equilibria in the sequential punishment model can be described in terms of an **initial path** and a **punishment path**. Along the initial path, no harm opportunities are permitted to be taken. If a player deviates from the initial path, then a punishment path tailored for that player is initiated. If a player deviates from an ongoing punishment path, then a new punishment path tailored for the most recent deviator is initiated.

A punishment path, denoted by ψ, is a vector of trigger levels for π above which harm opportunities are taken in a punishment equilibrium. Punishment paths provide a natural way to conceive of punishment equilibria in the sequential punishment model. If a punishment path, which was initiated in period j through a deviation by player j, is being followed in period t, then player t sets their trigger level T_t equal to ψ_{t-j} (so that player t takes the harm opportunity when $\pi_t > \psi_{t-j}$) and punishes player j by setting $A_t = j$.

Definition 2.2: A punishment path, denoted ψ, is a vector of trigger levels, subscripted by the point reached along the path.[18] Trigger levels must lie within the support for π, therefore $\forall_\psi \forall_k : \psi_k \in [\hat{\pi}, 1]$. The set of possible punishment paths is Ψ, so that $\forall_\psi : \psi \in \Psi$. A **flat punishment path**, $\bar{\psi}$, has the property that $\forall_k : \bar{\psi}_k = \bar{\psi}$.[19] The set of flat punishment paths is denoted $\bar{\Psi}$.

Following Abreu's argument, in order to find out if the socially efficient outcome is supportable for any given θ and δ, it is in general necessary to derive the **optimal punishment path**. Along a punishment path, it will be desirable to harm the most recent deviator as much as possible.

[18] We use the term "period" to refer to "game time" and "point" to refer to the current position along an ongoing path.

[19] We use $\bar{\psi}$ to denote *both* a flat path *and* the constant trigger level that defines it. This simplifies notation significantly in subsequent lemmas and theorems. A number of functions will be defined later on as taking a path (a vector of real numbers) or a constant trigger level along a flat path (a real number) as an input. When we are dealing with flat paths, the two interpretations of the notation can be used interchangeably without causing any problematic ambiguity. When dealing with non-flat paths, however, the distinction between the two types of input must be kept in mind.

Since players are indifferent as to whom they harm, any harm opportunities taken along an optimal punishment path will therefore be "focused" upon the most recent deviator.

We can imagine choosing a fixed punishment, and then finding out the most severe path we can support given the use of that fixed punishment for any deviation. However, as argued by Abreu, we will only have found the most severe path we can support if we are in fact using that path to punish any deviation from any ongoing punishment path. Hence the optimal punishment path must be used to punish any deviation from itself. This is a useful recursive symmetry which we exploit in constructing the conditions for supportability in Definition 2.3.

There are two constraints at each point along a punishment path. The first concerns the "squeamishness" of partially altruistic individuals in implementing the "stick". Individual t is only willing to take a harm opportunity when $\pi_t \leq \theta$ if he is himself threatened with punishment, in order to give him an incentive to inflict harm when it is unpleasant for him to do so. The second constraint concerns the "carrot" part of the path. In order to provide a carrot, it is necessary that trigger levels be higher later in the path (so that harm is inflicted only for high benefit values). This may involve individuals being required to *abstain* from taking a harm opportunity when $\pi_t > \theta$, for which they will also need to be given an incentive via carrot-and-stick punishment.

The second constraint turns out to be more difficult to deal with, but we are able to prove that, as $\theta \longrightarrow 1^-$, this constraint becomes insignificant, because even when it is not imposed, the socially efficient outcome becomes unsupportable using the optimal path anyway. Also, in many cases the second ("upper") constraint will not bind at any point along the path, whereas the first ("lower") constraint must always bind at the beginning of the optimal path. It is therefore the first constraint which primarily drives the shape of optimal punishment paths in the sequential punishment model.

Ignoring the upper constraint, optimal paths will be shown to have a **quasi-flat** structure, in that the trigger level is identical following the second point along the path. This is a surprising result, since optimal paths in the standard infinitely-repeated stage game models treated in the existing literature, such as the Cournot and Bertrand oligopoly models, involve a finite punishment phase followed by a return to full co-operation, where the Pareto efficient outcome in the stage game is restored (Abreu, 1986) (Lambson, 1987).[20] The different result in the sequential punishment model is driven by the presence of altruistic preferences, which cause "neutral observers", who are not being punished (but

[20] We discuss this further in subsection 2.9.5.

who are still affected by the "carrot" created by the remainder of the path) to be more sensitive to variation in the trigger levels than the individual being punished (the first have a more concave intertemporal utility function than the second). Intuitively, with partial altruism ($\theta < 1$), the individual being punished is hurt partly or primarily simply because *they* are being punished, whereas the benefit values for which harm opportunities are taken makes more relative difference to a "neutral observer".

The socially efficient outcome is **supportable** for given values of δ and θ if and only if there exists a punishment path ψ such that the corresponding strategy profile forms a subgame-perfect Nash equilibrium. Checking for supportability involves two conditions. Firstly, ψ must be **sustainable**. This requires that individuals be incentivized to co-operate with the punishment path, either by punishing when they would prefer not to in a single-move game, or by refraining from punishing when they would prefer to. Secondly, given a sustainable path, it must be of sufficient **severity** to incentivize all players to co-operate with the initial path, so that the socially efficient outcome occurs in equilibrium.

Definition 2.3: Let $U_k : \Psi \longrightarrow \mathbb{R}$ be the per-period average discounted expected utility of the individual being punished along path ψ, looking forward from point k. Let $V_k : \Psi \longrightarrow \mathbb{R}$ be the per-period average discounted expected utility of a "neutral observer" who is not being punished along path ψ.

$$U_k(\psi) \equiv \left(\frac{1-\delta}{\delta}\right) \sum_{i=k+1}^{\infty} \left[\delta^{i-k} \int_{\psi_i}^{1} (\theta\pi - 1) g(\pi) d\pi\right] \tag{2.5}$$

$$V_k(\psi) \equiv \left(\frac{1-\delta}{\delta}\right) \sum_{i=k+1}^{\infty} \left[\delta^{i-k} \int_{\psi_i}^{1} (\theta\pi - \theta) g(\pi) d\pi\right] \tag{2.6}$$

Note that, for a flat path, $\bar{\psi}$, these functions simplify to give (where $U : \mathbb{R} \longrightarrow \mathbb{R}$ and $V : \mathbb{R} \longrightarrow \mathbb{R}$):[21]

$$\forall_k : U_k\left(\bar{\psi}\right) \equiv U\left(\bar{\psi}\right) \equiv \int_{\bar{\psi}}^{1} \left(\theta\pi - 1\right) g(\pi) d\pi \qquad \forall_k : V_k\left(\bar{\psi}\right) \equiv V\left(\bar{\psi}\right) \equiv \int_{\bar{\psi}}^{1} \left(\theta\pi - \theta\right) g(\pi) d\pi \tag{2.7}$$

The **supportability constraints** are as follows. $\lambda_k : \Psi \longrightarrow \mathbb{R}$ is the lowest possible net loss of utility from refusing to punish when required to at point k along punishment path ψ (this only "bites" when $\psi_k < \theta$) and $\mu_k : \Psi \longrightarrow \mathbb{R}$ is the lowest possible net loss of utility from punishing when required not to along punishment path ψ (this only "bites" when $\psi_k > \theta$). $\kappa : \Psi \longrightarrow \mathbb{R}$, meanwhile, is the lowest possible net loss in utility from defecting from the initial path, given that path ψ is used to punish such a deviation.[22]

In order for punishment path ψ to support the socially efficient equilibrium, it must be the case that $\forall_k : \lambda_k(\psi) \geq 0$, $\forall_k : \mu_k(\psi) \geq 0$ and that $\kappa(\psi) \geq 0$. The optimal punishment path is the one which

[21] The functions defined in 2.8 through 2.11 are therefore also alternatively functions of a constant trigger level $\bar{\psi}$ along a flat path (a real number). Note also that we can suppress the subscript to indicate the point reached along a flat path, since a flat path looks identical at all points.

[22] These are also functions of δ and θ but we generally suppress this in the notation, for clarity and simplicity.

minimises $U_0(\psi)$ subject to these constraints, which is the same as maximising the severity of the punishment path for the punishee, denoted by $\phi : \Psi \longrightarrow \mathbb{R}$. The **optimal path**, ψ^* is therefore the path that maximizes $\phi(\psi)$ whilst satisfying all the constraints. The **optimal flat path**, $\bar{\psi}^*$ is defined analogously.

$$\lambda_k(\psi) \equiv \left(\frac{\delta}{1-\delta}\right) V_k(\psi) - \left(\frac{\delta}{1-\delta}\right) U_0(\psi) + \psi_k - \theta \tag{2.8}$$

$$\mu_k(\psi) \equiv \left(\frac{\delta}{1-\delta}\right) V_k(\psi) - \left(\frac{\delta}{1-\delta}\right) U_0(\psi) - \psi_k + \theta \tag{2.9}$$

$$\kappa(\psi) \equiv -\left(\frac{\delta}{1-\delta}\right) U_0(\psi) + \theta - 1 \tag{2.10}$$

$$\phi(\psi) \equiv -\left(\frac{\delta}{1-\delta}\right) U_0(\psi) \tag{2.11}$$

2.8 *The Optimal Flat Path*

In this section, we proceed to exhaustively derive the nature of the socially efficient equilibria which can be supported using the optimal flat punishment path. The uniform distribution is used as an illustrative example, but Theorems 2.I, 2.III and 2.IV apply for any continuous benefit distribution. Theorem 2.II only applies to equilibria supported by flat paths. It will be generalized later on.

Intuitively, the impact of the level of altruism on the severity of the optimal flat punishment path, and the resultant supportability of the socially efficient equilibrium, depends upon the interaction of the temptation, severity and willingness effects laid out in section 2.2. The temptation effect means that lower altruism makes a socially efficient outcome harder to support, *ceteris paribus*. The severity effect, by contrast, makes a given punishment path more effective with a lower coefficient of altruism, and so increases the supportability of the socially efficient equilibrium, *ceteris paribus*. The willingness effect makes individuals more willing to punish with lower altruism, thus also rendering punishment paths more effective, and social efficiency easier to support, *ceteris paribus*.

The key result to be established is that, as $\theta \longrightarrow 1^-$ and individuals become perfectly altruistic, the interaction of the three effects leads to a breakdown of the socially efficient equilibrium. The intuition is, firstly, that when $\theta = 1$, the severity of any punishment path will be 0, and the optimal path will not involve any harm being inflicted. This is because perfectly altruistic individuals do not mind harm being focused from other people onto them, and so there is no loss of utility from defecting from the punishment path, and therefore individuals cannot be incentivized to do any punishing at all. The constraint for supportability of the initial path must therefore be just fulfilled with equality at this point (because there is also no temptation to defect).

If the coefficient of altruism is reduced slightly below θ then, since very little punishment can be sustained with such a high coefficient of altruism, the willingness and severity effects must be negligible. Hence the temptation effect must dominate, and social efficiency must be rendered unsupportable. However, as θ is further reduced, provided δ is sufficiently high, the combined willingness and severity effects will eventually become large enough to offset the temptation effect and lead the condition for social efficiency to be supported again. For high enough values of δ, this can continue to occur as $\theta \longrightarrow -\infty$, meaning that social efficiency can be supported even with infinite malevolence. (This phenomenon is driven by the severity effect - see Theorem 2.III.)

There will in general exist an optimal level of altruism θ^*, in the sense that it allows the socially efficient outcome to be supported for the widest range of δ, δ^*. θ^* has the general property that it is low enough for the punishment path to involve inflicting harm for the widest possible range of benefit values, but not any lower, in order to avoid the temptation effect outweighing the severity effect as θ is further reduced. As we shall see later, for a "sufficiently flat" benefit distribution, the optimal punishment path is always flat (see Theorem 2.V). Even when the optimal path is not flat, however, Theorem 2.II provides an important intermediate step in proving the core result of the paper, that too high a level of altruism is socially detrimental, for these more general cases.

In this section, we will derive the properties of the optimal flat punishment path denoted by $\bar{\psi}^*$. Along such a path, the co-operation constraints at every point are identical, and so $\forall_k : \lambda_k(\bar{\psi}) \equiv \lambda(\bar{\psi})$ and $\forall_k : \mu_k(\bar{\psi}) \equiv \mu(\bar{\psi})$. Assuming that $\bar{\psi}^*$ is greater than $\hat{\pi}$ (i.e. that individuals cannot be incentivized to punish for *all* possible values of the benefit), then an expression for the optimal flat trigger level can be found by setting $\lambda(\bar{\psi}^*) = 0$ so that all individuals are being pushed right up against the limit of their willingness to punish given that they themselves are threatened with punishment if they refuse. Substituting in (2.5) and (2.6) into (2.8), equating with 0 and solving for $\bar{\psi}^*$ gives us the following:

$$\bar{\psi}^* = \theta - (1-\theta)\frac{\delta}{1-\delta}\int_{\bar{\psi}^*}^{1} g(\pi)d\pi \tag{2.12}$$

Although this only implicitly defines $\bar{\psi}^*$, and cannot be solved without making specific assumptions about the functional form of $g(\pi)$, it can be totally differentiated and rearranged to yield the following expression for the derivative $\frac{d\bar{\psi}^*}{d\theta}$. The shows the willingness effect - the impact of a change in the coefficient of altruism upon the optimal flat trigger level. We should note at this point that as $\theta \longrightarrow 1^-$, this expression becomes unambiguously positive. This means that the willingness effect negatively affects the severity of punishment as θ increases at an already high level of altruism.

$$\frac{d\bar{\psi}^*}{d\theta} = \frac{(1-\delta) + \delta \int_{\bar{\psi}^*}^{1} g(\pi)d\pi}{(1-\delta) - \delta(1-\theta)g(\bar{\psi}^*)} \tag{2.13}$$

If θ is low enough, individuals will be willing to punish for all possible vales of the benefit. This will imply that $\lambda(\bar{\psi}^*) \geq 0$ when $\bar{\psi}^* = \hat{\pi}$. Theorem 2.I derives the relevant "cut-off" value of θ.

Theorem 2.I

If $\theta \leq \delta + (1-\delta)\hat{\pi}$ then punishment will occur for all benefit values along the optimal path.

If $\theta \leq \delta + (1-\delta)\hat{\pi}$ then $\bar{\psi}^* = \hat{\pi}$, otherwise $\bar{\psi}^* = \theta - (1-\theta)\frac{\delta}{1-\delta}\int_{\bar{\psi}^*}^{1} g(\pi)d\pi$.

Proof: Substituting expressions (2.5) and (2.6) into (2.8) and making this greater than or equal to 0 where $\bar{\psi}^* = \hat{\pi}$ gives us $(1-\theta)\frac{\delta}{1-\delta}\int_{\hat{\pi}}^{1} g(\pi)d\pi + \hat{\pi} - \theta \geq 0$. Since $\int_{\hat{\pi}}^{1} g(\pi)d\pi = 1$, rearranging yields the stated inequality. If $\theta > \delta + (1-\delta)\hat{\pi}$, on the other hand, then $\bar{\psi}^*$ must be where $\lambda(\bar{\psi}^*) = 0$ in an interior solution as described by equation (2.12).

Having derived the optimal flat punishment path, we are now in a position to characterise the socially efficient equilibria which can be supported using it. Substituting in (2.5) into (2.10) gives us the following (along with its total derivative with respect to θ):

$$\kappa\left(\bar{\psi}^*\right) = -\frac{\delta}{1-\delta}\int_{\bar{\psi}^*}^{1} (\theta\pi - 1)g(\pi)d\pi + \theta - 1 \tag{2.14}$$

$$\frac{d\kappa}{d\theta} = 1 - \frac{\delta}{1-\delta}\left(\int_{\bar{\psi}^*}^{1} \pi g(\pi)d\pi + \left(1-\theta\bar{\psi}^*\right) g(\bar{\psi}^*)\frac{d\bar{\psi}^*}{d\theta}\right) \tag{2.15}$$

We can now prove that, for any functional form for $g(\pi)$, there will exist values of θ close to but below 1 for which social efficiency will not be supportable (i.e. for which $\kappa(\bar{\psi}^*) < 0$).[23]

Theorem 2.II

As altruism becomes perfect, the optimal flat punishment path cannot support the socially efficient equilibrium, for any value of the discount factor.

As $\theta \longrightarrow 1^-$, $\bar{\psi}^* \longrightarrow 1$, $\kappa\left(\bar{\psi}^*\right) \longrightarrow 0$ and $\frac{d\kappa}{d\theta} \longrightarrow 1$, therefore as $\theta \longrightarrow 1^-$, $\kappa\left(\bar{\psi}^*\right) \longrightarrow 0^-$.

[23] We will prove this result more generally for any generic optimal punishment path in section 2.10, Theorem 2.VII.

Proof: As $\theta \longrightarrow 1$, it can be seen from expression (2.12) that $\bar{\psi}^* \longrightarrow 1$. The RHS of (2.14) thus goes to 0. Meanwhile, the RHS of (2.15) goes to 1. Since $\kappa(\bar{\psi}^*)$ is a continuously differentiable function of $\bar{\psi}^*$ and θ, it must be the case that $\kappa(\bar{\psi}^*)$ falls below 0 for some values of θ close to but less than 1.

We will now show that, if δ is high enough, then, once $\bar{\psi}^* = \hat{\pi}$, so that punishment is occurring for all possible values of the benefit, the severity effect will dominate. This means that as $\theta \longrightarrow -\infty$, $\kappa(\bar{\psi}^*) \longrightarrow \infty$ and so social efficiency becomes unambiguously supportable. Theorem 2.III derives the required condition on δ.

Theorem 2.III

If $\delta > \frac{1}{1+\bar{\pi}}$ then the socially efficient equilibrium can be supported with infinite malevolence.

If $\delta > \frac{1}{1+\bar{\pi}}$ then $\kappa\left(\bar{\psi}^*\right) \longrightarrow \infty$ as $\theta \longrightarrow -\infty$.

Proof: By Theorem 2.I, when $\theta \leq \delta + (1-\delta)\hat{\pi}$ and so $\bar{\psi}^* = \hat{\pi}$, there is no further willingness effect and so $\frac{d\bar{\psi}^*}{d\theta} = 0$. As $\theta \longrightarrow -\infty$, this must occur. Therefore, as can be seen from (2.14), as $\theta \longrightarrow -\infty$, $\kappa(\bar{\psi}^*) \longrightarrow \infty$ provided that $\frac{\delta}{1-\delta} > \frac{1}{\int_{\hat{\pi}}^{1} \pi g(\pi) d\pi}$. (This is because, as $\theta \longrightarrow -\infty$, any part of (2.14) not containing θ becomes negligible.) It can similarly be seen from (2.15) that these same conditions will ensure that $\frac{d\kappa}{d\theta} < 0$ once $\theta \leq \delta + (1-\delta)\hat{\pi}$. Letting $\bar{\pi} \equiv \int_{\hat{\pi}}^{1} \pi g(\pi) d\pi$ and rearranging yields the stated result.

Theorem 2.III shows that as the coefficient of altruism becomes infinitely negative, the severity effect will dominate if $\delta > \frac{1}{1+\bar{\pi}}$. Since this lower bound for δ is less than 1, there will be a range where too high a level of altruism renders the socially efficient equilibrium unsupportable but, once θ is below the upper limit, no arbitrarily high degree of malevolence will do so. If, however, $\delta^* < \delta < \frac{1}{1+\bar{\pi}}$ then both too high and too low a level of altruism will cause a breakdown of social efficiency.[24]

There are a number of approaches which we could take in defining the socially optimal level of altruism in the sequential punishment model. In a world where we were unable to achieve the first-best solution, we could ask what the impact of a change in the coefficient of altruism is upon the efficiency of the second-best equilibrium. This we do in section 2.11. In this section, we concentrate on worlds where the first-best solution is available, and ask what value of θ allows the socially efficient outcome to be supportable for the widest range of δ. We thus not only consider the best we can do in each

[24] See Theorem 2.IV.

possible world, but begin by considering the broader and more "philosophical" issue of which is the best of all possible worlds to be in.

Theorem 2.IV defines the socially optimal level of altruism, θ^* and corresponding minimum δ, δ^*. It also establishes that $\delta^* \leq \frac{1}{1+\bar{\pi}}$ for any benefit distribution, so that both too high and too low a level of altruism relative to θ^* will cause a break-down of social efficiency, for values of δ close to, but above, δ^*. The optimal coefficient of altruism has a number of key features. Firstly, it must be "knife-edge" socially efficient equilibrium so that $\kappa(\bar{\psi}^*) = 0$. Secondly, it must be the case that δ is just high enough so that punishment can occur for all values of π, in order that punishment paths are maximally severe for the individual being punished.

Theorem 2.IV

The socially optimal level of altruism is always strictly positive and strictly less than 1.

The socially optimal level of altruism is $\theta^* = \frac{3+\bar{\pi}\hat{\pi}-\sqrt{5+2\,\bar{\pi}\hat{\pi}+\bar{\pi}^2\hat{\pi}^2-4\,\bar{\pi}-4\,\hat{\pi}}}{2(1+\bar{\pi})}$, where $0 < \theta^* < 1$.

Proof: The socially optimal level of altruism is where $\kappa(\bar{\psi}^*) = 0$ and $\theta = \delta + (1-\delta)\hat{\pi}$. The following two equations must therefore hold simultaneously: (Equation (2.16) is derived from Theorem 2.I. Equation (2.17) is derived from setting equation (2.14) equal to 0 and plugging in $\bar{\psi}^* = \hat{\pi}$ and $\bar{\pi} \equiv \int_{\hat{\pi}}^{1} \pi g(\pi) d\pi$.)

$$\theta = \delta + (1-\delta)\hat{\pi} \tag{2.16}$$

$$\frac{\delta}{1-\delta} = \frac{1-\theta}{1-\theta\bar{\pi}} \tag{2.17}$$

Equations (2.17) and (2.16) form a quadratic equation system, yielding the following solution:[25]

$$\delta^* = \frac{3-2\,\hat{\pi}-\bar{\pi}\hat{\pi}-\sqrt{5+2\,\bar{\pi}\hat{\pi}+\bar{\pi}^2\hat{\pi}^2-4\,\bar{\pi}-4\,\hat{\pi}}}{2\,(1-\hat{\pi})\,(1+\bar{\pi})} \tag{2.18}$$

$$\theta^* = \frac{3+\bar{\pi}\hat{\pi}-\sqrt{5+2\,\bar{\pi}\hat{\pi}+\bar{\pi}^2\hat{\pi}^2-4\,\bar{\pi}-4\,\hat{\pi}}}{2\,(1+\bar{\pi})} \tag{2.19}$$

To interpret (2.18) and (2.19), we first need to observe that $\sqrt{5+2\,\bar{\pi}\hat{\pi}+\bar{\pi}^2\hat{\pi}^2-4\,\bar{\pi}-4\,\hat{\pi}}$ is decreasing in $\hat{\pi}$ and $\bar{\pi}$ and so its value will lie between $\sqrt{5}$ (when $\hat{\pi} = 0$ and $\bar{\pi} = 0$) and 0 (when $\hat{\pi} = 1$ and $\bar{\pi} = 1$). Therefore, it can immediately be seen that, since $\sqrt{5} < 3$, the socially optimal level of altruism defined by (2.19) will always be positive. Also, since θ^* is increasing in $\hat{\pi}$, its upper limiting value will be 1, as $\hat{\pi} \longrightarrow 1^-$. Since $0 < \theta^* < 1$ and $0 < \bar{\pi} < 1$, it can therefore also be seen from (2.17) that $0 < \delta^* < 1$ (since the LHS must be positive in order to equal the RHS).

[25] Note that the second solution to the quadratic can be discounted since we require that $\theta^* < 1$ in order for (2.16) to be satisfied with $\delta^* < 1$.

In order to be certain that (2.19) defines a point where a further decrease in θ will render the socially efficient initial path unsupportable, we need to show that δ^* from (2.18) lies weakly below $\frac{1}{1+\bar{\pi}}$, derived in Theorem 2.III. Dividing the RHS of (2.18) by $\frac{1}{1+\bar{\pi}}$ yields the following ratio, which we need to show is always weakly less than 1:

$$\frac{3-2\,\hat{\pi}-\bar{\pi}\hat{\pi}-\sqrt{5+2\,\bar{\pi}\hat{\pi}+\bar{\pi}^2\hat{\pi}^2-4\,\bar{\pi}-4\,\hat{\pi}}}{2\,(1-\hat{\pi})} \tag{2.20}$$

Note first that when $\bar{\pi}=1$, (2.20) also equals 1. If the derivative of (2.20) with respect to $\bar{\pi}$ can be shown to be always positive when (2.20) is positive, then this will be sufficient to establish that (2.20) is always less than 1. Differentiating (2.20) with respect to $\bar{\pi}$ gives us:

$$\frac{2-\hat{\pi}-\bar{\pi}\hat{\pi}^2-\hat{\pi}\sqrt{5+2\,\bar{\pi}\hat{\pi}+\bar{\pi}^2\hat{\pi}^2-4\,\bar{\pi}-4\,\hat{\pi}}}{2\sqrt{5+2\,\bar{\pi}\hat{\pi}+\bar{\pi}^2\hat{\pi}^2-4\,\bar{\pi}-4\,\hat{\pi}}\,(1-\hat{\pi})} \tag{2.21}$$

Denoting $\sqrt{5+2\,\bar{\pi}\hat{\pi}+\bar{\pi}^2\hat{\pi}^2-4\,\bar{\pi}-4\,\hat{\pi}}$ by β, (2.20) and(2.21) become respectively:

$$\frac{3-2\,\hat{\pi}-\bar{\pi}\hat{\pi}-\beta}{2\,(1-\hat{\pi})} \tag{2.22}$$

$$\frac{2-\hat{\pi}-\bar{\pi}\hat{\pi}^2-\hat{\pi}\beta}{2\beta\,(1-\hat{\pi})} \tag{2.23}$$

Both (2.22) and (2.23) are decreasing in β. Therefore they can only be positive when β is low enough. Setting (2.22) equal to 0 and solving for β yields the following:

$$\beta = 3-2\,\hat{\pi}-\bar{\pi}\hat{\pi} \tag{2.24}$$

Since (2.20) must always be positive, β must be weakly lower than this.

Plugging (2.24) into (2.23) yields the following expression, which is always positive, showing that β can never be high enough to make (2.23) negative.

$$\frac{1-\hat{\pi}}{3-2\,\hat{\pi}-\bar{\pi}\hat{\pi}} \tag{2.25}$$

We have therefore established that $\delta^* \leq \frac{1}{1+\bar{\pi}}$ for all the relevant values of $\hat{\pi}$ and $\bar{\pi}$.[26] Thus, if θ is further reduced below θ^* with δ unchanged at δ^*, the supportability constraint on the socially efficient initial path will be broken. θ^* is therefore optimal in the sense that a lower level of altruism could, for some δ values, render the socially efficient outcome unsupportable. Also, by Theorem 2.II, a big enough *increase* in θ above θ^* will cause the socially efficient equilibrium to break down.

[26] If $\bar{\pi} < 1$ (as will hold in a standard non-limiting case - see subsection 2.8.1) then this inequality is strictly satisfied, i.e. $\delta^* < \frac{1}{1+\bar{\pi}}$.

2.8.1 Illustration: unitary distribution cases

In this subsection, we produce some graphics to compare the socially optimal level of altruism in a number of key cases. First, we assume that $\hat{\pi} = 0$ and we vary $\bar{\pi}$. (2.18) and (2.19) then become:

$$\theta^* = \delta^* = \frac{3 - \sqrt{5 - 4\bar{\pi}}}{2(1 + \bar{\pi})} \tag{2.26}$$

We compare this to the two extreme **unitary distribution** cases (where we vary $\hat{\pi}$). The first is where $\bar{\pi} \longrightarrow 1$, so that the probability density is infinitely concentrated around the point where the benefit of punishing to the punisher is equal to the cost of punishing to the punishee. The second is where $\bar{\pi} \longrightarrow \hat{\pi}$, so that the probability density is infinitely concentrated around $\hat{\pi}$, and thus the benefit is always less than the cost. By taking the limit of (2.16) and (2.17) as $\bar{\pi} \longrightarrow \hat{\pi}$, we get the following solution for the unitary distribution case where the benefit is always $\hat{\pi}$:

$$\theta^* = \frac{3 + \hat{\pi}^2 - \sqrt{\left(\hat{\pi}^2 + 2\hat{\pi} + 5\right)\left(1 - \hat{\pi}\right)^2}}{2\left(\hat{\pi} + 1\right)} \qquad \delta^* = \frac{3 + \hat{\pi} - \sqrt{\hat{\pi}^2 + 2\hat{\pi} + 5}}{2\left(\hat{\pi} + 1\right)} \tag{2.27}$$

For the case where $\bar{\pi} \longrightarrow 1$ and so the benefit is always 1, we get the following:

$$\theta^* = \frac{\hat{\pi}}{2} + \frac{1}{2} \qquad \delta^* = \frac{1}{2} \tag{2.28}$$

Finally, if we take a "symmetric" distribution where $\bar{\pi} = \frac{\hat{\pi}}{2} + \frac{1}{2}$ then we get the following:

$$\theta^* = \frac{6 + \hat{\pi}^2 + \hat{\pi} - \sqrt{\left(\hat{\pi}^2 + 4\hat{\pi} + 12\right)\left(1 - \hat{\pi}\right)^2}}{2\left(3 + \hat{\pi}\right)} \qquad \delta^* = \frac{6 - \hat{\pi}^2 - 5\hat{\pi} - \sqrt{\left(\hat{\pi}^2 + 4\hat{\pi} + 12\right)\left(1 - \hat{\pi}\right)^2}}{2\left(3 + \hat{\pi}\right)\left(1 - \hat{\pi}\right)} \tag{2.29}$$

Figures 2.2 and 2.3 illustrate the value of the RHS of (2.26) as $\bar{\pi}$ changes, and of (2.27) through (2.29) as $\hat{\pi}$ changes. The two unitary distribution cases represent the two extreme values of θ^* given a particular value of $\hat{\pi}$. The "symmetric" distribution case can be seen to lie somewhere between. A number of observations are worth noting. Firstly, θ^* is increasing in both $\hat{\pi}$ and $\bar{\pi}$. This can be related to the impact of the willingness, severity and temptation effects. When $\hat{\pi}$ is high, the point where optimal paths become maximal, and the willingness effect becomes 0, is reached more quickly as θ is reduced, and so the socially optimal level of altruism is higher. When $\hat{\pi}$ is fixed and $\bar{\pi}$ is increased, on the other hand, then the severity effect is reduced at all values of θ so that, when the willingness effect becomes 0, the value of δ must be higher in order to be at a "knife-edge" point where the social efficient outcome is just supportable. Again, the balance between the severity effect and the temptation effect is effectively tilted in favour of the latter, leading to a higher socially optimal level of altruism.

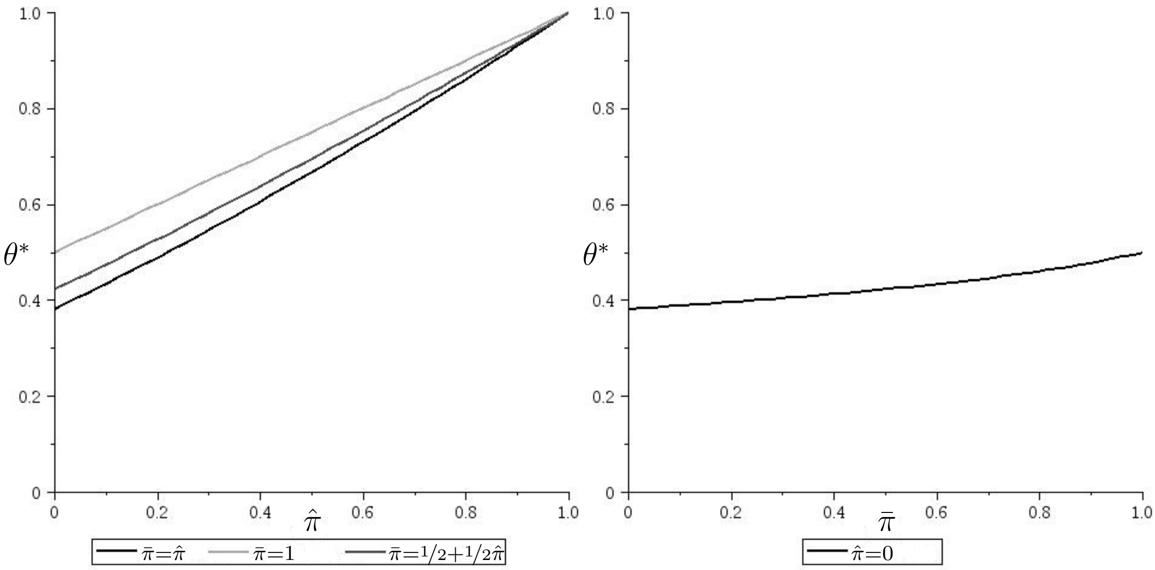

Figure 2.2: Range of θ^* given $\bar{\pi}$ or $\hat{\pi}$

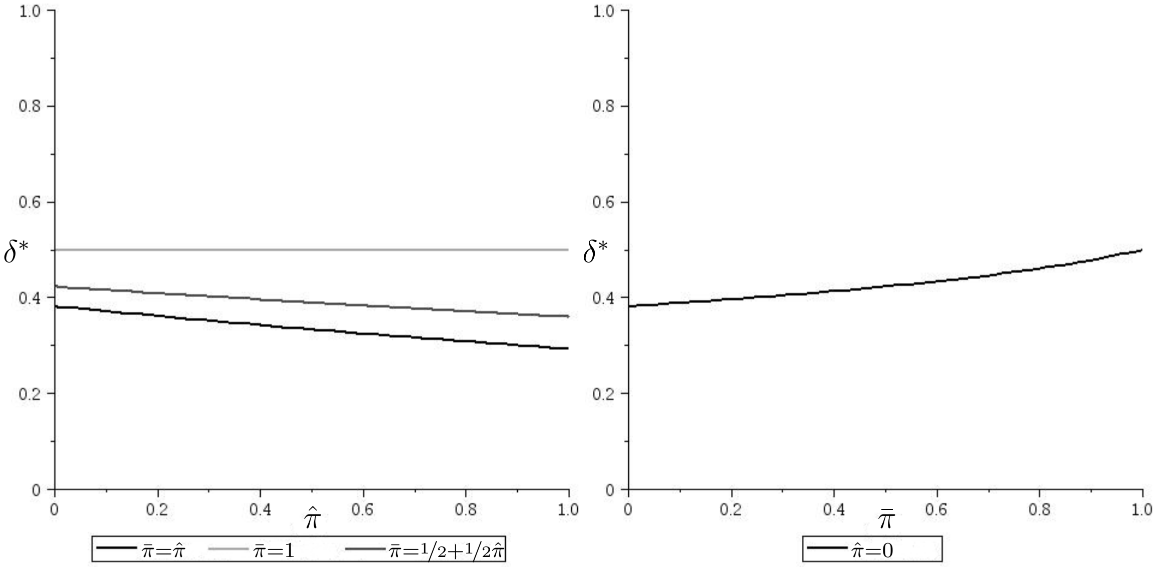

Figure 2.3: Range of δ^* given $\bar{\pi}$ or $\hat{\pi}$

The second key observation to make is that the range of θ^* is quite narrow for any given $\hat{\pi}$. Again, this is due to the interaction of the three effects. This can be most clearly seen by considering the unitary distribution where the benefit is always close to 1. This means that the willingness effect is always 0, because as soon as the flat trigger level $\bar{\psi}^*$ falls below 1, all of the probability mass lies above the trigger level, and so any further reduction in θ and $\bar{\psi}^*$ has no impact. This explains why δ^* is always $\frac{1}{2}$, because this is the value of δ that will lead the temptation and severity effects to exactly cancel out as θ is reduced. It is, however, still necessary that $\bar{\psi}^* = \hat{\pi}$ in order to reach θ^*.[27]

[27] Intuitively, if $\bar{\pi}$ is high then the willingness effect becomes negligible but, since the temptation and severity effects are then roughly equal, it still takes quite a low value of θ before maximal punishment occurs.

2.8.2 Illustration: continuous uniform distribution

Figure 2.4 illustrates the application of Theorems 2.II, 2.III and 2.IV to the case of a uniform distribution with support between 0 and 1 and therefore where $\forall_\pi : g(\pi) = 1$. The key features are the socially optimal level of altruism θ^* and corresponding minimum δ^* where, by substituting in $\bar{\pi} = \frac{1}{2}$, expression (2.26) tells us is at $\theta^* = \delta^* = 1 - \frac{1}{\sqrt{3}}$, and the value of $\delta = \frac{1}{1+\bar{\pi}} = \frac{2}{3}$ above which the severity effect dominates as $\theta \longrightarrow -\infty$, resulting in a horizontal asymptote for the black region where social efficiency is not supportable.

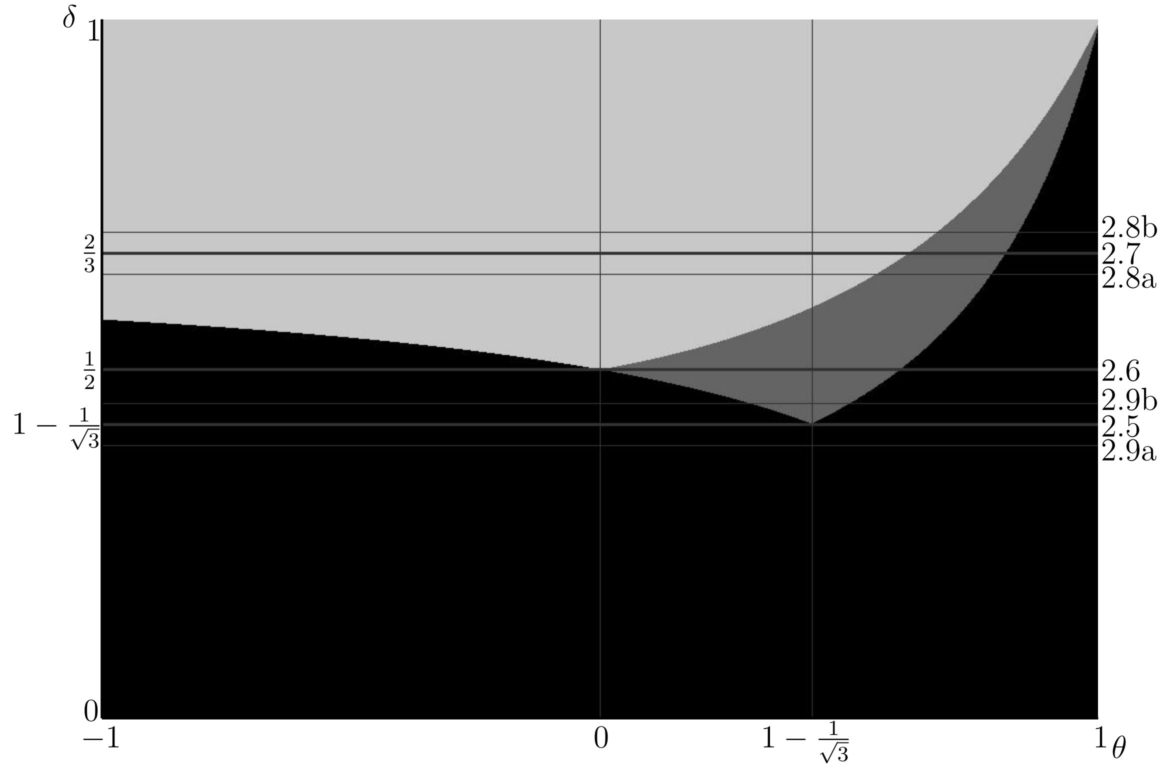

Figure 2.4: Socially efficient equilibria where $\hat{\pi} = 0$ and $g(\pi) = 1$

Figure 2.5 illustrates how the value of $\kappa(\bar{\psi}^*)$ changes (on the y-axis) for a horizontal "cross section" taken through figure 2.4 where $\delta = 1 - \frac{1}{\sqrt{3}}$ and θ is allowed to vary along the x-axis.[28] It can be seen that $\kappa(\bar{\psi}^*)$ lies below 0 for any value of θ apart from $\theta^* = \delta^* = 1 - \frac{1}{\sqrt{3}}$ and $\theta = 1$.[29]

Figure 2.6 shows a "cross-section" at a second key value of $\delta = 0.5$. The significance of this point is that it is where the co-operation constraint for the Nash-reversion punishment path (where $\bar{\psi} = \theta$) and the co-operation constraint for the optimal flat punishment path both bind at the boundary of the black region (which is, for this point only, also on the boundary of the dark grey region).

[28] The labels on the horizontal lines in figure 2.4 indicate the figure which shows the corresponding horizontal cross-section.

[29] When $\theta = 1$ there is no temptation to defect and so social efficiency can always be supported.

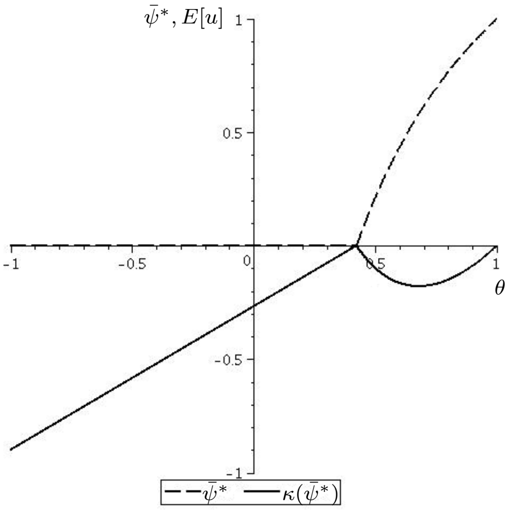

Figure 2.5: Values of $\bar{\psi}^*$ and $\kappa\left(\bar{\psi}^*\right)$ when $\delta = 1 - \frac{1}{\sqrt{3}}$

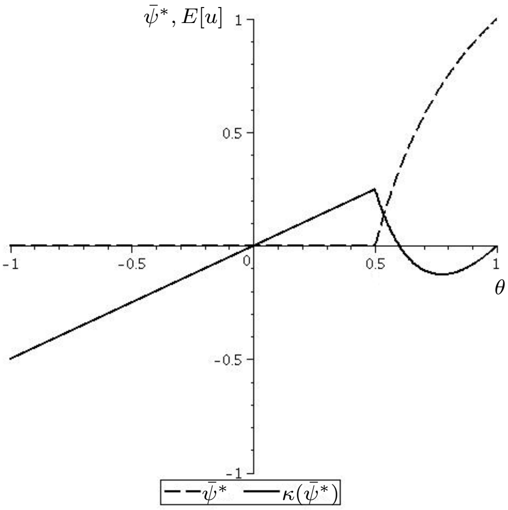

Figure 2.6: Values of $\bar{\psi}^*$ and $\kappa\left(\bar{\psi}^*\right)$ when $\delta = \frac{1}{2}$

A third important value of δ is that above which the temptation effect never outweighs the severity effect as $\delta \longrightarrow -\infty$. In the case of this model, this is where $\frac{\delta}{1-\delta} = 2 \Longrightarrow \delta = \frac{2}{3}$. This cross-section is illustrated graphically in figure 2.7. The property of this particular value of δ that the severity and temptation effects exactly cancel is the fact that the value of $\kappa(\bar{\psi}^*)$ is constant for any $\theta < \delta$. It is instructive to compare this diagram to figure 2.8, which illustrates values of δ slightly above and below $\frac{2}{3}$ respectively. Here we notice that the value of $\kappa(\bar{\psi}^*)$ increases as θ is reduced below δ when $\delta > \frac{2}{3}$, showing that the severity effect outweighs the temptation effect, and the opposite occurs when $\delta < \frac{2}{3}$. Figure 2.9 shows cross-sections for values of δ slightly above and below $1 - \frac{1}{\sqrt{3}}$. The key feature is that when δ is below $1 - \frac{1}{\sqrt{3}}$, the only value of θ for which $\kappa(\bar{\psi}^*)$ is not negative is 1.

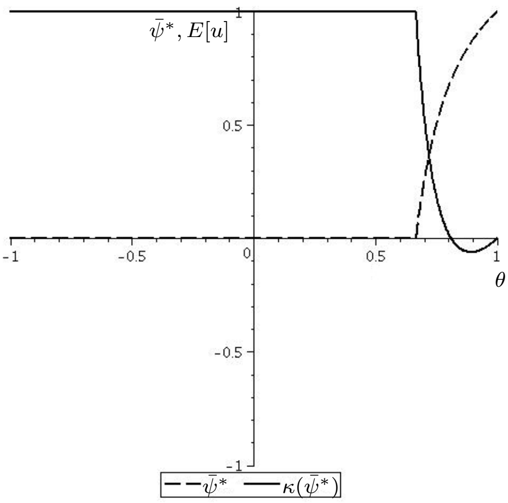

Figure 2.7: Values of $\bar{\psi}^*$ and $\kappa\left(\bar{\psi}^*\right)$ when $\delta = \frac{2}{3}$

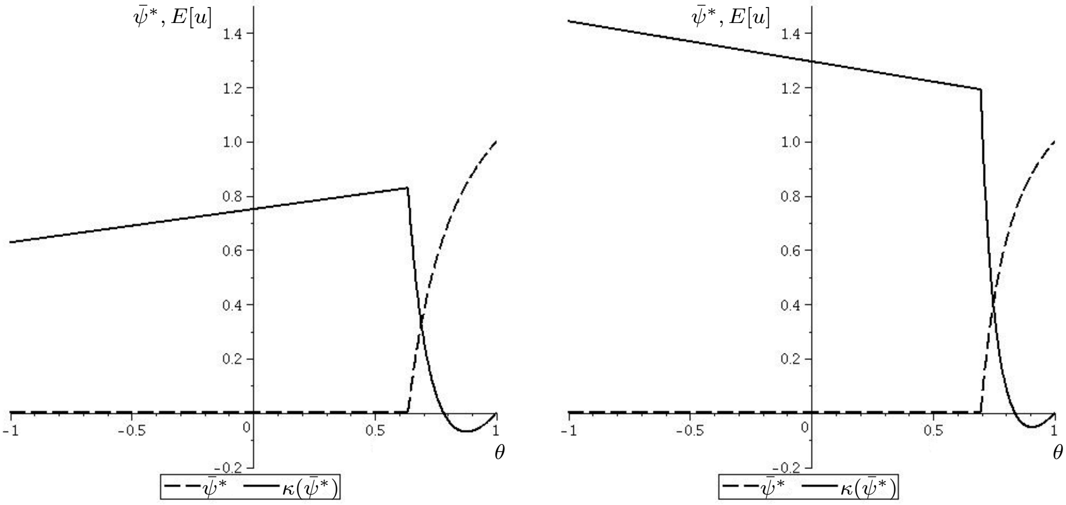

Figure 2.8: Values of $\bar{\psi}^*$ and $\kappa\left(\bar{\psi}^*\right)$ when $\delta = \frac{2}{3} \pm 0.03$

2.9 *The Optimal Quasi-Flat Path*

The next two sections will primarily be concerned with proving the result from Theorem 2.II for the general case where the optimal punishment path is not flat. We begin by showing that, as $\theta \longrightarrow 1^-$, the socially efficient equilibrium becomes unsupportable using the optimal **quasi-flat** punishment path. We then proceed, in section 2.10, to establish that the same result holds even when the optimal *generic* punishment path is used.

Definition 2.4: A **quasi-flat path** is one which is flat from point 2 onwards, and is denoted by $\tilde{\psi}$.[30] The point 1 trigger level is $\tilde{\psi}_1$. The point 2 and after trigger level is $\tilde{\psi}_2$.[31]

[30] The set of quasi-flat paths is denoted $\tilde{\Psi}$.

[31] This is the simplest structure enabling carrot-and-stick punishment, because the individual required to punish at point 1 will take into account the future they face if they co-operate, where the path continues to the less severe "carrot" part, whereas if they defect the path will reset and the "stick" at point 1 will be repeated.

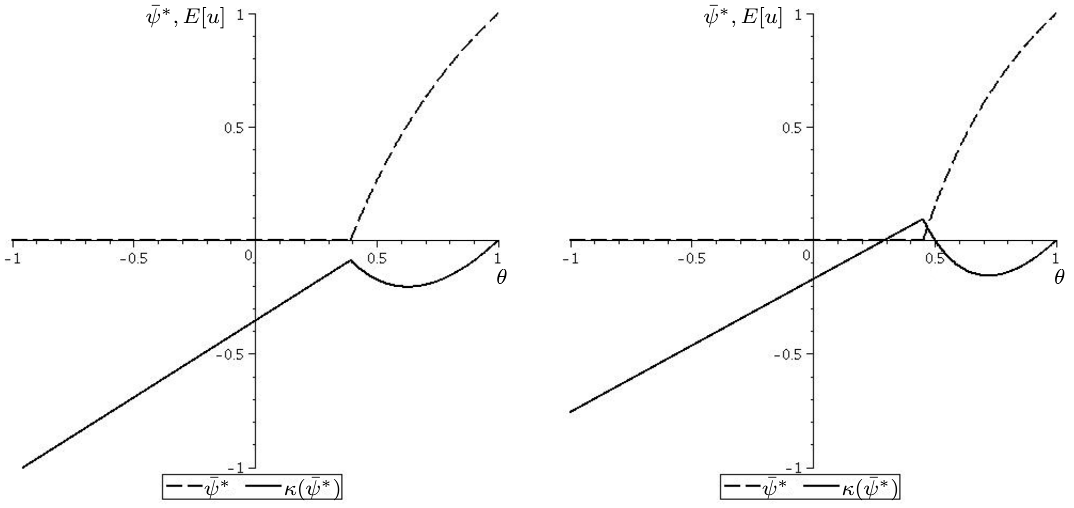

Figure 2.9: Values of $\bar{\psi}^*$ and $\kappa\left(\bar{\psi}^*\right)$ when $\delta = 1 - \frac{1}{\sqrt{3}} \pm 0.03$

An important concept that will be used repeatedly in the lemmas and theorems to follow is the definition of an average trigger level which defines a flat path which is equivalent in terms of per-period average discounted utility to a given non-flat path looking forward from a particular point. This average will in general be different for the individual being punished and for a "neutral observer".

Definition 2.5: Let the **U-average** and the **V-average** be respectively denoted, given Definition 2.3 (equations (2.5) through (2.7)), as $U^{-1}(U_k(\psi))$ and $V^{-1}(V_k(\psi))$. These two averages are given an expanded definition below, and total differentiation is also used to find their derivatives with respect to the trigger level at a particular point $i+k$ (where $i > 0$ since the average is "forward looking"), and the implicit derivative of $V^{-1}(V_k(\psi))$ with respect to $U^{-1}(U_k(\psi))$.[32]

$$U_k\left(\psi\right) \equiv \int_{U^{-1}(U_k(\psi))}^{1} (\theta\pi - 1)g(\pi)d\pi \equiv \left(\frac{1-\delta}{\delta}\right)\left(\sum_{i=1}^{\infty}\left[\delta^i\left(\int_{\psi_{i+k}}^{1}(\theta\pi - 1)g(\pi)d\pi\right)\right]\right) \tag{2.30}$$

$$V_k\left(\psi\right) \equiv \int_{V^{-1}(V_k(\psi))}^{1} (\theta\pi - \theta)g(\pi)d\pi \equiv \left(\frac{1-\delta}{\delta}\right)\left(\sum_{i=1}^{\infty}\left[\delta^i\left(\int_{\psi_{i+k}}^{1}(\theta\pi - \theta)g(\pi)d\pi\right)\right]\right) \tag{2.31}$$

$$\frac{d}{d\psi_{i+k}}U^{-1}(U_k(\psi)) = \left(\frac{1-\delta}{\delta}\right)\delta^i\left(\frac{g\left(\psi_{i+k}\right)}{g\left(U^{-1}(U_k(\psi))\right)}\right)\left(\frac{1-\theta\psi_{i+k}}{1-\theta U^{-1}(U_k(\psi))}\right)$$

$$\frac{d}{d\psi_{i+k}}V^{-1}(V_k(\psi)) = \left(\frac{1-\delta}{\delta}\right)\delta^i\left(\frac{g\left(\psi_{i+k}\right)}{g\left(V^{-1}(V_k(\psi))\right)}\right)\left(\frac{1-\psi_{i+k}}{1-V^{-1}(V_k(\psi))}\right)$$

$$\frac{\frac{d}{d\psi_{i+k}}V^{-1}(V_k(\psi))}{\frac{d}{d\psi_{i+k}}U^{-1}(U_k(\psi))} = \left(\frac{1-\psi_{i+k}}{1-\theta\,\psi_{i+k}}\right)\left(\frac{g\left(U^{-1}(U_k(\psi))\right)}{g\left(V^{-1}(V_k(\psi))\right)}\right)\left(\frac{1-\theta\,U^{-1}(U_k(\psi))}{1-V^{-1}(V_k(\psi))}\right) \tag{2.32}$$

[32] Note that (2.30) and (2.31) are identities rather than just equations, so that both sides can be differentiated with the identity continuing to hold.

The following two lemmas will prove very useful in this and subsequent sections. Importantly, Lemma 2.V.i applies to all optimal punishment paths, not just quasi-flat ones. It states that U_k must we weakly minimised at point 0 along an optimal path. Lemma 2.V.ii states that constraint (2.8) must bind at point 1 along an optimal quasi-flat path.[33]

Lemma 2.V.i

The U-average must be weakly minimized at the beginning of an optimal punishment path.

(a) If a punishment path ψ^* is optimal then $\forall_k : U_k\left(\psi^*\right) \geq U_0\left(\psi^*\right)$.

(b) If a punishment path ψ^* is optimal then $\psi_1^* \leq U^{-1}\left(U_0\left(\psi^*\right)\right) \leq U^{-1}\left(U_1\left(\psi^*\right)\right)$.

Proof: For the first claim, note that it would be possible to construct a new path ψ' identical to ψ^* except beginning at point k so that $\forall_i : \psi_i' = \psi_{k+i}^*$, resulting in the following sustainability constraints:

$$\lambda_i(\psi') \equiv \left(\frac{\delta}{1-\delta}\right) V_i(\psi') - \left(\frac{\delta}{1-\delta}\right) U_0(\psi') + \psi_i' - \theta$$

$$\mu_i(\psi') \equiv \left(\frac{\delta}{1-\delta}\right) V_i(\psi') - \left(\frac{\delta}{1-\delta}\right) U_0(\psi') - \psi_i' + \theta$$

These can be rewritten as:

$$\lambda_i(\psi') \equiv \left(\frac{\delta}{1-\delta}\right) V_{k+i}(\psi^*) - \left(\frac{\delta}{1-\delta}\right) U_k(\psi^*) + \psi_{k+i}^* - \theta$$

$$\mu_i(\psi') \equiv \left(\frac{\delta}{1-\delta}\right) V_{k+i}(\psi^*) - \left(\frac{\delta}{1-\delta}\right) U_k(\psi^*) - \psi_{k+i}^* + \theta$$

Now, since ψ^* must, by assumption, be sustainable, we know that, for any k and i:

$$\lambda_{k+i}(\psi^*) \equiv \left(\frac{\delta}{1-\delta}\right) V_{k+i}(\psi^*) - \left(\frac{\delta}{1-\delta}\right) U_0(\psi^*) + \psi_{k+i}^* - \theta \geq 0$$

$$\mu_{k+i}(\psi^*) \equiv \left(\frac{\delta}{1-\delta}\right) V_{k+i}(\psi^*) - \left(\frac{\delta}{1-\delta}\right) U_0(\psi^*) - \psi_{k+i}^* + \theta \geq 0$$

If we now suppose that there exists a k such that $U_k(\psi^*) < U_0(\psi^*)$, this would mean, by observation, that the supportability constraints for ψ' would unambiguously be fulfilled at every point. Also, this would mean that $\phi(\psi') > \phi(\psi^*)$. Therefore ψ' would be sustainable, and would be more severe than ψ^*. Hence ψ^* could not be optimal - a contradiction.

For the second claim, note that the following identity holds for any path ψ:

$$\frac{\delta}{1-\delta}\left(\int_{U^{-1}(U_0(\psi))}^{1} (\theta\pi - 1)g(\pi)d\pi\right) \equiv \delta \int_{\psi_1}^{1} (\theta\pi - 1)g(\pi)d\pi + \frac{\delta^2}{1-\delta} \int_{U^{-1}(U_1(\psi))}^{1} (\theta\pi - 1)g(\pi)d\pi \quad (2.33)$$

[33] We only need, given our overall strategy, to prove this result for optimal quasi-flat paths but, intuitively, it will also hold for all optimal paths.

This can be rewritten as:

$$\left(\frac{\delta}{1-\delta}\right) U_0\left(\psi\right) \equiv \left(\frac{\delta}{1-\delta}\right) U_1\left(\psi\right) + \delta \int_{\psi_1}^{U^{-1}(U_1(\psi))} (\theta\pi - 1)g(\pi)d\pi$$

Since we know from the argument made above that $U_0\left(\psi^*\right) \leq U_1\left(\psi^*\right)$, we also know that $\int_{\psi_1^*}^{U^{-1}(U_1(\psi^*))}(1-\theta\pi)g(\pi)d\pi \geq 0$, and therefore that $\psi_1^* \leq U^{-1}(U_1(\psi^*))$.

Finally, identity (2.33) can also be rewritten as:

$$0 \equiv \delta \int_{\psi_1}^{U^{-1}(U_0(\psi))} (\theta\pi - 1)g(\pi)d\pi + \frac{\delta^2}{1-\delta} \int_{U^{-1}(U_1(\psi))}^{U^{-1}(U_0(\psi))} (\theta\pi - 1)g(\pi)d\pi$$

Since $U^{-1}(U_0(\psi^*)) \leq U^{-1}(U_1(\psi^*))$, in order for this to hold it must follow that $U^{-1}(U_0(\psi^*)) \geq \psi_1^*$. Intuitively, the U-average must be "dragged down" from below by the trigger level at point 1.

Lemma 2.V.ii

The "lower" constraint must bind at the beginning of an optimal quasi-flat punishment path.

If a quasi-flat punishment path $\tilde{\psi}^*$ is optimal then $\lambda_1\left(\tilde{\psi}^*\right) = 0$.

Proof: Firstly, note that differentiating $\lambda_k(\tilde{\psi})$ or $\mu_k(\tilde{\psi})$ where $k > 1$ with respect to $\tilde{\psi}_1$ yields $\frac{d\lambda_k}{d\tilde{\psi}_1} = \frac{d\mu_k}{d\tilde{\psi}_1} = -\delta(1-\theta\tilde{\psi}_1)g(\tilde{\psi}_1)$. Since this is unambiguously negative, decreasing $\tilde{\psi}_1$ *improves* the incentive to co-operate at all later points along the punishment path. Since $\frac{d\phi}{d\tilde{\psi}_1} = -\delta(1-\theta\tilde{\psi}_1)g(\tilde{\psi}_1)$ is also negative, decreasing $\tilde{\psi}_1$ also makes the punishment path more severe. Therefore, to have reached an optimal path, $\tilde{\psi}_1$ should have been reduced until $\lambda_1(\tilde{\psi}^*) = 0$. It may, however, be the case that $\tilde{\psi}_1$ reaches $\hat{\pi}$ before the period 1 co-operation constraint binds. Since $\tilde{\psi}_2 \equiv U^{-1}(U_1(\tilde{\psi}))$, Lemma 2.V.i implies that $\tilde{\psi}_2^* \geq \tilde{\psi}_1^*$. This in turn means that, for a quasi-flat path $\tilde{\psi}^*$, $\lambda_1(\tilde{\psi}^*) \geq 0$ is a sufficient condition for $\lambda_k(\tilde{\psi}^*) \geq 0$ to hold for all k. Therefore, if $\lambda_1(\tilde{\psi}^*) > 0$ and $\tilde{\psi}_1^* = \hat{\pi}$, then (since $\frac{d\phi}{d\tilde{\psi}_2} = -\frac{\delta^2}{1-\delta}(1-\theta\tilde{\psi}_2)g(\tilde{\psi}_2)$ is negative) $\tilde{\psi}_2^*$ could instead be reduced, resulting in an improvement to the severity of the path - a contradiction.

Lemma 2.V.i implies that the optimal quasi-flat path $\tilde{\psi}^*$ must have a weakly lower trigger level at point 1 ($\tilde{\psi}_1^*$) than at point 2 and after ($\tilde{\psi}_2^*$), so that $\tilde{\psi}_1^* \leq \tilde{\psi}_2^*$. This fits the intuition that the "stick" of harsher punishment should come earlier in the punishment path so that the "carrot" of less harsh punishment later along the path will operate as an incentive for those doing the punishing to co-operate with the harsher punishment earlier on. Note also that, for any quasi-flat path $\tilde{\psi}$, $\forall_{k\geq 1} : \tilde{\psi}_2 \equiv V^{-1}(V_k(\tilde{\psi})) \equiv U^{-1}(U_k(\tilde{\psi}))$.

The trade-off which drives the optimal quasi-flat path is between a “bigger stick” at point one and the resulting “nicer carrot” at point two and after.[34] The gain in severity of the punishment path from a reduction in the point one trigger level depends upon the probability density at that benefit value, whilst the effectiveness of the carrot in offsetting this to ensure sustainability also depends on the probability density at the point two and after trigger level. The *cost* of incentivizing co-operation with a lower trigger level at point one, however, does *not* depend upon the probability density at the point one trigger level, since it is “paid” in full if the value of the benefit turns out to be in the relevant range. Intuitively, therefore, if the probability density function for the benefit is sufficiently flat, this cost will always outweigh the benefit of making the quasi-flat punishment path non-flat.[35]

2.9.1 A taxonomy of optimal quasi-flat paths

An important lesson to draw from the discussion so far is that the framework of strategy profiles constructed from punishment paths does not, in and of itself, provide enough structure to allow a complete and comprehensive solution to the problem of finding the form of optimal punishment in a specific context such as that of the sequential punishment model. Although punishment paths share a similar carrot-and-stick structure, each particular model requires its own toolkit of “tricks” to derive the precise shape of the optimal paths.

The concept of a quasi-flat punishment path turns out to be essential to analysing the equilibria supportable by optimal generic punishment paths in the sequential punishment model. This is because, as will be shown in section 2.10, it is always possible to construct a quasi-flat path whose severity value ϕ (see equation (2.11) in Definition 2.3) forms an upper bound for all sustainable generic paths, and to derive sufficient structure from this to extend Theorem 2.II to all the necessary general cases. Also, quasi-flat paths themselves come in a variety of “flavours”, the differences between them driven by the optimal structure of carrot-and-stick punishment in the sequential punishment model, and its interaction with the partially altruistic preferences of the players, along with the limits on the support for the distribution of the benefit π.

[34] See sections 2.9.5 and 2.9.6 for more detailed examples of this principle in action.

[35] Theorem 2.V will establish this sufficient condition on the probability density function $g(\pi)$ to ensure that the optimal quasi-flat punishment path will be flat.

It is useful to ask at this point why quasi-flat paths emerge naturally from the structure of the sequential punishment model whilst, as we shall discuss in more detail in section 2.9.5, optimal punishment in the infinitely-repeated Bertrand and Cournot games involves complete "front-loading", with as much punishment as possible packed into the early stages of the path (Abreu, 1986) (Lambson, 1987). The answer is that it is the presence of partial altruism which drives the "flattening-out" of the tail of paths in the sequential punishment model.

The positioning of trigger levels at point 2 and onwards involves a trade-off, allowing greater punishment to be "bought" at point 1, but at the cost of less severe punishment later. The "sacrifice ratio" will be given by expression (2.32), which measures the increase in "carrot" (measured as a higher V-average) for a given reduction in "stick" along the path (a higher U-average), brought about by a rise in a later trigger level, ψ_{k+i}. It can be seen that this ratio is more favourable when ψ_{k+i} is lower.[36] It is therefore optimal to "spread out" the punishment evenly over the entire tail, in order to provide the maximum incentive to co-operate for those doing the punishing.

There are a number of possibilities for the precise form that the optimal quasi-flat path might take. By Lemma 2.V.i, it is impossible for the trigger level at point 1 to be higher than at point 2 onwards. This leaves seven possibilities, types A-G. Type A is the **maximal** path characterised in Theorem 2.I. Possibility B is a **quasi-maximal path**, where the trigger level is "maxed-out" at point 1 but not from point 2 onwards. Type C is a **flat** path. With type D paths, on the other hand, the amount of punishment at point 1 runs up against constraint (2.9) in that we cannot further increase $\tilde{\psi}_2$ without rendering the path unsustainable. We shall refer to this case as a **carrot-constrained** path. A fifth possibility, E, is that there is an optimal marginal trade-off between punishment in point 1 and punishment at point 2 and after. We shall call this a **carrot-maximized** path. The next type, F, is a path where the amount of punishment at point 1 runs up against the constraint imposed by not being able to make the future "carrot" attractive enough to allow more severe punishment.[37] We shall call this a **quasi-minimal** path. The seventh and final case, G, is one where no punishment at all can be incentivized; this is a **minimal** path. It can immediately be seen, however that type G paths cannot possibly be optimal, since Theorem 2.I implies that there will always exist an optimal flat path with $\bar{\psi}^* < 1$, which is strictly more severe than the minimal path (where $\bar{\psi}^* = 1$).

[36] The benefits and costs are discounted, so the exchange ratio, given a particular trigger level, looks the same for all future periods.

[37] This happens because we reach the top of the support of the distribution for π (i.e. $\tilde{\psi}_2$ reaches 1).

We shall also find it essential in the lemmas and theorems to follow to distinguish between two different types of optimal path. **Fully-constrained paths** must satisfy all co-operation conditions defined by (2.8) and (2.9). **Semi-constrained paths** only need satisfy the conditions defined in (2.8). A fully-constrained optimal quasi-flat path satisfies co-operation conditions $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, $\mu_1 \geq 0$ and $\mu_2 \geq 0$. Given quasi-flatness, these form a sufficient condition for all the co-operation constraints to be fulfilled. A semi-constrained optimal quasi-flat path is one which is only constrained to satisfy $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ (though it may, by "chance", also satisfy $\mu_1 \geq 0$ and $\mu_2 \geq 0$, and therefore also be in the set of fully-constrained paths).

Definition 2.6: Let Ψ denote the set of unconstrained paths, in increasing order of ϕ. Let Ψ_{sc} and Ψ_{fc} be the sets of sustainable semi-constrained and fully-constrained paths respectively, so that $\Psi_{fc} \subset \Psi_{sc} \subset \Psi$. These are similarly ordered by ϕ. The optimal generic path can be defined as $\psi^* = \sup \Psi_{fc} = \max_{\psi \in \Psi_{fc}} \{\phi(\psi)\}$. Let $\tilde{\Psi} \subset \Psi$ be the set of quasi-flat paths and $\bar{\Psi} \subset \tilde{\Psi}$ be the set of flat paths. Let $\tilde{\Psi}_{sc} = \tilde{\Psi} \cap \Psi_{sc}$, $\bar{\Psi}_{sc} = \bar{\Psi} \cap \Psi_{sc}$, $\tilde{\Psi}_{fc} = \tilde{\Psi} \cap \Psi_{fc}$ and $\bar{\Psi}_{fc} = \bar{\Psi} \cap \Psi_{fc}$ denote analogous sets of semi-constrained and fully-constrained quasi-flat and flat paths. The optimal flat and quasi-flat paths are respectively defined as $\bar{\psi}^* = \sup \bar{\Psi}_{fc}$ and $\tilde{\psi}^* = \sup \tilde{\Psi}_{fc}$. Let $\bar{\psi}^*_{sc} = \sup \bar{\Psi}_{sc}$ denote the optimal semi-constrained flat path and $\tilde{\psi}^*_{sc} = \sup \tilde{\Psi}_{sc}$ denote the optimal semi-constrained quasi-flat path.

It should be noted at this point that the optimal semi-constrained quasi-flat path cannot be "carrot-constrained", since constraint (2.9) does not apply. Also, observe that the optimal semi-constrained path will be at least as effective as the optimal fully-constrained path. In other words, if all paths are ordered in ϕ, the optimal semi-constrained path will equal or beat the optimal fully-constrained path: $\phi(\sup \Psi_{sc}) \geq \phi\left(\sup \Psi_{fc}\right)$. Analogously, $\phi(\sup \tilde{\Psi}_{sc}) \geq \phi(\sup \tilde{\Psi}_{fc})$ and $\phi(\sup \bar{\Psi}_{sc}) \geq \phi(\sup \bar{\Psi}_{fc})$. This observation is key in enabling Theorem 2.II to be generalized to the case where the optimal generic path is used to punish deviations from the socially efficient initial path.

Define ψ_k^{λ} and ψ_k^{μ} as the values of ψ_k that respectively satisfy conditions (2.8) and (2.9) with equality.[38] These therefore represent the upper and lower limits for the trigger level that *could* be sustained at point k given the structure of the entire punishment path. Figures 2.10 through 2.15 illustrate the various possible structural types of optimal quasi-flat path.[39]

[38] Note that for a quasi-flat path this will be the same for any k because $\forall_{k \geq 1} : V_k(\tilde{\psi}) \equiv V_1(\tilde{\psi})$.

[39] We do not show case G, that of a minimal path, since we have already argued that this type of quasi-flat path cannot possibly be optimal.

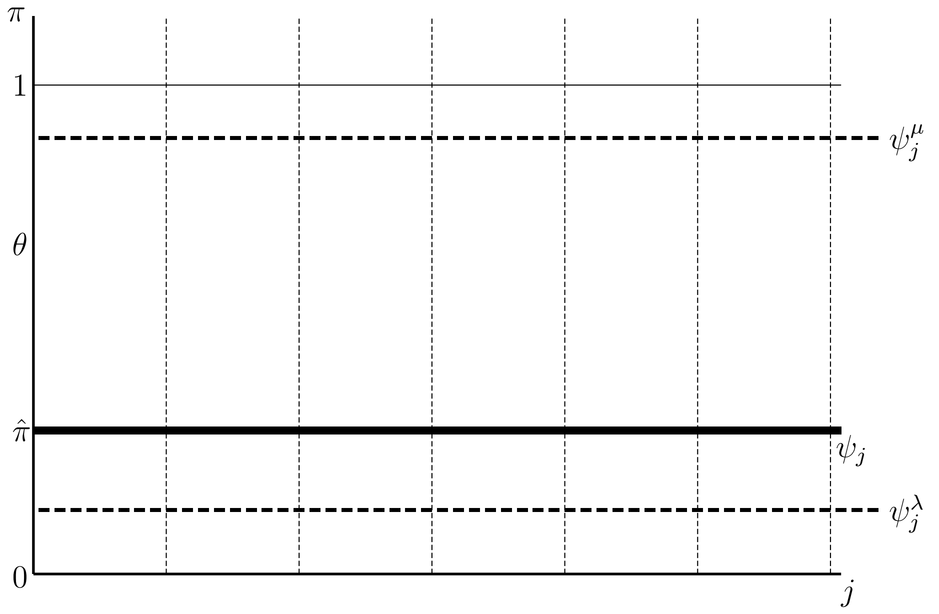

Figure 2.10: Type A: A maximal path

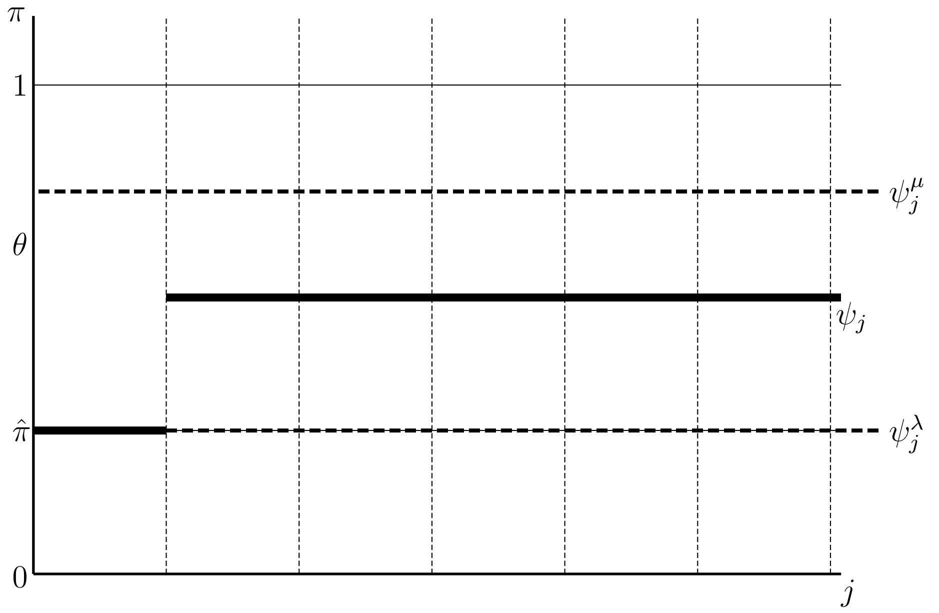

Figure 2.11: Type B: A quasi-maximal path

Theorems 2.I, 2.III and 2.IV continue to hold for the equilibria supportable by generic optimal paths (and, therefore, quasi-flat paths) without alteration. In general, the quasi-maximal case only occurs when θ is close to the boundary established in Theorem 2.I below which the optimal path is maximal. In section 2.9.5, we see an example of this with the triangular distribution. As we shall argue

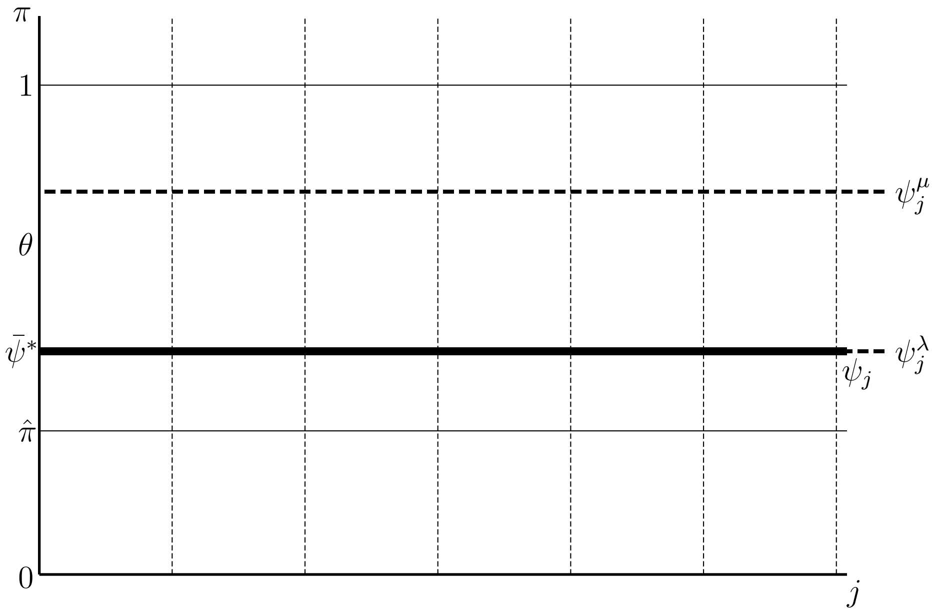

Figure 2.12: Type C: A flat path

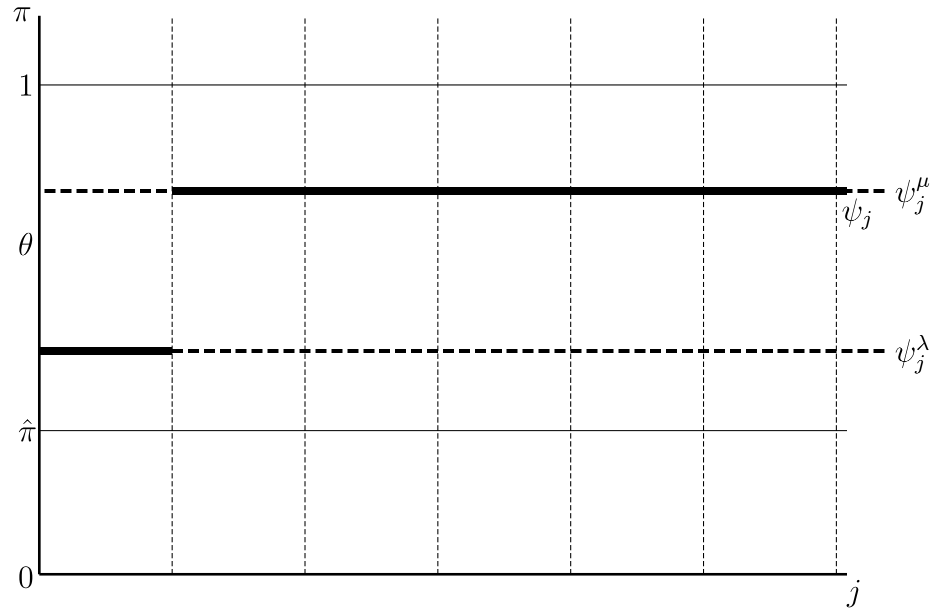

Figure 2.13: Type D: A carrot-constrained path

shortly, the quasi-minimal case cannot possibly be optimal.[40] The flat, carrot-constrained and carrot-maximized paths illustrated in figures 2.12, 2.13 and 2.14 respectively represent the three possibilities for an "interior solution".

[40] The proof of Lemma 2.IX.i (the appendix) verifies this.

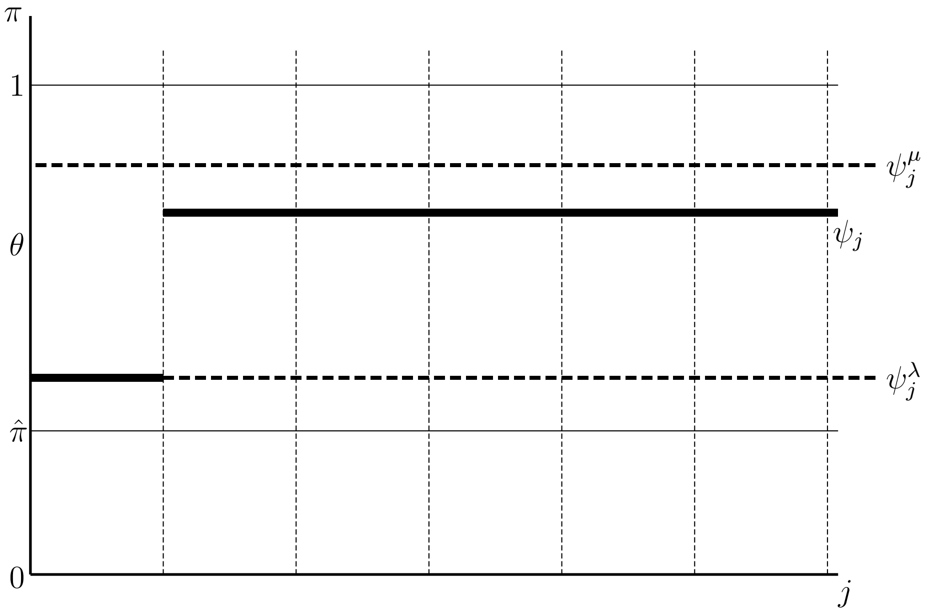

Figure 2.14: Type E: A carrot-maximized path

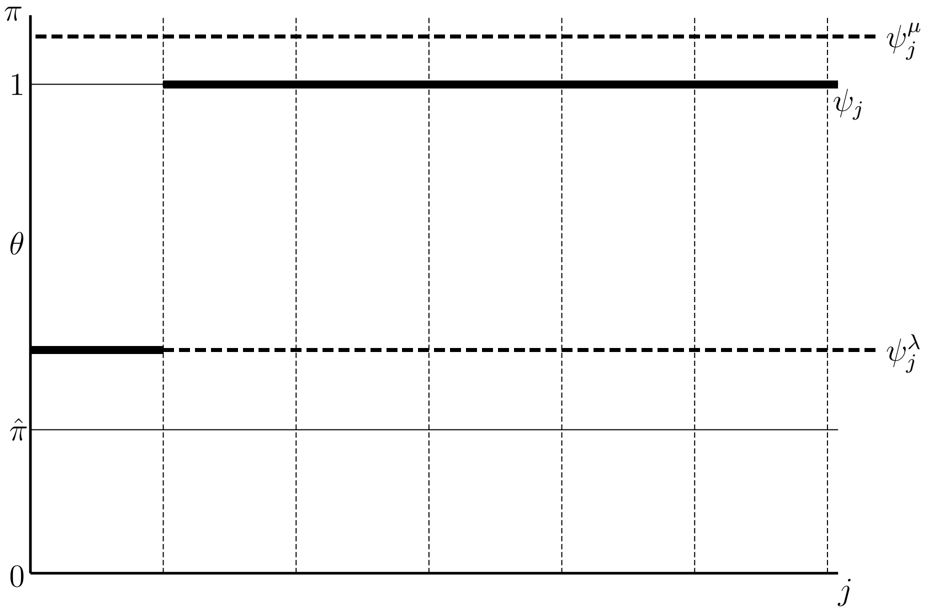

Figure 2.15: Type F: A quasi-minimal path

2.9.2 Conditions for a flat path

To intuitively derive necessary and sufficient conditions for the optimal quasi-flat path to be flat (type C), we can use a trick from Abreu by considering the optimal path we are able to construct using a fixed punishment for a deviation. If this can be shown to be flat, then the optimal path constructed using itself as a punishment will also be flat. Let $\underline{U} \equiv (\frac{\delta}{1-\delta})U_0\left(\underline{\psi}\right)$ be the expected utility for the punishee along the fixed punishment path $\underline{\psi}$. Let $\underline{\lambda}_k(\tilde{\psi}) \equiv (\frac{\delta}{1-\delta})V_k(\tilde{\psi}) - \underline{U} + \tilde{\psi}_k - \theta$ be the co-operation constraint at point k for quasi-flat path $\tilde{\psi}$ given the use of the fixed path $\underline{\psi}$ to punish any deviation. Since the trigger level at point 1 should be set so that $\underline{\lambda}_1(\tilde{\psi}) = 0$ for an optimal path, we know that the following condition must hold:

$$\begin{aligned} \tilde{\psi}_1 =& \theta - \left(\frac{\delta}{1-\delta}\right) V_1(\tilde{\psi}) + \underline{U} \\ =& \theta + \left(\frac{\delta}{1-\delta}\right) \int_{\tilde{\psi}_2}^{1} (\theta - \theta\pi) g(\pi) d\pi + \underline{U} \end{aligned} \tag{2.34}$$

We are seeking to maximize the disutility of the person being punished along the quasi-flat punishment path $\tilde{\psi}$. This will be given by:

$$\begin{aligned} \phi =& - \left(\frac{\delta}{1-\delta}\right) U_0\left(\tilde{\psi}\right) \\ =& \delta \int_{\tilde{\psi}_1}^{1} \left(1 - \theta\pi\right) g(\pi) d\pi + \left(\frac{\delta^2}{1-\delta}\right) \int_{\tilde{\psi}_2}^{1} \left(1 - \theta\pi\right) g(\pi) d\pi \end{aligned} \tag{2.35}$$

Totally differentiating (2.34) with respect to $\tilde{\psi}_1$ gives us:

$$\frac{d\tilde{\psi}_1}{d\tilde{\psi}_2} = - \left(\frac{\delta}{1-\delta}\right) \left(\theta - \theta\tilde{\psi}_2\right) g\left(\tilde{\psi}_2\right) \tag{2.36}$$

Totally differentiating (2.35) with respect to $\tilde{\psi}_2$ gives us:

$$\frac{d\phi}{d\tilde{\psi}_2} = \delta\left(\theta\tilde{\psi}_1 - 1\right) g\left(\tilde{\psi}_1\right) \frac{d\tilde{\psi}_1}{d\tilde{\psi}_2} - \left(\frac{\delta^2}{1-\delta}\right) \left(1 - \theta\tilde{\psi}_2\right) g\left(\tilde{\psi}_2\right)$$

Substituting in (2.36) and simplifying yields:

$$\frac{d\phi}{d\tilde{\psi}_2} = \left(\frac{\delta^2}{1-\delta}\right) g\left(\tilde{\psi}_2\right) \left(\left(1 - \theta\tilde{\psi}_1\right) \left(\theta - \theta\tilde{\psi}_2\right) g\left(\tilde{\psi}_1\right) - \left(1 - \theta\tilde{\psi}_2\right)\right)$$

This is unambiguously negative if the following condition holds:

$$g\left(\tilde{\psi}_1\right) < \frac{1 - \theta\tilde{\psi}_2}{\left(1 - \theta\tilde{\psi}_1\right) \theta \left(1 - \tilde{\psi}_2\right)} \tag{2.37}$$

The RHS of (2.37) is increasing in $\tilde{\psi}_1$ and $\tilde{\psi}_2$. This means that the most stringent condition will be where $\tilde{\psi}_1 = \tilde{\psi}_2 = \hat{\pi}$. Requiring that the probability density function $g(\pi)$ always be less than this ensures that the above condition will always hold. This yields the following condition:

$$\forall_\pi : g(\pi) < \frac{1}{\theta(1-\hat{\pi})} \tag{2.38}$$

Provided this condition holds, increasing $\tilde{\psi}_2$ in order to further reduce $\tilde{\psi}_1$ always makes the punishment path less effective by reducing ϕ. It is therefore optimal to set $\tilde{\psi}_2 = \tilde{\psi}_1$ since to set $\tilde{\psi}_2 < \tilde{\psi}_1$ will result in a clearly non-optimal path, by Lemma 2.V.i. Condition (2.38) is therefore sufficient for the optimal quasi-flat path to be flat. (This result is verified in Theorem 2.V below.)

A *necessary* condition for the optimal quasi-flat path to be flat can be found by substituting $\tilde{\psi}_1 = \tilde{\psi}_2 = \bar{\psi}^*$ into condition (2.37) to give the following:

$$g\left(\bar{\psi}^*\right) < \frac{1}{\theta\left(1-\bar{\psi}^*\right)} \tag{2.39}$$

Theorem 2.V

If the benefit distribution is sufficiently flat, then the optimal quasi-flat punishment path is flat.

If $\forall_\pi : g(\pi) < \frac{1}{\theta(1-\hat{\pi})}$, then $\phi\left(\bar{\psi}^*\right) \geq \phi\left(\tilde{\psi}^*_{sc}\right)$ and therefore $\phi\left(\bar{\psi}^*\right) \geq \phi\left(\tilde{\psi}^*\right)$.

Proof: Lemma 2.IX.i (the appendix) proves that if $\forall_\pi : g(\pi) < \frac{1}{\theta(1-\hat{\pi})}$ then the optimal semi-constrained quasi-flat path must be flat ($\sup(\tilde{\Psi}_{sc}) \in \bar{\Psi}_{sc}$). This implies that $\phi(\bar{\psi}^*_{sc}) \geq \phi(\tilde{\psi}^*_{sc})$. If we can show that $\sup \bar{\Psi}_{sc} \in \bar{\Psi}_{fc}$, then this will mean that $\phi(\bar{\psi}^*) \geq \phi(\bar{\psi}^*_{sc}) \geq \phi(\tilde{\psi}^*_{sc})$, and the result of the lemma will follow.

Crucially to the result of this lemma, the "upper" (μ) constraint never binds for an optimal flat path. To observe this, first note that it can be seen, from expressions (2.8) and (2.9) in Definition 2.3, that if $\bar{\psi} \leq \theta$ and $\lambda(\bar{\psi}) \geq 0$, then $\mu(\bar{\psi}) \geq 0$. Also, by observation of expression (2.12), it must be the case that $\bar{\psi}^* \leq \theta$ for any conceivable optimal flat path $\bar{\psi}^*$, since the "lower" (λ) constraint must bind. Therefore, it must follow that $\sup \bar{\Psi}_{sc} \in \bar{\Psi}_{fc}$.

Finally, to verify the second claim of the lemma, that $\phi(\bar{\psi}^*) \geq \phi(\tilde{\psi}^*)$, simply note that $\tilde{\Psi}_{fc} \subset \tilde{\Psi}_{sc}$ and thus $\phi(\tilde{\psi}^*) \leq \phi(\tilde{\psi}^*_{sc})$.

2.9.3 Conditions for non-flat paths

We now proceed to lay out the conditions which must hold for the various possible configurations (types A to F) of a fully-constrained optimal quasi-flat path. Firstly, we already know the necessary and sufficient condition for a maximal (type A) path from Theorem 2.I. Secondly, taking the case of a quasi-maximal (type B) path, we know that $\tilde{\psi}_1 = \hat{\pi}$. The value for $\tilde{\psi}_2$ can then be derived using the $\lambda_1(\tilde{\psi}) = 0$ condition. This should be checked as a candidate for the fully-constrained optimal quasi-flat path. We have already characterized flat (type C) paths in section 2.8.

For a carrot-constrained (type D) path, constraint (2.8) binds at point one and constraint (2.9) at points two and after. In this case, the limit to how much "carrot" can be created is imposed by the difficulty in incentivizing individuals to refrain from punishing when they would like to along the "tail" of the punishment path.[41] The optimal carrot-constrained path is characterised by the property that both $\lambda_1(\tilde{\psi}) = 0$ and $\mu_2(\tilde{\psi}) = 0$. Solving $\mu_2(\tilde{\psi}) - \lambda_1(\tilde{\psi}) = 0$ (applying identities (2.8) and (2.9) from Definition 2.3) gives us:

$$\tilde{\psi}_2 = 2\theta - \tilde{\psi}_1 \tag{2.40}$$

The fact that a carrot-maximized (type E) path is also a possibility can be seen by observation of condition (2.37). As $\tilde{\psi}_2 \longrightarrow 1$, the RHS of (2.37) goes to infinity. Therefore the inequality will definitely be fulfilled, and further increases in $\tilde{\psi}_2$ in order to decrease $\tilde{\psi}_1$ will no longer improve the severity of the path. If this happens before $\tilde{\psi}_2$ reaches ψ_j^{μ} then the optimal quasi-flat path will be "carrot-maximized". A carrot-maximized path must therefore have the property that condition (2.37) is satisfied as an equality. Rearranging this condition gives us:

$$\tilde{\psi}_2 = \frac{g\left(\tilde{\psi}_1\right)\left(1 - \theta\tilde{\psi}_1\right)\theta - 1}{g\left(\tilde{\psi}_1\right)\left(1 - \theta\tilde{\psi}_1\right)\theta - \theta} \tag{2.41}$$

The above argument from condition (2.37) also shows why a quasi-minimal (type F) path is impossible, because as $\tilde{\psi}_2 \longrightarrow 1^-$ the RHS of the inequality goes to ∞. Therefore, when raising $\tilde{\psi}_2$ in search of the optimal quasi-flat path, a path would always become carrot-maximized before it becomes quasi-minimal.

[41] Although it might be felt intuitively that if the socially efficient initial path is to be sustainable using a particular path, then co-operation with the "tail" of the punishment path would automatically also be sustainable, this does not necessarily follow because there is still a less attractive future along the tail of the punishment path if punishers co-operate, rendering the severity of punishment lower and thus making co-operation with the "tail" more difficult to incentivize than co-operation with the initial path.

2.9.4 The optimal quasi-flat path

We now have all the information we need to lay out the procedure for finding the optimal quasi-flat path. Assuming that the optimal path is not maximal, the only possibilities are a quasi-maximal, flat, carrot-constrained and carrot-maximized path. Solving (2.40) and (2.41) respectively simultaneously with equation (2.43) below provides a shortcut in finding the optimal quasi-flat path since, once the various solutions have been compared with the optimal flat path defined by (2.12) and with the quasi-maximal candidate where $\lambda_1(\tilde{\psi}) = 0$ and $\tilde{\psi}_1 = \hat{\pi}$, and the most severe path among them found,[42] we can be sure that it is optimal.[43] Most importantly, we can now prove the analogue of Theorem 2.II for the more general case of the optimal quasi-flat path. This requires the use of Lemma 2.IX.i (the appendix).[44] Once we present Theorem 2.VII, the result will be generalized to all globally optimal fully-constrained paths. This will allow us to substantiate generally, for any benefit distribution, the result that too high a level of altruism will cause the socially efficient equilibrium to break down.[45]

Theorem 2.VI

As altruism becomes perfect, the optimal quasi-flat punishment path cannot support the socially efficient equilibrium, for any value of the discount factor.

As $\theta \longrightarrow 1^-$, $\kappa\left(\tilde{\psi}^*\right) < 0$.

Proof: Firstly, observe that the most severe path in $\tilde{\Psi}_{sc}$, $\sup\tilde{\Psi}_{sc}$ must be at least as severe as the most severe path in $\tilde{\Psi}_{fc}$, $\sup\tilde{\Psi}_{fc}$. Therefore $\phi(\sup\tilde{\Psi}_{sc}) \geq \phi(\sup\tilde{\Psi}_{fc})$ and so $\phi\left(\tilde{\psi}^*_{sc}\right) \geq \phi\left(\tilde{\psi}^*\right)$. Now, combining with expressions (2.10) and (2.11), we know that:

$$\kappa\left(\tilde{\psi}^*\right) \leq \phi\left(\tilde{\psi}^*_{sc}\right) + \theta - 1 \tag{2.42}$$

Theorem 2.II, combined with Lemma 2.IX.i, has already established that the RHS of expression (2.42) goes to 0^- as $\theta \longrightarrow 1^-$. Therefore the LHS of (2.42) must be strictly negative as $\theta \longrightarrow 1^-$.

---

[42] Note that we must also check the sustainability of each "candidate" because these are necessary rather than sufficient conditions.

[43] We use this method in subsection 2.9.6 to analyse a specific example of a carrot-maximized and a carrot-constrained path.

[44] In Lemma 2.IX.i, we impose only that the optimal path be semi-constrained, partly to simplify the proof but, more importantly, because we are later able to generate semi-constrained quasi-flat paths by "flattening-out" generic paths.

[45] In the course of the proof for Lemma 2.IX.i, we exhaustively derive the conditions on $g(\pi)$ under which the optimal semi-constrained quasi-flat path can be maximal, quasi-maximal, carrot-maximized and flat. The key result is that as $\theta \longrightarrow 1^-$, the optimal semi-constrained quasi-flat punishment path must be flat.

2.9.5 Illustration: quasi-maximal paths

Given assumption (2.3), condition (2.38) will definitely hold for a uniform distribution with support between $\hat{\pi}$ and 1. It is however, instructive to look at some examples where the optimal quasi-flat path is not flat. As we have seen, this requires a distribution with a high enough probability density around the optimal flat trigger level. The simplest distribution for the benefit which can produce this result is a triangular distribution. The triangular probability density function $g(\pi) = 50(1 - \pi)$ with support between 0.8 and 1 and illustrated below is one such example.

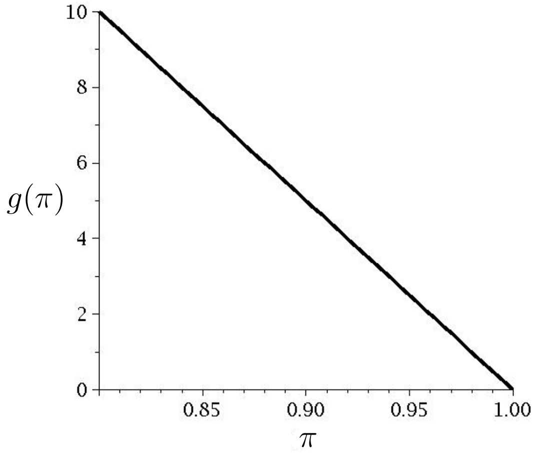

Figure 2.16: Triangular probability density function

The optimal quasi-flat path can be computed by using the $\lambda_1(\tilde{\psi}) = 0$ condition to derive the following expression for $\tilde{\psi}_1$, and then numerically solving for $\tilde{\psi}_1$ given each particular $\tilde{\psi}_2$:

$$\tilde{\psi}_1 = \theta - \frac{\delta}{1-\delta}\int_{\tilde{\psi}_2}^{1}(\theta\pi - \theta)g(\pi)d\pi + \delta\int_{\tilde{\psi}_1}^{1}(\theta\pi - 1)g(\pi)d\pi + \frac{\delta^2}{1-\delta}\int_{\tilde{\psi}_2}^{1}(\theta\pi - 1)g(\pi)d\pi \qquad (2.43)$$

Figures 2.17 through 2.22 illustrate the results of this exercise for $\delta = 0.48$ and $\delta = 0.49$, with $\theta = 0.9$. Figures 2.17 and 2.20 show the value of $\tilde{\psi}_2$ imputed from (2.43) for each value of $\tilde{\psi}_1$. They also show the optimal flat trigger level $\bar{\psi}^*$ (numerically solved using (2.12)), and the $\tilde{\psi}_2 = \tilde{\psi}_1$ line.

If $\delta = 0.5$, then solving equation (2.12) to find the optimal flat path for the assumed triangular benefit distribution yields $\bar{\psi}^* = 0.8$. This would be a maximal (type A) path involving punishment for all benefit values all of the time, as $\hat{\pi} = 0.8$. By making δ slightly lower than this, we prevent the optimal flat path from being maximal because the future is no longer "important enough" to sustain it. For the triangular distribution, this results in the optimal path being quasi-maximal (type B).

Figures 2.18, 2.19, 2.21 and 2.22 show the values of the various co-operation constraints, and the severity of the punishment paths (ϕ).[46] In both cases, all of the constraints are positive, showing that

[46] Note that, because (2.8) binds at point 1 (and so (2.9) cannot), λ_1 and μ_1 do not need to be displayed.

the punishment path is itself sustainable and supports the initial path. In both cases, the highest value of $\phi(\tilde{\psi})$ occurs where $\tilde{\psi}_1 = 0.8$. We have, therefore, two cases where the optimal quasi-flat punishment path is "maxed-out" at point 1, so that $\tilde{\psi}_1^* = \hat{\pi}$. The "tail" of the path is, however, not maxed-out.

A notable feature of the graph of the imputed $\tilde{\psi}_2$ in figure 2.17 is that it initially climbs as $\tilde{\psi}_1$ is reduced below $\bar{\psi}^*$. This means that $\tilde{\psi}_2$ must initially be raised to compensate for a reduction in $\tilde{\psi}_1$ if the point 1 co-operation constraint is to remain unbroken. Eventually, on the other hand, the imputed $\tilde{\psi}_2$ graph begins to fall as ψ_1 is reduced. Since this line continues to lie above the dashed $\tilde{\psi}_2 = \tilde{\psi}_1$ line, these more severe quasi-flat punishment paths can only be achieved by introducing some carrot-and-stick element into the punishment path. By doing this, however it is possible to reduce *both* trigger levels below the trigger level of the optimal flat punishment path.

Figure 2.20 shows a situation where the graph of the imputed $\tilde{\psi}_2$ has a positive gradient for any change in $\tilde{\psi}_1$. This means that it is now possible to reach more effective paths which remain sustainable by simultaneously reducing *both* trigger levels below the optimal flat trigger level, *provided* that $\tilde{\psi}_2$ is reduced by less than $\tilde{\psi}_1$. Here we have a classic illustration of Abreu's principle that, when being optimally punished, individuals can be persuaded to put up with an even more unpleasant present in exchange for a relatively less unpleasant future.

The carrot-and-stick structure of optimal quasi-flat punishment paths illustrated here is similar to the optimal penal codes derived for infinitely repeated oligopoly games. Abreu himself derived the properties of optimal penal codes in the infinitely repeated Cournot game (Abreu, 1986). He found that, provided there is sufficiently low discounting, optimal punishment paths would involve one period of very high output and negative profits followed by a return to fully collusive behaviour. Such paths cannot be any worse than the 0 discounted stream of profits that could be achieved by shutting down.

Lambson built on Abreu's work to derive optimal penal codes in the infinitely repeated homogeneous product Bertrand model with identical firms and capacity constraints (Lambson, 1987). Optimal punishment paths in this model involve a number of periods of low profits followed by a return to fully collusive behaviour. This is because the losses per firm in any one period cannot be infinite in the Bertrand model. The sequential punishment model is more like Bertrand in that the severity of the punishment path at point 1 is constrained by the fact that $\tilde{\psi}_1$ cannot be any lower than $\hat{\pi}$. This means that generally there will also be some punishment along the "tail" of an optimal quasi-flat path.

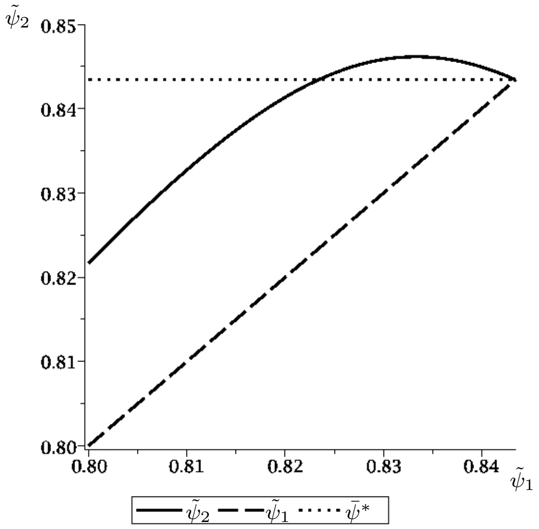

Figure 2.17: Trigger levels for a quasi-maximal path - $\theta = 0.9$, $\delta = 0.48$.

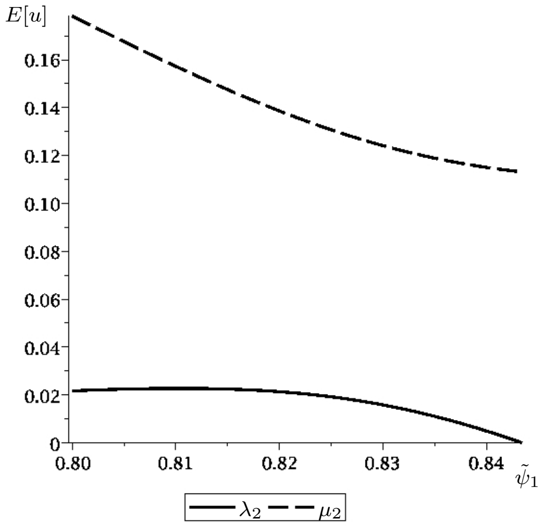

Figure 2.18: Co-operation constraints for a quasi-maximal path - $\theta = 0.9$, $\delta = 0.48$.

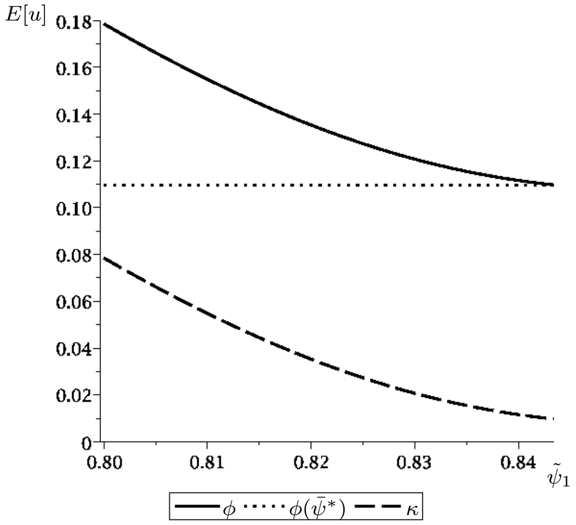

Figure 2.19: Severity of a quasi-maximal path - $\theta = 0.9$, $\delta = 0.48$.

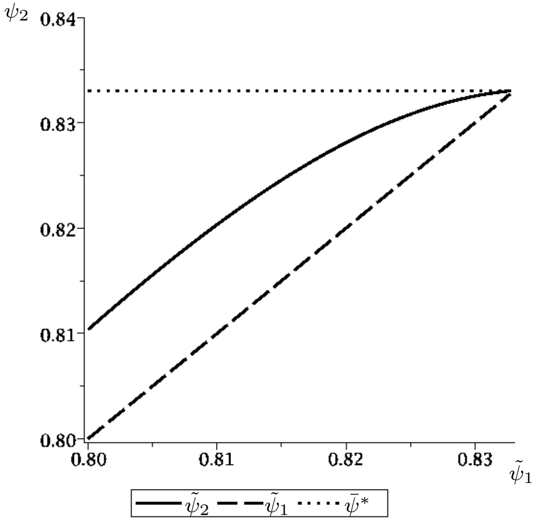

Figure 2.20: Trigger levels for a quasi-maximal path - $\theta = 0.9$, $\delta = 0.49$.

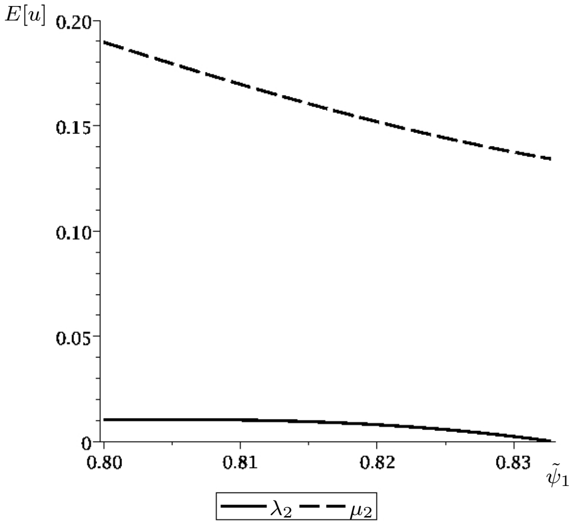

Figure 2.21: Co-operation constraints for a quasi-maximal path -$\theta = 0.9$, $\delta = 0.49$.

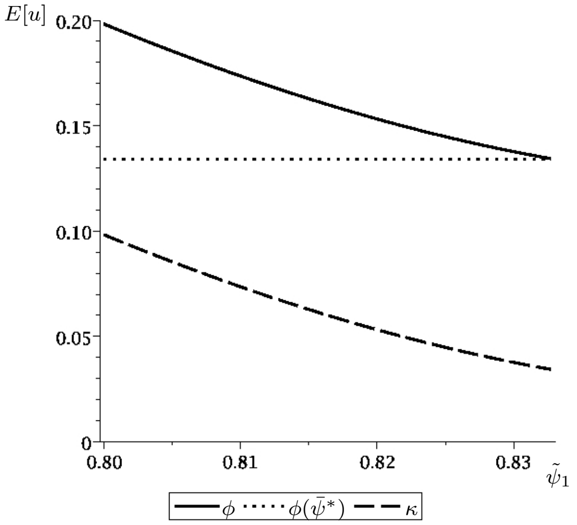

Figure 2.22: Severity of a quasi-maximal path - $\theta = 0.9$, $\delta = 0.49$.

2.9.6 Illustration: carrot-constrained and carrot-maximized paths

In this subsection, we will present specific numerical examples of a flat, a carrot-constrained and a carrot-maximized path, to illustrate more of the general principles discussed in subsection 2.9.1. Figures 2.24, 2.25 and 2.26 respectively show the numerically calculated values for $\tilde{\psi}_2$ given $\tilde{\psi}_1$, the various co-operation constraints[47] and the effectiveness of the punishment path ϕ for the triangular distribution shown in figure 2.16, with $\delta = \frac{1}{3}$ and $\theta = 0.9$. Figures 2.27, 2.28 and 2.29 show the results for the same exercise conducted using the probability density function given in (2.44) below, and shown in figure 2.23, with $\delta = \frac{1}{8}$ and $\theta = 0.9$, $\hat{\pi} = 0.8$ and support between 0.8 and 1. Figures 2.30, 2.31 and 2.32 show the analysis for the same distribution but with $\delta = \frac{1}{16}$.

$$g(\pi) = 17.5\left(1 - \left(\left(\frac{2\pi - \hat{\pi} - 1}{1 - \hat{\pi}}\right)^2\right)^{\frac{1}{5}}\right) \tag{2.44}$$

The key features to note are, firstly, that in all three cases μ_2 reaches 0 before $\tilde{\psi}_1$ reaches $\hat{\pi} = 0.8$, showing that the optimal path cannot possibly be quasi-maximal. Secondly, whereas in figure 2.26, ϕ decreases as $\tilde{\psi}_1$ is reduced, showing that the optimal quasi-flat path is flat, in figure 2.29, ϕ reaches an interior maximum, showing that the optimal quasi-flat path is carrot-maximized. In figure 2.32, meanwhile, ϕ reaches a constrained maximum at the carrot-constrained value of $\tilde{\psi}_1$, where the x-axis begins. Note, finally, that the starting value on the x-axis, corresponding to the carrot-maximized "candidate" for the optimal path in each case, and the "carrot-maximum" denoted by the vertical dotted line in figure 2.29 were derived using the shortcut method described earlier in section 2.9.4.

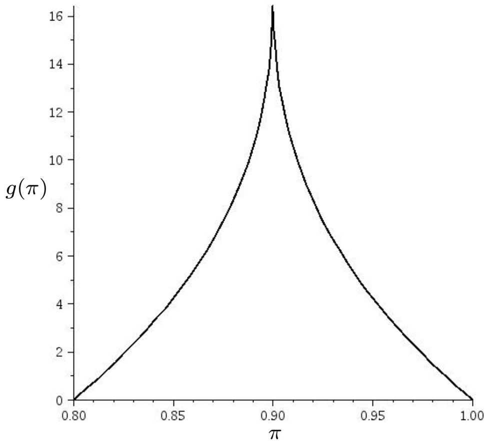

Figure 2.23: Probability density for a carrot-maximized or carrot-constrained path

[47] Note that we do not show $\kappa(\tilde{\psi})$ here as, in some of these cases, the optimal quasi-flat path is not severe enough to enable to socially efficient initial path to be supported. However, an optimal punishment path would still be needed to support the optimal second-best equilibrium, as we will discuss in section 2.11.

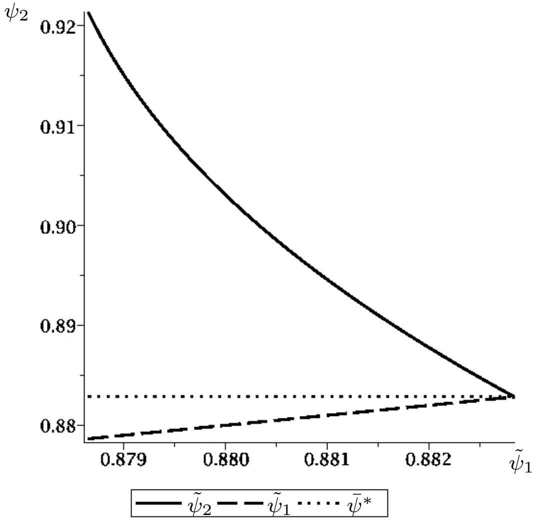

Figure 2.24: Trigger levels for a flat path

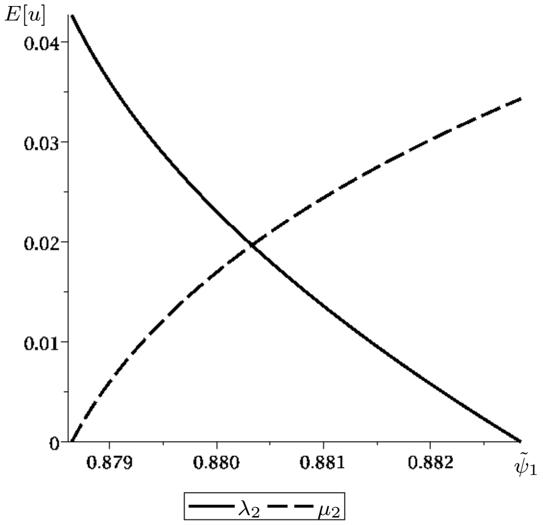

Figure 2.25: Co-operation constraints for a flat path

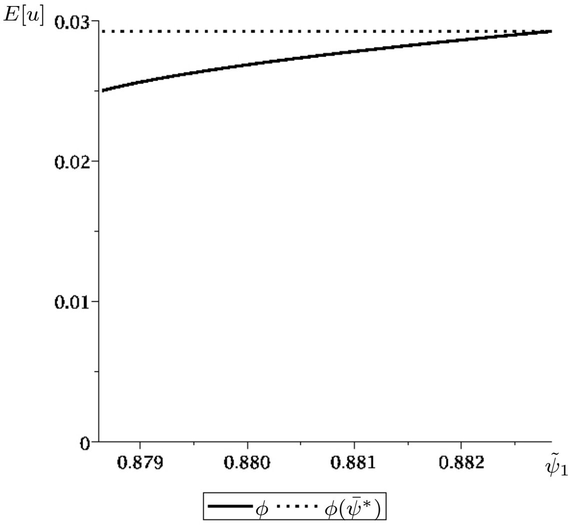

Figure 2.26: Severity of a flat path

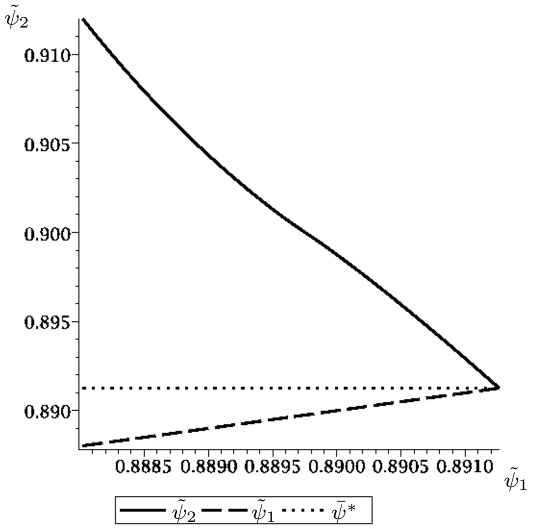

Figure 2.27: Trigger levels for a carrot-maximized path

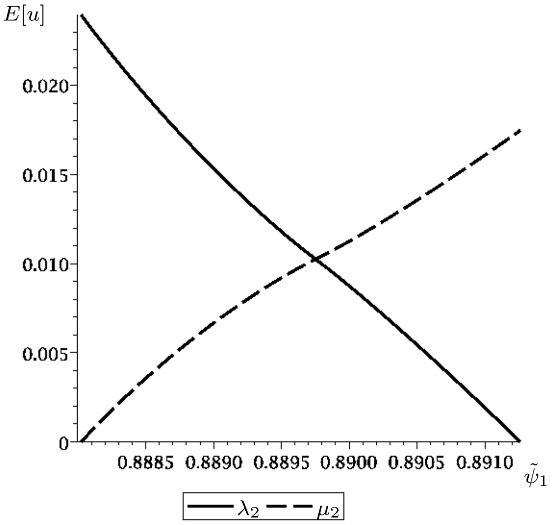

Figure 2.28: Co-operation constraints for a carrot-maximized path

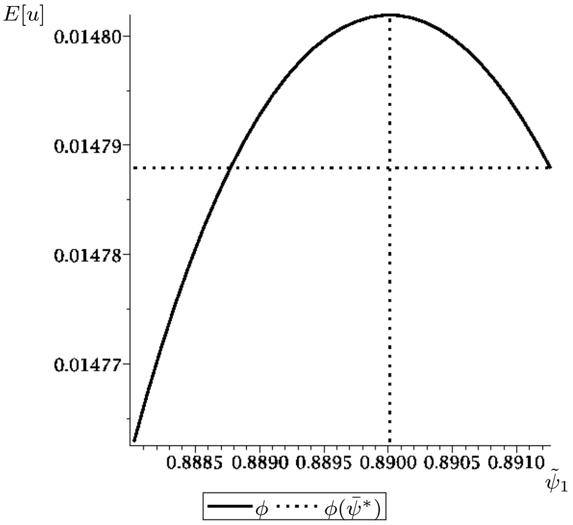

Figure 2.29: Severity of a carrot-maximized path

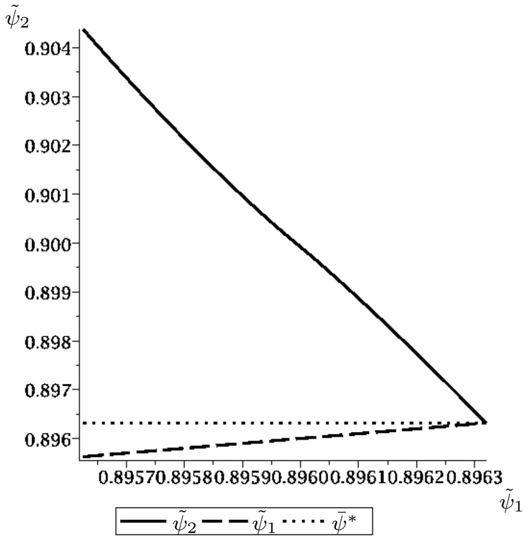

Figure 2.30: Trigger levels for a carrot-constrained path

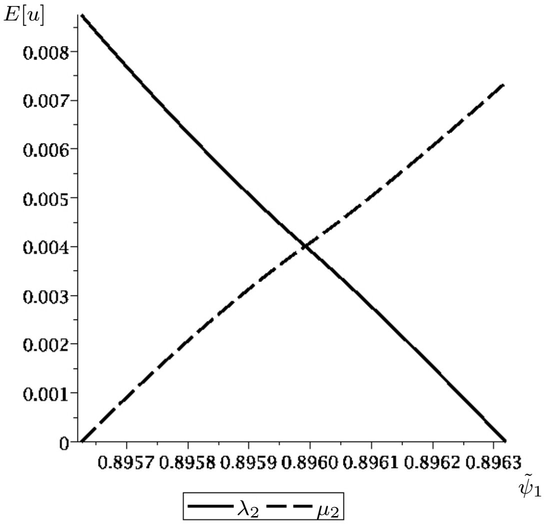

Figure 2.31: Co-operation constraints for a carrot-constrained path

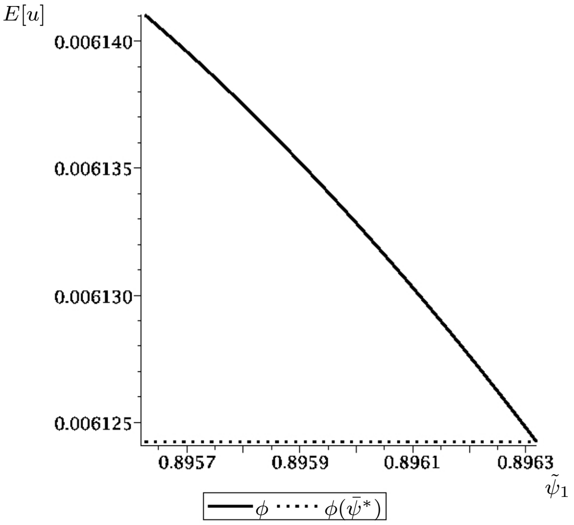

Figure 2.32: Severity of a carrot-constrained path

2.10 *The Optimal Generic Path*

We are now ready to further generalize Theorems 2.II and 2.VI to the case where the optimal generic punishment path is used to support the initial path. The result hinges upon three intuitive observations. Firstly, the optimal semi-constrained path is quasi-flat, because it is always possible to take any given optimal semi-constrained path and "flatten out" the tail, producing an equally severe path without breaking any of the co-operation constraints. Secondly, as $\theta \longrightarrow 1^-$, it is impossible to support the socially efficient equilibrium using the optimal semi-constrained quasi-flat path, because it must become flat (this is proved in Lemma 2.IX.i), and we already know (from Theorem 2.II) that the result holds for flat paths. Thirdly, since the optimal fully-constrained generic path must be weakly less severe than the optimal semi-constrained path, then it must also follow that, as $\theta \longrightarrow 1^-$, the socially efficient equilibrium cannot be supported by *any* path.

Theorem 2.VII will work by arguing, firstly, that any generic optimal punishment path can be replaced by a semi-constrained quasi-flat path constructed by "flattening out" to the point 1 U-average from point 2 onwards. This newly constructed quasi-flat path will continue to fulfil the point 1 and point 2 cooperation conditions[48] $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$. It will therefore be in the set of semi-constrained paths, and will be as severe as the path it was generated from.

Applying Lemma 2.IX.i (the appendix), it will therefore be shown that the supportability constraint on the socially efficient initial path, given the use of the generic optimal path to punish a deviation, is broken as $\theta \longrightarrow 1^-$. Whether or not globally optimal paths are flat, it is thus established that intermediate values of the coefficient of altruism are best able to support a socially efficient equilibrium. Theorem 2.VII will, as a result, be crucial to establishing the general applicability of the key results from section 2.8, which form the core contribution of this paper. We also show, in Theorem 2.VIII, that the globally optimal path is definitely flat if the benefit distribution is sufficiently flat.

Lemma 2.VII.i below is an essential building block. It establishes that for any generic punishment path looking forward from any point k, the U-average *must* lie equal to or above the V-average. It should be noted that the result from Lemma 2.VII.i is based on the assumption of risk neutral agents. Although mathematically analogous to differing levels of risk aversion, the result is in fact generated by differing attitudes towards the benefit values under which punishment is inflicted.

[48] The co-operation conditions for point 3 and after are identical to those at point 2.

Lemma 2.VII.i

For any punishment path, at any point, the U-average is weakly greater than the V-average.

$\forall_{\psi}\forall_{k} : U^{-1}\left(U_k\left(\psi\right)\right) \geq V^{-1}\left(V_k\left(\psi\right)\right)$.

Proof: This result follows from an application of the principle of stochastic dominance, a key principle in the economics of risk (Rothschild & Stiglitz, 1970). First, note from Definition 2.3 that the per period average utility functions for the person being punished and a neutral observer can be rewritten as $U_k(\psi) \equiv \sum_{i=1}^{\infty}[p_i U(\psi_{i+k})]$ and $V_k(\psi) \equiv \sum_{i=1}^{\infty}[p_i V(\psi_{i+k})]$ where $p_i = (1-\delta)\delta^{i-1}$ and $\sum_{i=1}^{\infty}[p_i] = 1$. This means that $U(\psi)$ and $V(\psi)$ can be thought of as "expected utility functions", with the discount factor for each point along the path taking the role of probabilities for different outcomes of a lottery. The "expected value" of a particular path ψ looking forward from point k can then be defined as $E_k[\psi] = \sum_{i=1}^{\infty}[p_i\psi_{i+k}]$. Since this "expected value" is equal for both the punishee and the neutral observer, the discounted expected utility along a path is exactly analogous to the expected utility of a risky prospect. Therefore if we can show that the "neutral observer" is more "risk averse" than the punishee, the result of the lemma will follow.

The coefficient of absolute risk aversion $R_a = -\frac{U''}{U'}$ measures the degree of concavity of a utility function. If it is always higher for one function than another, then the corresponding agent is the more risk averse (Diamond & Stiglitz, 1974). For the two types of agent under consideration (with utility functions $U(\bar{\psi})$ and $V(\bar{\psi})$ respectively), the CARA works out as the following.

$$R_a^U = \frac{\theta}{1-\theta\bar{\psi}} - \frac{g'(\bar{\psi})}{g(\bar{\psi})} \qquad R_a^V = \frac{\theta}{\theta-\theta\bar{\psi}} - \frac{g'(\bar{\psi})}{g(\bar{\psi})} \tag{2.45}$$

Since R_a^V is always unambiguously greater than R_a^U, the neutral observer is more "risk averse" than the punishee, and hence will always have a lower "certainty equivalent" from a given path looking forward, which, from Definition 2.3, is precisely analogous to $V^{-1}(V_k(\psi))$, as opposed to $U^{-1}(U_k(\psi))$.

The sensitivity of a "neutral observer" to more "wasteful" punishment when watching others being punished is greater than the sensitivity of the punishee. This makes intuitive sense, since, to take the example of a fine, the person being fined is mainly affected in social utility terms by the fact that they are fined, whereas altruistic neutral observers who value the felicity of the person fined and the recipient of the revenue equally will be more sensitive to any deadweight loss from punishment. The intuition for the result can also be related to the "sacrifice ratio" derived in expression (2.32). Punishing within a certain "bracket" of values of π, with a fixed width, has a greater effect on $U^{-1}(U_k(\psi))$ (increasing the "stick") relative to $V^{-1}(V_k(\psi))$ (decreasing the "carrot") the higher the bracket. $V^{-1}(V_k(\psi))$ is maximized when a particular value of $U^{-1}(U_k(\psi))$ is generated by a flat path looking forwards.

Definition 2.7: Let $\gamma : \Psi \longrightarrow \tilde{\Psi}$ be the function which constructs a quasi-flat path from a generic path by "flattening-out" the trigger levels from point 2 onwards to the point 1 U-average. This means that $\gamma_1(\psi) = \psi_1$ and $\gamma_2(\psi) = U^{-1}(U_1(\psi))$. Note also that $\phi(\gamma(\psi)) = \phi(\psi)$.[49]

Lemma 2.VII.ii

If a punishment path is optimal, then the quasi-flat path constructed by "flattening it out" will be in the set of semi-constrained paths.

If a path ψ^* is optimal then $\forall_k : \lambda_k(\gamma(\psi^*)) \geq 0$, therefore $\gamma(\psi^*) \in \tilde{\Psi}_{sc}$.

Proof: Firstly, observe that, given the result from Lemma 2.VII.i, $U_0(\gamma(\psi^*)) = U_0(\psi^*)$ and $\forall_k : V_k(\gamma(\psi^*)) \geq V_1(\psi^*)$. To see this, note the following:

$$\left(\frac{\delta}{1-\delta}\right) U_0(\psi) \equiv \delta \int_{\psi_1}^{1} (\theta\pi - 1) g(\pi) d\pi + \frac{\delta^2}{1-\delta} \int_{U^{-1}(U_1(\psi))}^{1} (\theta\pi - 1) g(\pi) d\pi$$

$$\left(\frac{\delta}{1-\delta}\right) U_0(\gamma(\psi)) \equiv \delta \int_{\gamma_1(\psi)}^{1} (\theta\pi - 1) g(\pi) d\pi + \frac{\delta^2}{1-\delta} \int_{\gamma_2(\psi)}^{1} (\theta\pi - 1) g(\pi) d\pi$$

$$\left(\frac{\delta}{1-\delta}\right) V_1(\psi) \equiv \frac{\delta}{1-\delta} \int_{V^{-1}(V_1(\psi))}^{1} (\theta\pi - \theta) g(\pi) d\pi$$

$$\forall_k : \left(\frac{\delta}{1-\delta}\right) V_k(\gamma(\psi)) \equiv \frac{\delta}{1-\delta} \int_{\gamma_2(\psi)}^{1} (\theta\pi - \theta) g(\pi) d\pi$$

The claim is then straightforward to verify once it is noted that $\gamma_1(\psi) = \psi_1$ and $\gamma_2(\psi) = U^{-1}(U_1(\psi))$, since, by Lemma 2.VII.i, $U^{-1}(U_1(\psi)) \geq V^{-1}(V_1(\psi))$.

Now we can proceed to note that:

$$\lambda_1(\gamma(\psi)) \equiv \left(\frac{\delta}{1-\delta}\right) V_1(\gamma(\psi)) - \left(\frac{\delta}{1-\delta}\right) U_0(\gamma(\psi)) + \gamma_1(\psi) - \theta \tag{2.46}$$

$$\forall_{k\geq 2} : \lambda_k(\gamma(\psi)) \equiv \left(\frac{\delta}{1-\delta}\right) V_k(\gamma(\psi)) - \left(\frac{\delta}{1-\delta}\right) U_0(\gamma(\psi)) + \gamma_2(\psi) - \theta \tag{2.47}$$

Now take an optimal path ψ^*. By assumption, ψ^* is sustainable, and so:

$$\lambda_1(\psi^*) \equiv \left(\frac{\delta}{1-\delta}\right) V_1(\psi^*) - \left(\frac{\delta}{1-\delta}\right) U_0(\psi^*) + \psi_1^* - \theta \geq 0 \tag{2.48}$$

Given the result from Lemma 2.V.i that $\psi_1^* \leq U^{-1}(U_1(\psi^*))$, along with all the observations noted so far, condition (2.48) is sufficient for (2.46) and (2.47) to be weakly positive for all relevant k.

[49] Note that $\gamma_k(\psi)$ is shorthand for $(\gamma(\psi))_k$ - the k^{th} trigger level in the path $\gamma(\psi)$.

Lemma 2.VII.iii

The optimal semi-constrained quasi-flat punishment path is at least as severe as the optimal fully-constrained generic punishment path.

$\phi\left(\tilde{\psi}^*_{sc}\right) \geq \phi\left(\psi^*\right)$.

Proof: Consider the optimal generic path $\psi^* \in \Psi_{fc}$. By Lemma 2.VII.ii, $\gamma(\psi^*) \in \tilde{\Psi}_{sc}$. Also, by definition, $\phi(\gamma(\psi^*)) = \phi\left(\psi^*\right)$. Therefore the most severe path in $\tilde{\Psi}_{sc}$, $\sup \tilde{\Psi}_{sc}$ must be at least as severe as the most severe path in Ψ_{fc}, $\sup \Psi_{fc}$. Thus $\phi(\sup \tilde{\Psi}_{sc}) \geq \phi(\sup \Psi_{fc})$ and so $\phi(\tilde{\psi}^*_{sc}) \geq \phi\left(\psi^*\right)$.

Theorem 2.VII

As altruism becomes perfect, the optimal generic punishment path cannot support the socially efficient equilibrium, for any value of the discount factor.

As $\theta \longrightarrow 1^-$, $\kappa\left(\psi^*\right) < 0$.

Proof: First, note, from expressions (2.10) and (2.11) in Definition 2.3, that:

$$\kappa\left(\psi^*\right) = \phi\left(\psi^*\right) + \theta - 1$$

By Lemma 2.VII.iii, this implies that:

$$\kappa\left(\psi^*\right) \leq \phi\left(\tilde{\psi}^*_{sc}\right) + \theta - 1 \tag{2.49}$$

Theorem 2.II, combined with Lemma 2.IX.i, has already established that the RHS of expression (2.49) goes to 0^- as $\theta \longrightarrow 1^-$. Therefore the LHS of (2.49) must be strictly negative as $\theta \longrightarrow 1^-$.

Theorem 2.VIII

If the benefit distribution is sufficiently flat, then the optimal generic punishment path is flat.

If $\forall_\pi : g(\pi) < \frac{1}{\theta(1-\hat{\pi})}$ then $\phi\left(\bar{\psi}^*\right) \geq \phi\left(\psi^*\right)$.

Proof: From Theorem 2.V, we know that if $\forall_\pi : g(\pi) < \frac{1}{\theta(1-\hat{\pi})}$, then $\phi\left(\bar{\psi}^*\right) \geq \phi\left(\tilde{\psi}^*_{sc}\right)$. Applying Lemma 2.VII.iii, the result follows straightforwardly.

2.11 *Second-Best Equilibria*

For any particular level of the coefficient of altruism θ, if the discount factor δ is low enough, so that players are sufficiently impatient, then the socially efficient initial path will not be supportable. There will still, however, exist a second-best optimal subgame-perfect Nash equilibrium, supported by the optimal punishment path, in the sense that the associated initial path maximizes efficiency by minimising the range of benefit values for which punishment occurs. Given condition (2.38) then, as shown in Theorem 2.VIII, the optimal punishment path which supports such an optimal equilibrium will be flat. For the remainder of the paper, we will assume that this condition is satisfied.

Along the second-best optimal initial path, the punishment path will be triggered when an individual punishes for a value of the benefit below trigger level $\bar{\psi}_h$. (Intuitively, the most attractive punishment opportunities will be the most difficult to deter along the initial path.) All players will then switch to a path where punishment is carried out above trigger level $\bar{\psi}^*$, derived from the optimal flat punishment path (which is globally optimal, under the assumptions made). Assuming an interior solution for the optimal flat path where $\bar{\psi}^* > \hat{\pi}$, $\bar{\psi}^*$ and its total derivative with respect to θ will be respectively given by (2.12) and (2.13).

The supportability constraint on the initial path will therefore be:

$$\bar{\psi}_h - \theta \leq \frac{\delta}{1-\delta}\int_{\bar{\psi}_h}^{1}(\theta\pi-\theta)g(\pi)d\pi - \frac{\delta}{1-\delta}\int_{\bar{\psi}^*}^{1}(\theta\pi-1)g(\pi)d\pi \tag{2.50}$$

When $\bar{\psi}_h$ is set just high enough to make (2.50) bind, to give us the optimal second-best equilibrium, we will have the following implicit definition of $\bar{\psi}_h^*$ in terms of itself, $\bar{\psi}^*$, θ and δ:

$$\bar{\psi}_h^* = \theta + \frac{\delta}{1-\delta}\left(\int_{\bar{\psi}^*}^{1} g\left(\pi\right)d\pi - \theta\int_{\bar{\psi}^*}^{\bar{\psi}_h^*}\pi g\left(\pi\right)d\pi - \theta\int_{\bar{\psi}_h^*}^{1} g\left(\pi\right)d\pi\right) \tag{2.51}$$

Totally differentiating (2.51) with respect to θ and solving for $\frac{d\bar{\psi}_h^*}{d\theta}$ gives us the following expression for the overall derivative of $\bar{\psi}_h^*$ with respect to the coefficient of altruism:

$$\frac{d\bar{\psi}_h^*}{d\theta} = \frac{(1-\delta) - \delta\left(\int_{\bar{\psi}^*}^{\bar{\psi}_h^*} g\left(\pi\right)\pi d\pi + \int_{\bar{\psi}_h^*}^{1} g\left(\pi\right)d\pi\right) - \delta\left(1-\theta\bar{\psi}^*\right)g\left(\bar{\psi}^*\right)\frac{d\bar{\psi}^*}{d\theta}}{(1-\delta) - \delta\theta\left(1-\bar{\psi}_h^*\right)g\left(\bar{\psi}_h^*\right)} \tag{2.52}$$

By observation of (2.12) and (2.51), as $\theta \longrightarrow 1^-$, $\bar{\psi}^* \longrightarrow 1^-$ and therefore $\bar{\psi}_h^* \longrightarrow 1^-$. Additionally, note that as $\theta \longrightarrow 1^-$, the RHS of (2.52) goes to 1. These observations imply that there must be a region where increasing the coefficient of altruism results in a decreasing $\bar{\psi}_h^*$, followed by a region of increasing $\bar{\psi}_h^*$. Note also that as $\delta \longrightarrow 0$, the RHS of (2.52) again goes to 1. This implies that, for a low enough δ, increasing θ always increases $\bar{\psi}_h^*$ (intuitively, this is where the future counts for sufficiently little that the severity effect, even in combination with the willingness effect, is never sufficient to outweigh the temptation effect).

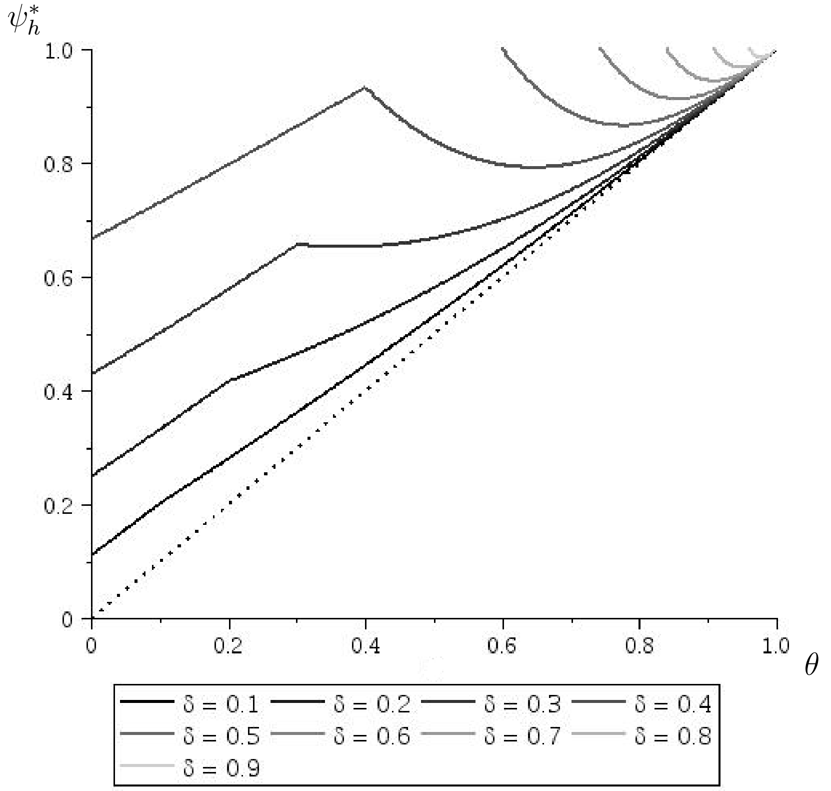

Figure 2.33: Second-best equilibria, $\hat{\pi} = 0$ and $g(\pi) = 1$

Figure 2.33 shows the highest $\bar{\psi}_h^*$ which is supportable along the initial path given different values of θ (along the x-axis) and δ. It can be seen that the curve always has a slope of 1 as $\theta \longrightarrow 1$, and for all θ when δ is low. Importantly, there is always a level of θ high enough but lower than 1 where $\bar{\psi}_h^*$ falls below one and then back up to one as $\theta \longrightarrow 1$. This corresponds to the black region in figure 2.4 and derived analytically in Theorems 2.I through 2.IV. Finally, for high enough δ with a low enough θ, $\bar{\psi}_h^*$ goes above one (i.e. the graph gets "cut off").

An important conclusion to draw from figure 2.33 is that the efficiency loss from too high a level of altruism can be non-negligible. Although as $\theta \longrightarrow 1$, $\bar{\psi}_h^* \longrightarrow 1$, there will exist intermediate levels of altruism where an increase in the coefficient of altruism to a higher intermediate level (which is still less than 1) could make the efficiency of the optimal second-best outcome significantly lower. Altruism is in many realistic cases a “double-edged sword” in the sequential punishment model, and greater altruism will in general be socially harmful in a significant way.

2.12 *Conclusion*

This paper has taken two areas of economic theory, the modelling of altruistic preferences and the structure of optimal punishment paths, and shown how they can interact to produce interesting results in a new type of model, the sequential punishment model - a simple infinite-move sequential game with perfect information and discounting - where players move by choosing whether or not to take opportunities to inflict harm upon others with benefit to themselves. Essentially, the model is an abstract representation of the fundamentally vicarious nature of human interaction in any kind of society, whatever its organizational principles.

The central implication of the analysis is that excessive altruism will interfere detrimentally with punishment systems, “denting” them in such a manner that social welfare is reduced compared to a situation with lower altruism. Greater “intrinsic” altruistic motivation generally weakens the effectiveness of “extrinsic” social incentive mechanisms in inducing socially beneficial altruistic behaviour. Since societal punishment systems “amplify” the impact of altruistic preferences upon individual altruistic behaviour, the relationship between altrustic preferences, altruistic behaviour and social efficiency is not straightforward. The sequential punishment model shows that perverse interaction effects are not only possible, but likely and, indeed, general.

The concepts of willingness, severity and temptation effects which were used to analyse the subgame perfect equilibria of the sequential punishment model, and to establish the existence of a socially optimal level of altruism, should present themselves in other contexts. They would, for instance, be highly relevant to the analysis of optimal taxation, the economic theory of criminal punishment, and to issues surrounding the evolution of altruistic preferences; see, for example, “Punishment and the Potency of Group Selection” (Chapter 3).

Sections 2.1 through 2.7 set up the notation for the sequential punishment model, and related this to Abreu's framework of optimal punishment paths, originally developed for repeated stage games. The main body of novel results for this paper is in sections 2.8, 2.9 and 2.10, which progressively generalize the core result of the paper, that as altruism becomes perfect the socially efficient equilibrium breaks down, to the equilibria supportable by the optimal flat punishment path, the optimal quasi-flat path and, finally, the optimal generic punishment path.

In the sequential punishment model, the interaction between the severity, willingness and temptation effects can be conclusively seen to lead to the result that an intermediate "Goldilocks" level of altruism is socially optimal. This result was established (initially for equilibria supportable by flat paths), in Theorem 2.II. The key intuition for this result is that, for a low enough value of the discount factor δ, the temptation effect must initially dominate the severity and willingness effects as θ is reduced from below 1. Since social efficiency is only barely supportable at $\theta = 1$, the constraint for supportability *must* be broken as $\theta \longrightarrow 1^-$ for the relevant values of δ.

If individuals are sufficiently impatient, excessive malevolence will, on the other hand, be socially damaging, due to the dominance of the temptation effect over the severity effect as $\theta \longrightarrow -\infty$. In contrast however, if the discount factor δ is high enough, then even infinite malevolence does not break the socially efficient equilibrium, because the severity effect will outweigh the temptation effect as $\theta \longrightarrow -\infty$ (Theorem 2.III). The sequential punishment model also demonstrates, therefore, an important asymmetry, in that high altruism is in general more damaging than extreme malevolence. With sufficiently effective monitoring (leading to a δ close to 1) malevolent preferences are not of concern from a social welfare perspective, but excessively altruistic ones remain so.

The analysis in section 2.9 involved the investigation of quasi-flat paths, which maintain a simple structure of carrot-and-stick punishment, and are an interesting illustration of the general principles governing optimal penal codes in and of themselves. Quasi-flat paths are a specific feature which emerges from the application of the principles of optimal carrot-and-stick punishment to the sequential punishment model. They occur because the presence of partially altruistic preferences leads to a "flattening-out" of the tail of the optimal punishment path. This is a key difference between the nature of optimal punishment paths in the sequential punishment model and those found in the infinitely-repeated Cournot and Bertrand simultaneous oligopoly stage games.

The optimality of quasi-flat paths is driven by the higher relative concavity of the "inter-temporal" utility function of a "neutral observer" relative to the punishee in a punishment equilibrium. This is also an interesting result which should have wider implications. The broader society is more sensitive to the inefficiencies brought about by "wasteful" punishment technology than the individual being punished. This is because the individual cares primarily about the fact that they are punished, and so places less relative weighting on the deadweight loss to society from punishment. A similar phenomenon should emerge in any model with altruistic agents and punishment technologies. There are therefore potential applications in optimal taxation theory and in rational choice models of criminality and punishment.

Theorem 2.VII completes the generalization of the results from section 2.8 so that they apply to any continuous distribution of the benefit with support between $\hat{\pi}$ and 1 (where $0 \leq \hat{\pi} < 1$), and when the optimal generic punishment path is used to support the socially efficient equilibrium by punishing any deviation from the socially efficient initial path. This is a satisfyingly general result, although it does require ruling out situations where individuals would actually *enjoy* being punished. Another limitation of this paper is that the case where $\theta \geq 1$, so that individuals are *martyrs*, who care about others *more* than themselves, has been excluded from the outset.

Section 2.11 demonstrated that there is a potentially significant loss of social efficiency from the detrimental effects of too high a level of altruism on the social incentive systems used to induce co-operation via the punishment of transgressors. This was an important final piece of the argument, as it is necessary not only to show that too high a level of altruism will break the supportability constraint on the socially efficient outcome, but also that the resultant loss of social welfare in a second-best world will often be non-negligible.

Appendix

Lemma 2.IX.i

(a) As altruism becomes perfect, the optimal semi-constrained quasi-flat path becomes flat.

(b) If the benefit distribution is sufficiently flat, the optimal semi-constrained quasi-flat path becomes flat.

(a) As $\theta \longrightarrow 1^{-}$, $\tilde{\psi}^{*}_{sc} \in \bar{\Psi}_{sc}$.

(b) If $\forall_{\pi} : g(\pi) < \frac{1}{\theta(1-\hat{\pi})}$ then $\tilde{\psi}^{*}_{sc} \in \bar{\Psi}_{sc}$.

Proof: The optimal semi-constrained quasi flat path $\tilde{\psi}^{*}_{sc}$ can be found by solving a constrained optimization problem (Lagrange, 1806) (Simon & Blume, 1994) using the following Lagrangean function:

$$
\begin{aligned}
\mathcal{L} = & - \left(\frac{\delta}{1-\delta}\right) U_0\left(\tilde{\psi}\right) + \bar{\lambda}_1 \left(\left(\frac{\delta}{1-\delta}\right) V_1\left(\tilde{\psi}\right) - \left(\frac{\delta}{1-\delta}\right) U_0\left(\tilde{\psi}\right) + \tilde{\psi}_1 - \theta \right) \\
& + \bar{\lambda}_2 \left(\left(\frac{\delta}{1-\delta}\right) V_1\left(\tilde{\psi}\right) - \left(\frac{\delta}{1-\delta}\right) U_0\left(\tilde{\psi}\right) + \tilde{\psi}_2 - \theta \right) \\
& + \bar{\eta}_1 \left(1 - \tilde{\psi}_1\right) + \bar{\eta}_2 \left(1 - \tilde{\psi}_2\right) + \bar{\zeta}_1 \left(\tilde{\psi}_1 - \hat{\pi}\right) + \bar{\zeta}_2 \left(\tilde{\psi}_2 - \hat{\pi}\right)
\end{aligned}
$$

We know that the following first order conditions must hold at the constrained maximum: $\frac{\partial \mathcal{L}}{\partial \tilde{\psi}_1} = 0$, $\frac{\partial \mathcal{L}}{\partial \tilde{\psi}_2} = 0$, $\frac{\partial \mathcal{L}}{\partial \bar{\lambda}_1} \geq 0$, $\bar{\lambda}_1 \geq 0$, $\frac{\partial \mathcal{L}}{\partial \bar{\lambda}_1}\bar{\lambda}_1 = 0$, $\frac{\partial \mathcal{L}}{\partial \bar{\lambda}_2} \geq 0$, $\bar{\lambda}_2 \geq 0$, $\frac{\partial \mathcal{L}}{\partial \bar{\lambda}_2}\bar{\lambda}_2 = 0$, $\frac{\partial \mathcal{L}}{\partial \bar{\eta}_1} \geq 0$, $\bar{\eta}_1 \geq 0$, $\frac{\partial \mathcal{L}}{\partial \bar{\eta}_1}\bar{\eta}_1 = 0$, $\frac{\partial \mathcal{L}}{\partial \bar{\eta}_2} \geq 0$, $\bar{\eta}_2 \geq 0$, $\frac{\partial \mathcal{L}}{\partial \bar{\eta}_2}\bar{\eta}_2 = 0$, $\frac{\partial \mathcal{L}}{\partial \bar{\zeta}_1} \geq 0$, $\bar{\zeta}_1 \geq 0$, $\frac{\partial \mathcal{L}}{\partial \bar{\zeta}_1}\bar{\zeta}_1 = 0$, $\frac{\partial \mathcal{L}}{\partial \bar{\zeta}_2} \geq 0$, $\bar{\zeta}_2 \geq 0$, $\frac{\partial \mathcal{L}}{\partial \bar{\zeta}_2}\bar{\zeta}_2 = 0$.

Substituting in to the Lagrangean from (2.5) and (2.6) and rearranging yields the following:

$$
\begin{aligned}
\mathcal{L} = & - \delta \left(1 + \bar{\lambda}_1 + \bar{\lambda}_2\right) \int_{\tilde{\psi}_1}^{1} (\theta\,\pi - 1)\, g\left(\pi\right) d\pi \\
& - \frac{\delta^2 \left(1 + \bar{\lambda}_1 + \bar{\lambda}_2\right) \int_{\tilde{\psi}_2}^{1} (\theta\,\pi - 1)\, g\left(\pi\right) d\pi}{1-\delta} + \frac{\delta \left(\bar{\lambda}_1 + \bar{\lambda}_2\right) \int_{\tilde{\psi}_2}^{1} (\theta\,\pi - \theta)\, g\left(\pi\right) d\pi}{1-\delta} \\
& + \left(\bar{\zeta}_1 + \bar{\lambda}_1 - \bar{\eta}_1\right) \tilde{\psi}_1 + \left(\bar{\zeta}_2 + \bar{\lambda}_2 - \bar{\eta}_2\right) \tilde{\psi}_2 - \left(\bar{\zeta}_1 + \bar{\zeta}_2\right) \hat{\pi} - \left(\bar{\lambda}_1 + \bar{\lambda}_2\right) \theta + \bar{\eta}_1 + \bar{\eta}_2
\end{aligned}
$$

The relevant derivatives for the first order conditions are:

$$
\frac{d\mathcal{L}}{d\tilde{\psi}_1} = \delta \left(1 + \bar{\lambda}_1 + \bar{\lambda}_2\right) \left(\theta\,\tilde{\psi}_1 - 1\right) g\left(\tilde{\psi}_1\right) + \bar{\zeta}_1 + \bar{\lambda}_1 - \bar{\eta}_1
$$

$$
\frac{d\mathcal{L}}{d\tilde{\psi}_2} = \frac{\delta^2 \left(1 + \bar{\lambda}_1 + \bar{\lambda}_2\right) \left(\theta\,\tilde{\psi}_2 - 1\right) g\left(\tilde{\psi}_2\right)}{1-\delta} - \frac{\delta \left(\bar{\lambda}_1 + \bar{\lambda}_2\right) \left(\theta\,\tilde{\psi}_2 - \theta\right) g\left(\tilde{\psi}_2\right)}{1-\delta} + \bar{\lambda}_2 - \bar{\eta}_2 + \bar{\zeta}_2
$$

$$
\frac{d\mathcal{L}}{d\bar{\lambda}_1} = -\delta \int_{\tilde{\psi}_1}^{1} (\theta\,\pi - 1)\, g\left(\pi\right) d\pi - \frac{\delta^2 \int_{\tilde{\psi}_2}^{1} (\theta\,\pi - 1)\, g\left(\pi\right) d\pi}{1-\delta} + \frac{\delta \int_{\tilde{\psi}_2}^{1} (\theta\,\pi - \theta)\, g\left(\pi\right) d\pi}{1-\delta} + \tilde{\psi}_1 - \theta
$$

$$
\frac{d\mathcal{L}}{d\bar{\lambda}_2} = -\delta \int_{\tilde{\psi}_1}^{1} (\theta\,\pi - 1)\, g\left(\pi\right) d\pi - \frac{\delta^2 \int_{\tilde{\psi}_2}^{1} (\theta\,\pi - 1)\, g\left(\pi\right) d\pi}{1-\delta} + \frac{\delta \int_{\tilde{\psi}_2}^{1} (\theta\,\pi - \theta)\, g\left(\pi\right) d\pi}{1-\delta} + \tilde{\psi}_2 - \theta
$$

In order to prove the result, we must exhaustively consider the various possibilities for the different constraints, and whether or not they bind. We also see that this procedure relates straightforwardly and directly to the taxonomy of quasi-flat paths laid out in section 2.9.1.[50]

Maximal paths

Suppose, first of all, that $0 < \bar{\zeta}_1$ and $0 < \bar{\zeta}_2$. This means that both "lower constraints" on the trigger levels bind, so that $\tilde{\psi}_1 = \hat{\pi}$ and $\tilde{\psi}_2 = \hat{\pi}$. Therefore, the upper constraints cannot possibly bind, and so $\bar{\eta}_1 = 0$ and $\bar{\eta}_2 = 0$. The requirements that $\frac{\partial \mathcal{L}}{\partial \bar{\lambda}_1} \geq 0$ and $\frac{\partial \mathcal{L}}{\partial \bar{\lambda}_2} \geq 0$ therefore both become the following:

$$0 \leq \int_{\hat{\pi}}^{1} \frac{g(\pi)\, \delta\, (1-\theta)}{1-\delta} d\pi - \theta + \hat{\pi} \tag{2.53}$$

This simplifies to give $\theta \leq \delta + (1-\delta)\,\hat{\pi}$. So, only when the inequality condition from Theorem 2.I is either not fulfilled, or just fulfilled with equality[51], can we have a maximal path. Rearranging this condition for δ, we know that the maximal path will be the constrained optimum if and only if:

$$\frac{\theta - \hat{\pi}}{1 - \hat{\pi}} \leq \delta \tag{2.54}$$

If (2.54) is satisfied as a strict inequality, then, since $\frac{\partial \mathcal{L}}{\partial \bar{\lambda}_1} > 0$ and $\frac{\partial \mathcal{L}}{\partial \bar{\lambda}_2} > 0$, we know that $\bar{\lambda}_1 = 0$ and $\bar{\lambda}_2 = 0$. The first order conditions on $\tilde{\psi}_1$ and $\tilde{\psi}_2$ therefore yield the following solution for $\bar{\zeta}_1$ and $\bar{\zeta}_2$.

$$\frac{\partial \mathcal{L}}{\partial \tilde{\psi}_1} = \delta\, (\theta\, \hat{\pi} - 1)\, g(\hat{\pi}) + \bar{\zeta}_1 = 0$$

$$\frac{\partial \mathcal{L}}{\partial \tilde{\psi}_2} = \frac{\delta^2\, (\theta\, \hat{\pi} - 1)\, g(\hat{\pi})}{1-\delta} + \bar{\zeta}_2 = 0$$

$$\bar{\zeta}_1 = \delta\, (1 - \theta\, \hat{\pi})\, g(\hat{\pi}) \qquad \bar{\zeta}_2 = \frac{\delta^2\, (1 - \theta\, \hat{\pi})\, g(\hat{\pi})}{1-\delta}$$

The solutions for $\bar{\zeta}_1$ and $\bar{\zeta}_2$ are positive, which is consistent with our initial assumptions.

If we take the limit of the LHS of condition (2.54), we can see that a maximal path is not possible as $\theta \longrightarrow 1^-$, since, by assumption, $\delta < 1$.

$$\lim_{\theta \to 1^-} \left\{ \frac{\theta - \hat{\pi}}{1 - \hat{\pi}} \right\} = 1^-$$

Suppose instead that $\bar{\zeta}_1 = 0$ and $0 < \bar{\zeta}_2$. This implies that $\tilde{\psi}_2 = \hat{\pi}$ and that $\bar{\eta}_2 = 0$. However, by Lemma 2.V.i, for a quasi-flat path to be optimal, it must be the case that $\tilde{\psi}_1^* \leq \tilde{\psi}_2^*$. Also, it must hold that $\tilde{\psi}_1^* \geq \hat{\pi}$. Therefore $\tilde{\psi}_1^* = \hat{\pi}$, and so this case collapses into the previous one.

[50] We do not need to consider second order conditions since the first order conditions are necessary for an optimal path, and the results of the Lemma follow from these necessary conditions.

[51] If (2.53) is just fulfilled with equality, then the path is also flat (type C), so we are really only interested here in the case where (2.53) is fulfilled as a strict inequality.

Quasi-maximal paths

Suppose that $0 < \bar{\zeta}_1$ and $\bar{\zeta}_2 = 0$. This implies that $\tilde{\psi}_1 = \hat{\pi}$ and $\bar{\eta}_1 = 0$. The requirements that $\frac{\partial \mathcal{L}}{\partial \lambda_1} \geq 0$ and $\frac{\partial \mathcal{L}}{\partial \lambda_2} \geq 0$ become the following:

$$0 \leq -\int_{\tilde{\psi}_2}^{1} \frac{g(\pi)\,\delta\,(-\theta\,\pi + \theta + \delta\,\theta\,\pi - \delta)}{1-\delta} d\pi - \int_{\hat{\pi}}^{1} \delta\,(\theta\,\pi - 1)\,g(\pi)\,d\pi + \hat{\pi} - \theta \tag{2.55}$$

$$0 \leq -\int_{\tilde{\psi}_2}^{1} \frac{g(\pi)\,\delta\,(-\theta\,\pi + \theta + \delta\,\theta\,\pi - \delta)}{1-\delta} d\pi - \int_{\hat{\pi}}^{1} \delta\,(\theta\,\pi - 1)\,g(\pi)\,d\pi + \tilde{\psi}_2 - \theta \tag{2.56}$$

Assuming that $\tilde{\psi}_2 > \hat{\pi}$ (otherwise we would have a maximal path) then, since (2.55) must be fulfilled as a weak inequality, it follows that (2.56) is fulfilled as a strict inequality. Therefore, $\bar{\lambda}_2 = 0$.

Condition (2.55) can be rearranged to give:

$$\int_{\tilde{\psi}_2}^{1} \frac{\delta\,g(\pi)\,(-\theta\,\pi + \theta + \delta\,\theta\,\pi - \delta)}{1-\delta} d\pi + \delta\,\theta\,\bar{\pi} - \delta + \theta \leq \hat{\pi} \tag{2.57}$$

The first order conditions on $\tilde{\psi}_1$ and $\tilde{\psi}_2$ become the following:

$$\frac{\partial \mathcal{L}}{\partial \tilde{\psi}_1} = \delta\,\left(1 + \bar{\lambda}_1\right)(\theta\,\hat{\pi} - 1)\,g(\hat{\pi}) + \bar{\zeta}_1 + \bar{\lambda}_1 = 0$$

$$\frac{\partial \mathcal{L}}{\partial \tilde{\psi}_2} = \frac{\delta^2\left(1 + \bar{\lambda}_1\right)\left(\theta\,\tilde{\psi}_2 - 1\right)g\left(\tilde{\psi}_2\right)}{1-\delta} - \frac{\delta\,\bar{\lambda}_1\left(\theta\,\tilde{\psi}_2 - \theta\right)g\left(\tilde{\psi}_2\right)}{1-\delta} - \bar{\eta}_2 = 0$$

We now need to consider two possible cases, one where $\bar{\lambda}_1 = 0$ and one where $\bar{\lambda}_1 > 0$. Taking the first, the first order conditions on $\tilde{\psi}_1$ and $\tilde{\psi}_2$ become:

$$\delta\,(\theta\,\hat{\pi} - 1)\,g(\hat{\pi}) + \bar{\zeta}_1 = 0$$

$$\frac{\delta^2\left(\theta\,\tilde{\psi}_2 - 1\right)g\left(\tilde{\psi}_2\right)}{1-\delta} - \bar{\eta}_2 = 0$$

The second of these cannot possibly be fulfilled, since the LHS is unambiguously negative. Therefore, it must be the case that $\bar{\lambda}_1 > 0$. The two first order conditions then yield the following solutions for $\bar{\lambda}_1$:

$$\bar{\lambda}_1 = -\frac{\delta\,(\theta\,\hat{\pi} - 1)\,g(\hat{\pi}) + \bar{\zeta}_1}{1 + \delta\,(\theta\,\hat{\pi} - 1)\,g(\hat{\pi})} \tag{2.58}$$

$$\bar{\lambda}_1 = -\frac{\delta^2\left(\theta\,\tilde{\psi}_2 - 1\right)g\left(\tilde{\psi}_2\right) - \bar{\eta}_2(1-\delta)}{\delta\,g\left(\tilde{\psi}_2\right)\left(\left(\theta\,\tilde{\psi}_2 - 1\right)\delta - \theta\,\left(\tilde{\psi}_2 - 1\right)\right)} \tag{2.59}$$

In order for (2.59) to be weakly positive, we must have that:

$$\delta < \frac{\theta - \theta\tilde{\psi}_2}{1 - \theta\,\tilde{\psi}_2} \tag{2.60}$$

Now consider whether $\bar{\eta}_2 = 0$ or $\bar{\eta}_2 > 0$. If $\bar{\eta}_2 > 0$ then $\tilde{\psi}_2 = 1$ and so (2.60) cannot be fulfilled. Hence $\bar{\eta}_2 = 0$. In that case, solving (2.58) and (2.59) simultaneously for $\bar{\zeta}_1$ yields:

$$\bar{\zeta}_1 = \frac{\delta\left(\theta\,(\theta\,\hat{\pi} - 1)\left(\tilde{\psi}_2 - 1\right)g\,(\hat{\pi}) + \left(\theta\,\tilde{\psi}_2 - 1\right)\right)}{\left(\theta\,\tilde{\psi}_2 - 1\right)\delta - \theta\,\left(\tilde{\psi}_2 - 1\right)}$$

Given condition (2.60), the denominator is positive. The numerator will be positive if and only if the following condition holds:

$$g\,(\hat{\pi}) \geq \frac{1 - \theta\,\tilde{\psi}_2}{\theta\,\left(1 - \tilde{\psi}_2\right)(1 - \theta\,\hat{\pi})} \tag{2.61}$$

Note that this is simply condition (2.37) derived via a more intuitive argument in the main text.

Since condition (2.55) must be satisfied with equality, we have that:

$$-\int_{\tilde{\psi}_2}^{1} \frac{\delta\,g\,(\pi)\,(-\theta\,\pi + \theta + \delta\,\theta\,\pi - \delta)}{1 - \delta} d\pi = \delta\,\left(\theta\,\bar{\pi} - 1\right) - \hat{\pi} + \theta$$

Taking limits of both sides as $\theta \longrightarrow 1^-$ yields:

$$-\delta \int_{\tilde{\psi}_2}^{1} (1 - \pi)\,g\,(\pi)\,d\pi = 1 - \hat{\pi} - \delta\,(1 - \bar{\pi})$$

This is not possible, since the LHS is unambiguously negative whilst the RHS is unambiguously positive. Therefore, a quasi-maximal path is not possible as $\theta \longrightarrow 1^-$.

Minimal paths

Now take the case where $0 < \bar{\eta}_1$ and $0 < \bar{\eta}_2$. This implies that $\tilde{\psi}_1 = 1$, $\tilde{\psi}_2 = 1$, $\bar{\zeta}_1 = 0$ and $\bar{\zeta}_2 = 0$. Thus we have a minimal path, where no punishment occurs at all. $\frac{\partial \mathcal{L}}{\partial \bar{\lambda}_1}$ and $\frac{\partial \mathcal{L}}{\partial \bar{\lambda}_2}$ both simplify to give $1 - \theta$. Clearly, therefore, $\bar{\lambda}_1 = 0$ and $\bar{\lambda}_2 = 0$. The first order conditions on $\tilde{\psi}_1$ and $\tilde{\psi}_2$ become:

$$\frac{\partial \mathcal{L}}{\partial \tilde{\psi}_1} = -\delta\,\left(1 - \theta\right) g\,(1) - \bar{\eta}_1 = 0$$

$$\frac{\partial \mathcal{L}}{\partial \tilde{\psi}_2} = -\frac{\delta^2\,(1 - \theta)\,g\,(1)}{1 - \delta} - \bar{\eta}_2 = 0$$

Neither of these can possibly be fulfilled for $\theta < 1$, so a minimal path can never be optimal.

Consider instead the case where $0 < \bar{\eta}_1$ and $\bar{\eta}_2 = 0$. This implies that $\tilde{\psi}_1 = 1$ and $\bar{\zeta}_1 = 0$. However, in that case, by Lemma 2.V.i, $\tilde{\psi}_2 \geq 1$ and so $\tilde{\psi}_2 = 1$ (since it must also hold that $\tilde{\psi}_2 \leq 1$). Therefore this case collapses into the previous one (a minimal path, which can never be optimal).

Quasi-minimal paths

Now we consider the case where $\bar{\eta}_1 = 0$ and $0 < \bar{\eta}_2$. This implies that $\tilde{\psi}_2 = 1$ and $\bar{\zeta}_2 = 0$. The requirements that $\frac{\partial \mathcal{L}}{\partial \lambda_1} \geq 0$ and $\frac{\partial \mathcal{L}}{\partial \lambda_2} \geq 0$ become the following:

$$0 \leq \int_{\tilde{\psi}_1}^{1} -\delta\ (\theta\,\pi - 1)\, g\,(\pi)\, d\pi + \tilde{\psi}_1 - \theta$$

$$0 \leq \int_{\tilde{\psi}_1}^{1} -\delta\ (\theta\,\pi - 1)\, g\,(\pi)\, d\pi + 1 - \theta$$

Assuming $\tilde{\psi}_1 < 1$ (otherwise we would have a minimal path) then, since the first of these is fulfilled as a weak inequality, the second is fulfilled as a strict inequality. Therefore, $\frac{\partial \mathcal{L}}{\partial \lambda_2} > 0$ and so $\bar{\lambda}_2 = 0$. The first order conditions on $\tilde{\psi}_1$ and $\tilde{\psi}_2$ thus become:

$$\frac{\partial \mathcal{L}}{\partial \tilde{\psi}_1} = \delta\ \left(1 + \bar{\lambda}_1\right) \left(\theta\,\tilde{\psi}_1 - 1\right) g\left(\tilde{\psi}_1\right) + \bar{\zeta}_1 + \bar{\lambda}_1 = 0 \tag{2.62}$$

$$\frac{\partial \mathcal{L}}{\partial \tilde{\psi}_2} = -\frac{\delta^2 \left(1 + \bar{\lambda}_1\right)(1 - \theta)\, g\,(1)}{1 - \delta} - \bar{\eta}_2 = 0 \tag{2.63}$$

The second expression is unambiguously negative, so this shows that quasi-minimal paths cannot possibly ever be optimal, regardless of the value of θ.

Carrot-maximized paths

The case where $\bar{\zeta}_1 = 0$, $\bar{\zeta}_2 = 0$, $\bar{\eta}_1 = 0$, $\bar{\eta}_2 = 0$ is the most interesting one, where the carrot-maximized and flat path possibilities can occur. The requirements that $\frac{\partial \mathcal{L}}{\partial \lambda_1} \geq 0$ and $\frac{\partial \mathcal{L}}{\partial \lambda_2} \geq 0$ then become:

$$0 \leq -\int_{\tilde{\psi}_2}^{1} \frac{\delta\, g\,(\pi)\,(-\theta\,\pi + \theta + \delta\,\theta\,\pi - \delta)}{1 - \delta} d\pi + \int_{\tilde{\psi}_1}^{1} -\delta\ (\theta\,\pi - 1)\, g\,(\pi)\, d\pi + \tilde{\psi}_1 - \theta \tag{2.64}$$

$$0 \leq -\int_{\tilde{\psi}_2}^{1} \frac{\delta\, g\,(\pi)\,(-\theta\,\pi + \theta + \delta\,\theta\,\pi - \delta)}{1 - \delta} d\pi + \int_{\tilde{\psi}_1}^{1} -\delta\ (\theta\,\pi - 1)\, g\,(\pi)\, d\pi + \tilde{\psi}_2 - \theta \tag{2.65}$$

By Lemma 2.V.i, we know that an optimal path must have the property that $\tilde{\psi}_1 \leq \tilde{\psi}_2$. There are therefore two cases to consider. First, suppose that $\tilde{\psi}_1 \neq \tilde{\psi}_2$ and so $\tilde{\psi}_1 < \tilde{\psi}_2$. In that case, only (2.64) can bind, and so $\frac{\partial \mathcal{L}}{\partial \lambda_2} > 0$ and $\bar{\lambda}_2 = 0$. The first order conditions on $\tilde{\psi}_1$ and $\tilde{\psi}_2$ then become the following:

$$\frac{\partial \mathcal{L}}{\partial \tilde{\psi}_1} = \left(\delta\ \left(\theta\,\tilde{\psi}_1 - 1\right) g\left(\tilde{\psi}_1\right) + 1\right) \bar{\lambda}_1 + \delta\ \left(\theta\,\tilde{\psi}_1 - 1\right) g\left(\tilde{\psi}_1\right) = 0$$

$$\frac{\partial \mathcal{L}}{\partial \tilde{\psi}_2} = \left(-\,\theta\,\tilde{\psi}_2 + \frac{(\theta - \delta)}{1 - \delta}\right) \delta g\left(\tilde{\psi}_2\right) \bar{\lambda}_1 + \frac{\delta^2}{1 - \delta} \left(\theta\,\tilde{\psi}_2 - 1\right) g\left(\tilde{\psi}_2\right) = 0$$

Clearly, $\bar{\lambda}_1$ must be positive, or the first of these could not possibly be fulfilled. Solving these two equations respectively for $\bar{\lambda}_1$ gives us:

$$\bar{\lambda}_1 = -\frac{\delta\,\left(\theta\,\tilde{\psi}_1 - 1\right)g\left(\tilde{\psi}_1\right)}{\delta\,\left(\theta\,\tilde{\psi}_1 - 1\right)g\left(\tilde{\psi}_1\right) + 1}$$

$$\bar{\lambda}_1 = -\frac{\left(\theta\,\tilde{\psi}_2 - 1\right)\delta}{-\theta\,\left(1-\delta\right)\tilde{\psi}_2 - \delta + \theta}$$

Finally, solving these two equations simultaneously for $g\left(\tilde{\psi}_1\right)$ gives us the following necessary condition for a carrot-maximized path, which, once rearranged to make $\tilde{\psi}_2$ the subject, verifies condition (2.41) derived in the main text in section 2.9.2:

$$g\left(\tilde{\psi}_1\right) = \frac{1-\theta\,\tilde{\psi}_2}{\theta\,\left(1-\tilde{\psi}_2\right)\left(1-\theta\,\tilde{\psi}_1\right)} \tag{2.66}$$

Since $\bar{\lambda}_1 > 0$, condition (2.64) must be satisfied with equality. This gives us:

$$0 = -\int_{\tilde{\psi}_2}^{1}\frac{\delta\,g\left(\pi\right)\left(-\theta\,\pi+\theta+\delta\,\theta\,\pi-\delta\right)}{1-\delta}d\pi - \int_{\tilde{\psi}_1}^{1}\delta\,\left(\theta\,\pi-1\right)g\left(\pi\right)d\pi + \tilde{\psi}_1 - \theta$$

Taking the limit of both sides as $\theta \longrightarrow 1^-$ yields:

$$0 = \delta\int_{\tilde{\psi}_1}^{\tilde{\psi}_2}\left(1-\pi\right)g\left(\pi\right)d\pi - \left(1-\tilde{\psi}_1\right)$$

This condition cannot possibly be fulfilled, so a carrot-maximized path is also impossible as $\theta \longrightarrow 1^-$.

Flat paths

Consider, finally, the case where $\bar{\zeta}_1 = 0$, $\bar{\zeta}_2 = 0$, $\bar{\eta}_1 = 0$, $\bar{\eta}_2 = 0$ and $\tilde{\psi}_1 = \tilde{\psi}_2 = \bar{\psi}$. The requirements that $\frac{\partial\mathcal{L}}{\partial\lambda_1} \geq 0$ and $\frac{\partial\mathcal{L}}{\partial\lambda_2} \geq 0$ then both become:

$$0 \leq \int_{\bar{\psi}}^{1}\frac{\delta\,g\left(\pi\right)\left(1-\theta\right)}{1-\delta}d\pi - \theta + \bar{\psi}$$

This must hold with equality, yielding condition (2.12), which defines the optimal flat path $\bar{\psi}^*$.

The first order conditions on $\tilde{\psi}_1$ and $\tilde{\psi}_2$ become the following:

$$\left(\delta\,\left(\theta\,\bar{\psi}-1\right)g\left(\bar{\psi}\right)+1\right)\bar{\lambda}_1 + \delta\,\left(\theta\,\bar{\psi}-1\right)g\left(\bar{\psi}\right)\left(1+\bar{\lambda}_2\right) = 0$$

$$\frac{\delta\,\left(\delta\,\theta\,\bar{\psi}-\delta-\theta\,\bar{\psi}+\theta\right)g\left(\bar{\psi}\right)\bar{\lambda}_1}{1-\delta} + \left(1+\frac{\delta\,\left(\delta\,\theta\,\bar{\psi}-\delta-\theta\,\bar{\psi}+\theta\right)g\left(\bar{\psi}\right)}{1-\delta}\right)\bar{\lambda}_2 + \frac{\delta^2\,\left(\theta\,\bar{\psi}-1\right)g\left(\bar{\psi}\right)}{1-\delta} = 0$$

Solving these simultaneously for $\bar{\lambda}_1$ and $\bar{\lambda}_2$ yields

$$\bar{\lambda}_1 = \frac{\delta\, g\left(\bar{\psi}\right)\left(\theta\,\delta\,\left(\bar{\psi}-1\right)\left(\theta\,\bar{\psi}-1\right)g\left(\bar{\psi}\right)-\left(1-\delta\right)\left(\theta\,\bar{\psi}-1\right)\right)}{\delta\,\left(\theta-1\right)g\left(\bar{\psi}\right)+1-\delta}$$

$$\bar{\lambda}_2 = -\frac{\left(\theta\,\left(\bar{\psi}-1\right)\left(\theta\,\bar{\psi}-1\right)g\left(\bar{\psi}\right)+\theta\,\bar{\psi}-1\right)\delta^2 g\left(\bar{\psi}\right)}{\delta\,\left(\theta-1\right)g\left(\bar{\psi}\right)+1-\delta}$$

The following two conditions are together sufficient for both of these expressions to be finite and positive.

$$\delta < \frac{1}{1+(1-\theta)g\left(\bar{\psi}\right)} \tag{2.67}$$

$$g\left(\bar{\psi}\right) \leq \frac{1}{\left(1-\bar{\psi}\right)\theta} \tag{2.68}$$

Note now that the RHS of (2.37) from the main text (and (2.66) from this proof) is increasing in $\tilde{\psi}_1$ and $\tilde{\psi}_2$, and is the same as the RHS of (2.61) when $\tilde{\psi}_1 = \hat{\pi}$ and (2.68) when $\tilde{\psi}_1 = \tilde{\psi}_2 = \bar{\psi}$. Hence, condition (2.38) (repeated below as (2.69)) is sufficient for (2.68) to definitely hold, and for (2.61) and (2.66) to never hold. Also, condition (2.69) is sufficient for (2.67) to hold provided that (2.54) does not. (Subtracting the LHS of (2.54) from (2.67), requiring that this be positive and solving for $g(\bar{\psi})$ yields $g(\bar{\psi}) < \frac{1}{\theta-\hat{\pi}}$, which will definitely be fulfilled if (2.69) is.)

$$\forall_\pi : g(\pi) < \frac{1}{\theta\left(1-\hat{\pi}\right)} \tag{2.69}$$

Condition (2.69) is therefore sufficient for the optimal semi-constrained quasi-flat path to be flat (type C), provided that it is not maximal (type A).

Punishment and the Potency of Group Selection

> It is an important fact of human life that individuals can often greatly increase (reward) or decrease (punish) the fitness of others at trivial costs to themselves...Secondary behaviours evolve *more easily* by group selection than primary behaviours because they are less strongly opposed by within-group selection, but they still evolve by group selection. The package of primary and secondary behaviours therefore remains a group-level adaptation.
>
> (Sober & Wilson, 1999)

> The indirect evolutionary approach is based on the assumption that players behave rationally for given preferences but that their preferences change through an evolutionary process...While preferences might be inherited literally in a genetical sense, one could also think of it in terms of social evolution since preferences and value judgements of children are shaped by taking parents or peers as role models.
>
> (Huck & Oechssler, 1999)

3.1 *Overview*

It is a well-established result in evolutionary theory that altruism can in principle be sustained by a process of group selection if a population is split into groups whose members interact disproportionately with one another, provided that there is migration between groups. The level of altruism which can be sustained depends upon the relative strength of the evolutionary forces benefiting the more selfish individuals at the expense of altruists within groups, and that favouring the more altruistic groups over the less altruistic ones. There is an evolutionary "tug-of-war" between individual-level selection and group-level selection (Sober & Wilson, 1999).

This paper embeds a simple model of a punishment system within an indirect cultural evolution framework. The result is that the availability of punishment as a device for social control is shown to drastically reduce the potency of the group selection mechanism, and thus the average equilibrium level of altruism. A normative analysis of the outcome shows that the use of the punishment system can sometimes increase social welfare at the evolutionary equilibrium, by inducing selfish individuals to behave better. However, by weakening the group selection mechanism, it can also under some conditions cause a reduction in social welfare at the equilibrium, by causing less altruism to evolve.

The idea of group selection originates with Darwin, but the contemporary formulation was developed in the twentieth century literature on evolutionary biology, most famously in the work of W. D. Hamilton (Hamilton, 1963) (Hamilton, 1972). The mathematical framework was originally devised by Price (Price, 1970). Still controversial among some biologists (but more widely accepted as a useful practical theory in social science fields), the multilevel selection paradigm has recently been popularised within and beyond the biological field by Sober and Wilson. They have provided a survey article (Sober & Wilson, 1994) and a book-length treatment of the subject (Sober & Wilson, 1999).

Group selection has also become a popular framework in theoretical anthropology, from which the fruitful suggestion that we may see cultural as well as genetic characteristics as evolving through natural selection has been developed and employed (Boyd & Richerson, 1982) (Soltis et al., 1995) (Blackmore, 1999). This idea has a pedigree going back to Darwin in biology, but arguably he was heavily influenced (Hirshleifer, 1977) (Hayek, 1988) by the application of the same principle to social institutions by the philosophers of the Scottish Enlightenment, most famously Adam Smith (Smith, 1976). Economists have also made important contributions to the modern theory of group selection, particularly in clarifying issues regarding the mathematical analysis of the different types of group structure that can enable this phenomenon to arise (Bergstrom, 2002) (Cooper & Wallace, 2004).

There is also an existing literature on the role that punishment, such as in the form of informal sanctions or a legal system, can play as an "altruism amplification device" that allows selfish individuals to be induced to behave more like altruistic ones. It has been shown that, because punishing others is often "cheaper" in terms of cost to onesself than benefiting them, the emergence of the ability to carry out altruistic punishment (which Sober and Wilson refer to as secondary behaviours) can explain how the evolution of primary altruism is made possible in a much wider variety of cases. This hypothesis fits the empirically-observed phenomenon that the ability to punish transgressors in simple experimental games such as the public goods game results in more co-operation being sustained (usually in models where the standard Nash equilibrium with self-interested individuals leads to a complete break-down of co-operation) (Fehr & Gächter, 2000b) (Fehr & Gächter, 2000a) (Fehr & Gächter, 2002a) (Fehr & Fischbacher, 2003) (Fehr & Gächter, 2002b). There are two dimensions to this impact. Firstly, altruistic punishment improves "static" outcomes by making selfish individuals behave better, because they are afraid of being punished. Secondly, the evolution of altruistic punishment can also make it easier for altruism to evolve as a primary behaviour, by reducing the gain in fitness by selfish individuals relative to the altruists in the group (Sober & Wilson, 1999) (Boyd et al., 2003).

This paper aims to make a contribution to the theoretical understanding of the connection between group selection and punishment by applying a third conceptual strand; that of indirect evolution. Most models of the cultural evolution of altruism model cultural norms in a "mechanical" way in the sense that individuals blindly carry out their "programmed" behaviour, whereas economic theory seeks to explain phenomena from a wide variety of cultural scenarios as caused by the same underlying human rationality. The alternative is to assume that it is the weightings that individuals place on the felicity of others that form the evolving phenotype, rather than specific altruistic behaviours directly. In other words, *preferences* evolve but behaviour within the games being played is rational and forward looking, and therefore modelled in the standard manner in which game theory is applied in economic theory.

The indirect evolution approach was first proposed by Güth and Yaari (Güth & Yaari, 1992) following a suggestion originally made by Becker (Becker, 1976). It has been profitably applied to explaining the evolution of preferences for fairness in the ultimatum game (Huck & Oechssler, 1999), in which context it has been shown that "vengeful" individuals who gain utility from reducing the payoffs of others even at cost to themselves can survive and spread in an evolutionary context because, provided the damage they inflict by punishing is high enough relative to the loss to the punisher, they will still have a relatively higher fitness than the "non-vengeful" types.

The application of the indirect evolution methodology to modelling the relationship between punishment and group selection enables an original contribution to be made to an already vast literature. In the standard direct evolution group selection models, there is no distinction between "altruistic preferences" and "altruistic behaviour"; both are modelled as a simple "programmed" phenotype. The presence of altruistic punishers in the population can only lead to more altruistic behaviour in the evolutionary equilibrium by causing the evolved proportion of selfish types to reduce, by weakening the fitness differential between selfish and altruistic individuals (see section 3.2). By contrast, an indirect evolution approach allows an analysis of punishment mechanisms that can alter the static equilibrium in the game being played at each stage of the evolutionary process. Since the use of punishment improves the static outcome for selfish phenotypes by making them behave better towards one another, it can, paradoxically, potentially result in more selfish individuals evolving.

The sequential punishment model which forms the workhorse model in this paper has the property that *only* the outcome for the selfish phenotypes is improved, because the altruistic phenotypes are unwilling to carry out punishment. Thus the result is unambiguous that fewer altruists are able to survive. The normative consequences are, however, ambiguous, because it may be that despite having more selfish individuals, the gain in static efficiency for any given proportion of selfish individuals outweighs this. The balance of normative effects can also, however, go in the opposite direction. Although the result that the evolution of altruism is unambiguously weakened is a strong one, and dependent on the specific form of model used, the phenomenon described is arguably quite general.

The analysis of the sequential punishment model uses standard conceptual tools from economics such as social welfare, subgame-perfect Nash equilibria and utility functions for individuals which are weighted sums of the felicities of other individuals (a common way to model altruism). This enables normative analysis to be carried out in the usual way. The unconventional aspect of the model is the fact that the level of altruism exhibited by the players (the weighting they place upon the consumption of others in their utility function) is, instead of being exogenously determined, endogenously evolving. The key result that punishment reduces the potency of group selection is demonstrated analytically for any population structure, but specific simulations are also carried out for illustrative purposes.

Evolutionary game theory provides the theoretical tools we can use to analyse the selection pressures that will lead to the pattern of behavioural phenotypes at an evolutionarily stable equilibrium. In general, the level of altruism which would evolve would not be expected to be that which is socially optimal, because selection at the individual level creates pressure upon individual phenotypes in the direction of those which are economically non-altruistic. With a population structure split into groups, however, there can be selection pressure at the group-level as well as the individual level. The resulting phenotype pattern then represents a kind of "compromise" between the two. This is the essential conceptual scheme provided by multilevel selection theory (Sober & Wilson, 1999).

3.2 *The Standard Direct Evolution Model*

There are many ways to set up an evolutionary model where the evolution of a phenotype which engages in altruistic punishment makes it easier for primary altruism to evolve. A common set-up is a two-stage prisoners' dilemma, where individuals have a chance to punish their opponent at cost to themselves if they cheat in the first phase or, in the context of a group public goods game, where individuals have a chance, after having observed the contribution of other individuals, to carry out costly punishment on under-contributors. In general these models have the property that altruistic punishers do equally well against each other as against pure altruists, and that, although, in a static sense, they do worse relative to selfish individuals than pure altruists (because they carry out costly punishment), the long-run dynamic evolutionary impact of their presence is to make it much harder for selfish individuals to survive.

The intuition for this result is that the fitness reduction for the selfish types is the dominant effect on relative fitness. It is therefore possible, in a group with a sufficient number of altruistic punishers (which can be maintained by "genetic drift" through mutations), for the selfish types to be rapidly driven out of the population. It has been generally found that a three-phenotype model (selfish types, altruists, altruistic punishers) enables the population to be dominated by a mixture of altruists and altruistic punishers in a much wider variety of cases than the two-phenotype model (Boyd et al., 2003).

An indirect evolution approach, however, can provide a radically different perspective on this issue, because it can recognise the distinction between altruistic preferences and altruistic behaviour. Punishment is not carried out "blindly" but when the evolved preferences of the punisher make it rational for them to carry out the punishment. This means that there is no longer a simple connection between altruistic preferences and altruistic behaviour. More individuals with altruistic preferences in a population will not necessarily lead to more altruistic behaviour, because altruistic individuals who care about others may not be willing to go through with punishment. On the other hand, more altruistic behaviour may occur without an increase in altruistic preferences, because selfish individuals may be incentivized to behave better by the credible threat of punishment.

The results of this approach support the existing view that pure altruism is unlikely to survive evolutionarily. However, punishment is modelled not as a phenotype but as a potential equilibrium in the game being played in an evolutionary context. This means that punishment, although improving static outcomes, may worsen the dynamic outcome by helping more selfish individuals to evolve. Even if punishment is dynamically beneficial, this approach can also help explain why only imperfect forms of altruism appear to evolve in the real world. (See "The Limits to Altruism - A Survey" (Chapter 1) for a summary of the empirical results in this area.)

3.3 *The Sequential Punishment Model*

This paper uses a simplified two-move three-player version of the sequential punishment model which has been analysed extensively it its infinite-player form in "The Socially Optimal Level of Altruism" (Chapter 2). The analysis of the sequential punishment model there shows that there is a complex relationship between the altruism embodied in individual preferences and the social efficiency of the resulting outcomes. Sometimes these interactions can perversely result in too much altruism making it *harder* to support a socially efficient outcome. More frequently, the use of a self-supporting system of punishment means that, beyond a certain level, greater altruism is not necessary, as a socially efficient outcome can already be supported. These results were driven by the combination of the temptation effect (more altruistic individuals are less tempted to do harm to others), the willingness effect (more altruistic individuals are less willing to inflict punishment), and the severity effect (punishments, such as a fine where the revenue is redistributed, are less severe for more altruistic individuals, because they value the contribution of the revenue to the welfare of others).

In the version of the sequential punishment model analysed here, there are three players. Players 1 and 2 each get a opportunity in sequence to inflict harm upon[1] another individual. If they take the opportunity, they gain a benefit $\hat{\pi}$ (where $0 < \hat{\pi} < 1$) and the individual they harm suffers a felicity loss of 1. Player 1 first chooses whether or not to harm player 2. Player 2, observing player 1's action, can either choose not to inflict harm, to harm player 3, or to harm player 1. Player 2 is assumed to be indifferent as to whom they harm.[2] Player 2's ability to focus harm onto player 1 *if* they inflict harm

[1] Throughout the paper, we will use "harm" to refer to the infliction of a negative externality and "punish" to refer to the specific use of such harm opportunities to construct punishment equilibria.

[2] If individuals are indifferent between inflicting harm and not inflicting harm, they are assumed not to inflict harm.

creates the potential for a simple punishment scheme to be used to support a subgame perfect Nash equilibrium where a selfish player 1 is deterred from inflicting harm.

Imagine there is a large population of individuals, who differ in their level of altruism. This is designated by the coefficient of altruism, θ_i, which is the weighting placed on the felicity of other individuals in individual i's utility function. Since the benefit from inflicting harm always takes the value $\hat{\pi}$, we only need two distinct phenotypes, H and L, which correspond to $\hat{\pi} \leq \theta_H < 1$ and $\theta_L < \hat{\pi}$ respectively.[3] Suppose that the proportion of individuals in the population with phenotype H is q, so that $(1 - q)$ have phenotype L.

Each period, individuals are randomly chosen to play the sequential punishment game. Individuals are formed into triplets, where two of the individuals are able to actually make a move whilst a third individual is randomly selected to be player 3. This third individual does not play any role except to act as a passive receptacle for the harm inflicted by player 2 if player 1 co-operates by not inflicting harm. Nature randomly determines, with equal probability, which individuals will receive their harm opportunity first and second. All individuals are assumed to have full knowledge of the coefficient of altruism of the others with whom they interact. We will begin by assuming that the most socially efficient available equilibrium is played, and consider the consequences of dropping it later on.

3.4 *Derivation of Payoff Matrices*

We can think of the sequential punishment game as a sub-game nested within a supergame in which the coefficients of altruism chosen by individuals A and B[4] are a simultaneous move made before the sequential punishment game is played. The choice of the coefficients of altruism by the players then determines the payoffs, and therefore the outcome, of the sequential punishment game nested within.[5] It is, of course, not really appropriate to think of the coefficient of altruism as a strategy chosen, but rather as a phenotype which can be altered via mutations.[6] Also, whereas it is the social utility payoff that determines each player's behaviour in the nested game, it is the felicity payoff that determines

[3] We assume that individuals are always only "partially altruistic".

[4] We refer to the two individuals who are chosen to be players 1 and 2 as A and B before they know who will go first. Player A and player B both have a probability of 0.5 of being in each position.

[5] Additional assumptions about the properties of the equilibrium that will be played in the sequential punishment game are also required to make the outcome determinate. Initially we are assuming that it will be the most socially efficient one, where player 2 harms player 1 only if player 1 harms player 2 (otherwise player 2 either harms player 3, or does not inflict harm).

[6] In the context of cultural selection theory, mutations are not genetic but represent a kind of cultural random drift as behaviours are imperfectly mimicked or new ways of doing things are tried out.

the evolutionary stable equilibrium in the supergame, and the more complex evolutionarily dynamics involved in group selection that we analyse later.

Since there are two possible phenotypes for each individual, there are four possibilities when three individuals meet and interact.[7] Firstly, if both individuals have high altruism[8] then they both behave efficiently by never inflicting harm. Therefore whichever individual goes first, the felicity payoff to each individual is zero. The value of the social welfare function is also zero because no harm is inflicted, and therefore all three individuals get a felicity payoff of 0.[9] This can be seen in the upper payoff matrix in figure 3.1 in which the top left square shows the zero payoffs of individuals 1 and 2, and the resulting zero social welfare in the box in the middle of the square. Similarly, the corresponding square in the lower matrix[10] looks identical, because the payoffs are still zero for each player regardless of who gets to move first.

Suppose instead that player 2 has phenotype L and player 1 has phenotype H. Since there is no future in which they can be punished, player 2 will inefficiently inflict harm. Player 1 will still not inflict harm because he is sufficiently altruistic not to do this in a single-move game anyway. Therefore player 2 will get a felicity payoff of $\hat{\pi}$ and player 1 will get a felicity payoff of 0, because he co-operates and so player 2 follows her default behaviour and harms player 3. Total social welfare is therefore $\hat{\pi} - 1$.

On the other hand, supposing that player 1 has phenotype L and player 2 has phenotype H, player 2 will not inflict harm, and so there is then no credible threat to punish player 1 for inflicting harm, and so player 1 will do so. In this case, player 1 gets a felicity payoff of $\hat{\pi}$ and player 2 gets -1 because she is punished by player 1. Again, social welfare is $\hat{\pi} - 1$.

In the lower matrix, the payoffs for individuals A and B in the bottom left and top right squares are found by averaging the payoffs in the corresponding squares from the first matrix to produce a new symmetric matrix, because players A and B have an equal chance of being player 1 or 2.

[7] The phenotype of the individual selected to be player 3 is unimportant because they do not have any opportunity to act.

[8] Meaning that they both "play" strategy H in the supergame.

[9] The per-period social welfare function sums the felicity of the two individuals who get a harm opportunity along with the felicity of the third individual who acts as a "passive receptacle".

[10] Which shows the felicity payoffs for players A and B, once the chance of being player 1 or 2 has been randomized, and is therefore symmetric.

	1's Phenotype	**H**	**L**
2's Phenotype		$\theta_1 \geq \hat{\pi}$	$\theta_1 < \hat{\pi}$
H	$\theta_2 \geq \hat{\pi}$	0 $\boxed{0}$ 0	$\hat{\pi}$ $\boxed{\hat{\pi} - 1}$ -1
L	$\theta_2 < \hat{\pi}$	0 $\boxed{\hat{\pi} - 1}$ $\hat{\pi}$	0 $\boxed{\hat{\pi} - 1}$ $\hat{\pi}$

	A's Phenotype	**H**	**L**
B's Phenotype		$\theta_A \geq \hat{\pi}$	$\theta_A < \hat{\pi}$
H	$\theta_B \geq \hat{\pi}$	0 $\boxed{0}$ 0	$\underline{\hat{\pi}}$ $\boxed{\hat{\pi} - 1}$ $-\frac{1}{2}$
L	$\theta_B < \hat{\pi}$	$-\frac{1}{2}$ $\boxed{\hat{\pi} - 1}$ $\underline{\hat{\pi}}$	$\underline{\frac{\hat{\pi}}{2}}$ $\boxed{\hat{\pi} - 1}$ $\underline{\frac{\hat{\pi}}{2}}$

Figure 3.1: Sequential-move game

Finally, we have the case where both individuals have phenotype L. Here, player 2 will definitely inflict harm because there is no future. However, this allows a credible threat to be made to player 1 that if he harms socially inefficiently, the harm inflicted by player 2 will be switched from player 3 onto him. If this occurs, player 1 loses social utility of $1 - \theta_1$.[11] However, the gain in social utility he gets by inflicting harm is only $\hat{\pi} - \theta_1$. Player 1 will therefore be effectively deterred from inflicting harm. Player 1's felicity payoff is therefore 0 and player 2's is $\hat{\pi}$. Social welfare will be $\hat{\pi} - 1$. The payoffs for the second matrix are again found by averaging, in order to take into account the equal chance of players A and B being player 1 or 2 in the first matrix.

[11] This is assuming, for simplicity, no discounting. Permitting discounting would be problematic because we would then have to decide whether or not to discount felicity payoffs as well as social utility payoffs. It would also not really add anything insightful to the analysis of a finite-move sequential game.

Here we see an example of the interplay between the willingness effect and the temptation effect. If we examine the social impact of changing person 2 from a low altruism individual to a high altruism individual (with person 1 remaining a low altruism individual), we see that although the temptation effect leads person 2 not to inflict harm when she would have done so before, the willingness effect completely counteracts this by leading person 1 to defect, because he no longer faces the threat of being punished by person 2. The overall impact upon social welfare is therefore neutral.

The best response payoffs in the second matrix are underlined, and the pure strategy Nash equilibrium[12] is for both individuals A and B to have phenotype L. Since each player is always better off in felicity terms by having low altruism, regardless of whether the other individual has high or low altruism, the individual level selection pressure in this simple model leads to a socially inefficient evolutionarily stable equilibrium, in a similar manner to the standard prisoners' dilemma. This comes about because individuals with low altruism receive higher felicity payoffs and therefore reproduce faster than high altruism individuals, thus coming to dominate the population.

Before further analysing the properties of this evolutionary equilibrium, it is instructive to compare it to that of an identical model, except that rather than having a two-move sequential punishment model nested within the supergame, there is instead a game where each individual chooses whether or not to inflict harm in a single-move game simultaneously.[13] So, person A inflicts harm if and only if $\theta_A < \hat{\pi}$ and person B if and only if $\theta_B < \hat{\pi}$. (We continue to assume that person 1 will harm person 2 by default and that person 2 will harm person 3 by default. Thus individuals A and B only take the felicity loss of -1 if they turn out to be person 2, with probability $\frac{1}{2}$.) The payoff matrix for this model is shown in figure 3.2. Although the evolutionarily stable equilibrium is again for all individuals to have phenotype L, the important difference compared to the case where the nested game is sequential is that in the evolutionary equilibrium for this model, both individuals will inflict harm, whereas in the case of the two-move sequential game, although all individuals have low altruism in the evolutionary equilibrium, the individual who has a chance to inflict harm first does not inflict harm, due to the threat of having the harm inflicted by player 2 focused on to him if he defects by inflicting harm. This difference between the two models will turn out to be of crucial importance in determining the nature of their evolutionarily stable equilibria when group selection effects can occur.

[12] Which is also a dominant strategy equilibrium and therefore the unique Nash equilibrium.

[13] When we analyse this model is more detail later on, we will see that in terms of the evolutionary pressures, this model is essentially the same as the standard prisoners' dilemma set-up.

	H $\theta_A \geq \hat{\pi}$	**L** $\theta_A < \hat{\pi}$
H $\theta_B \geq \hat{\pi}$	0 $\boxed{0}$ 0	$\underline{\hat{\pi}}$ $\boxed{\hat{\pi} - 1}$ $-\frac{1}{2}$
L $\theta_B < \hat{\pi}$	$-\frac{1}{2}$ $\boxed{\hat{\pi} - 1}$ $\underline{\hat{\pi}}$	$\underline{\hat{\pi} - \frac{1}{2}}$ $\boxed{2\hat{\pi} - 2}$ $\underline{\hat{\pi} - \frac{1}{2}}$

Figure 3.2: Simultaneous-move game

The relevant difference between the sequential-move and simultaneous-move versions of the model can be brought out if we consider the effect on social welfare of a marginal increase in the proportion of the population with high altruism (phenotype H) from the evolutionarily stable equilibrium in a single homogeneous population. The expected value of the social welfare function, $E(W)$, depends upon the proportion of each phenotype in the population. In the case of the nested sequential-move punishment model, in a finite population of size n, this will be given by:

$$E\left(W\right) = \frac{q\left(nq-1\right)0}{n-1} + 2\,\frac{q\left(1-q\right)n\left(\hat{\pi}-1\right)}{n-1} + \frac{\left(1-q\right)\left(n\left(1-q\right)-1\right)\left(\hat{\pi}-1\right)}{n-1}$$

In the case of the nested simultaneous-move punishment model, this will be:

$$E\left(W\right) = \frac{q\left(nq-1\right)0}{n-1} + 2\,\frac{q\left(1-q\right)n\left(\hat{\pi}-1\right)}{n-1} + \frac{\left(1-q\right)\left(n\left(1-q\right)-1\right)\left(2\,\hat{\pi}-2\right)}{n-1}$$

If we now differentiate these expressions with respect to q, we can find an expression for the gains in social welfare from a marginal increase in the proportion of altruists. For the nested sequential punishment model, we get:

$$\frac{d}{dq}E\left(W\right)\left(q\right) = \frac{\left(1-\hat{\pi}\right)\left(2\,nq-1\right)}{n-1} \tag{3.1}$$

For the simultaneous-move model, we get:

$$\frac{d}{dq}E\left(W\right)\left(q\right) = 2\,\left(1-\hat{\pi}\right) \tag{3.2}$$

As $n \longrightarrow \infty$, (3.1) goes to:

$$\frac{d}{dq}E\left(W\right)\left(q\right) = 2\,\left(1-\hat{\pi}\right)q \tag{3.3}$$

The diagram below shows social welfare as a function of q for both types of nested model, letting $\hat{\pi} = \frac{1}{2}$ and taking the limit as $n \longrightarrow \infty$. We see that at the evolutionary equilibrium where $q = 0$, the marginal increase in social welfare when q increases is positive for the simultaneous-move model but 0 for the sequential-move model. This is because introducing a small number of high altruism individuals into a population of low altruism individuals means that they are almost certain to interact with low altruism individuals. In the nested sequential move game, however, this means that if the new high altruism individual inflicts harm first, they do not change their behaviour, whereas if they go second, although they do not inflict harm, this causes the low altruism individual to defect and inflict harm, whereas they would not do so if the second individual had low altruism instead of high altruism.

So, altruism is only socially beneficial in the sequential punishment model when altruists encounter each other rather than low altruism individuals. In the simultaneous-move game, by contrast, the presence of even a small number of high altruism individuals is socially beneficial because even if they do interact with a low altruism individual, their behaviour is changed because they now do not inflict harm, and this increases social efficiency, even though the low altruism individual they interact with still defects and inflicts harm.

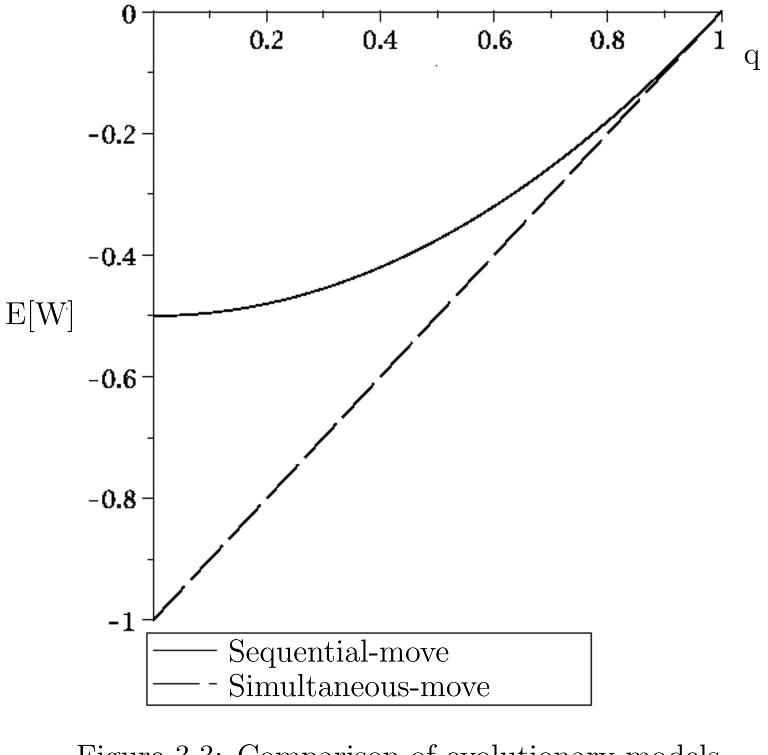

Figure 3.3: Comparison of evolutionary models

What this analysis shows us is that by introducing a change in the way social interactions are modelled so that there are sequential moves enabling conditional punishment, the properties of the evolved equilibrium pattern of individual altruism are altered in the sense that marginal injections of additional altruism into the population are not as socially beneficial. This gives us the intuition for the result that the group selection mechanism is weaker in the sequential-move model.

3.5 *Price Equations*

The conditions under which group-level selection pressures will dominate, and altruism will evolve, can be described using Price equations. Price was an evolutionary biologist and the first person to outline the mathematical conditions required for altruism to evolve by group selection (Price, 1970). It is clear that altruists will always be wiped out in the long run in a single isolated group because the selfish individuals always get better felicity payoffs from their interactions on average, and thus breed more rapidly. However, if altruists are sufficiently concentrated together in sub-groups within the population, whose members interact only (or mainly) with one another, then altruists can get better average payoffs in the population as a whole, and thus outbreed selfish types, because the benefits of altruism will be focused (mainly) upon other altruists. For this to work in the long run, there must be a dispersal mechanism which has the property that it allows altruists to migrate between groups whilst maintaining sufficient inter-group variance of the altruistic phenotype relative to the intra-group variance that altruists are fitter on average than selfish types. The dispersal mechanism determines the manner in which migration of individuals between groups occurs.[14] The Price Equation for a particular model establishes the minimum variance ratio required to enable altruism to survive.

In this section, we will show that the sequential-move game described above leads to a more stringent requirement on the variance ratio achieved by the dispersal mechanism than the simultaneous-move game. This means that altruism will, under general conditions, evolve to a higher degree if the social control mechanism provided by person 2's threat to punish person 1 is removed. Although the presence of this threat prevents person 1 from inflicting harm, and therefore causes a social welfare gain, ceteris paribus, the potential for the use of punishment also weakens the group selection mechanism, and thus the ability to achieve a socially efficient high altruism equilibrium "anarchically". Potentially,

[14] There must always be some migration of altruists in order for altruism to survive, since otherwise they would always be extinguished within each isolated group.

	H	**L**
H	$b-c$ $b-c$	$\underline{b}$ $-c$
L	$-c$ $\underline{b}$	$\underline{0}$ $\underline{0}$

Figure 3.4: Standard prisoners' dilemma

therefore, removing the social control mechanism might improve social welfare by giving a boost to altruism sufficient to cause a net rise in social welfare. After deriving analytic results regarding the Price equation which apply to any dispersal mechanism, we will proceed to show, using some simple specific simulations, that this can indeed be the case.

We will proceed by first showing that the simultaneous-move version of the punishment game has essentially the same Price equation as the standard prisoners' dilemma, which is the classic example of a Price equation in the literature. We will then derive the Price equation for the sequential-move punishment game, and proceed to prove that it always involves a more stringent requirement upon the inter-group variance. Finally, we will illustrate the analysis using a computer simulation of a specific dispersal mechanism.

3.6 *The Standard Prisoners' Dilemma*

The standard model involves a population split into m groups with average size n, so that $n = \frac{1}{m}\sum_{i=1}^{m} n_i$, where n_i is the number of individuals in group i. Individuals in each group only interact with members of their own group, and do so by playing a 2-person prisoners' dilemma game with the payoff matrix shown below (where $b > c$). Now, since the only Nash equilibrium is the dominant strategy equilibrium where both individuals play L (low altruism, equivalent to "defect" in the usual parlance), the only evolutionarily stable equilibrium if all individuals are in a single group is for all individuals to have phenotype L.[15]

[15] If there are random mutations, there may be some type Hs but they are always evolutionarily less fit than the type Ls and so are always in the process of dying out.

When there is more than one group, however, we can show that conditions exist under which the altruistic phenotype can spread through the population. We do this by deriving the change which will occur in the number of high altruism individuals and the total population, thus enabling the derivation of the condition for the proportion of high altruism individuals to increase, assuming that the number of offspring is equal to the felicity payoff. The payoffs for a member of a particular group depend upon the proportion in the group of each type, q_i being the proportion of high altruism types in group i and $q = \frac{1}{mn}\sum_{i=1}^{m} n_i q_i$ being the proportion of altruists in the population.

The payoffs of high and low altruism individuals in group i will be:[16]

$$U_i{}^H = f + \frac{(q_i n_i - 1)\, b}{n_i - 1} - c \tag{3.4}$$

$$U_i{}^L = f + \frac{q_i n_i b}{n_i - 1} \tag{3.5}$$

From the above, we can see that, once interactions have occurred and breeding has taken place, the new proportion of altruists in the overall population after one phase of interactions will be given by the following expression, derived by dividing the new number of high altruism individuals in the population by the new total population:

$$q' = \frac{\sum_{i=1}^{m}\left(\left(f + \frac{(q_i n_i - 1)b}{n_i - 1} - c\right) q_i n_i\right)}{\sum_{i=1}^{m}\left(\left(f + \frac{(q_i n_i - 1)b}{n_i - 1} - c\right) q_i n_i + \left(f + \frac{q_i n_i b}{n_i - 1}\right)(1 - q_i)\, n_i\right)} \tag{3.6}$$

Dividing the numerator and denominator through by n and collecting like terms gives us:

$$q' = \frac{\left(\sum_{i=1}^{m} \frac{q_i n_i f}{n} + \sum_{i=1}^{m} \frac{q_i{}^2 n_i{}^2 b}{n(n_i - 1)} - \sum_{i=1}^{m} \frac{q_i n_i b}{n(n_i - 1)} - \sum_{i=1}^{m} \frac{q_i n_i c}{n}\right)}{\left(\sum_{i=1}^{m} \frac{f n_i}{n} - \sum_{i=1}^{m} \frac{q_i n_i b}{n(n_i - 1)} - \sum_{i=1}^{m} \frac{q_i n_i c}{n} + \sum_{i=1}^{m} \frac{q_i n_i{}^2 b}{n(n_i - 1)}\right)} \tag{3.7}$$

The following expressions are used now to simplify the above expression, and also subsequently for the same purpose later in this section:

$$\begin{gathered}
\sum_{i=1}^{m} \frac{q_i n_i{}^2}{n_i - 1} = m Cov\left(\frac{q_i n_i}{n_i - 1}, n_i\right) + \sum_{i=1}^{m} \frac{q_i n_i}{n_i - 1} n \\
\sum_{i=1}^{m} \frac{q_i{}^2 n_i{}^2}{n_i - 1} = m Cov\left(\frac{q_i n_i}{n_i - 1}, q_i n_i\right) + \sum_{i=1}^{m} \frac{q_i n_i}{n_i - 1} n q \\
\sum_{i=1}^{m} \frac{n_i{}^2}{n_i - 1} = m Cov\left(\frac{n_i}{n_i - 1}, n_i\right) + n \sum_{i=1}^{m} \frac{n_i}{n_i - 1}
\end{gathered} \tag{3.8}$$

$$\begin{gathered}
\sum_{i=1}^{m} \frac{q_i n_i}{n_i - 1} = m E\left(\frac{q_i n_i}{n_i - 1}\right) \qquad \sum_{i=1}^{m} q_i n_i = q n m \\
\sum_{i=1}^{m} n_i = m n \qquad \sum_{i=1}^{m} \frac{n_i}{n_i - 1} = m E\left(\frac{n_i}{n_i - 1}\right)
\end{gathered} \tag{3.9}$$

[16] U_i^H and U_i^L are the expected felicity payoffs in group i. Note also the introduction of a fixed payoff f. This is the same for both phenotypes and thus has no effect on relative fitness, but is needed to ensure that both types always gain a strictly positive payoff. It will be set to a fixed value in the simulations later on.

In order for the proportion of high altruism individuals to grow in the population, we require that $q' - q > 0$, where $q' - q$ can be derived to be the following:

$$q' - q = \frac{\left(-qbCov\left(\frac{q_i n_i}{n_i - 1}, n_i\right) + bCov\left(\frac{q_i n_i}{n_i - 1}, q_i n_i\right) - b\left(1-q\right)E\left(\frac{q_i n_i}{n_i - 1}\right) - qcn\left(1-q\right)\right)}{\left(fn + bCov\left(\frac{q_i n_i}{n_i - 1}, n_i\right) + b\left(n-1\right)E\left(\frac{q_i n_i}{n_i - 1}\right) - cqn\right)} \tag{3.10}$$

Provided f is set high enough to ensure a positive payoff for both phenotypes, the denominator of (3.10) will be positive. Therefore the sign of the numerator will determine whether $q' - q > 0$ is positive or negative. It will therefore be the case that $q' - q > 0$ if and only if the following is fulfilled:

$$\frac{c}{b} < \frac{Cov\left(\frac{q_i n_i}{n_i - 1}, q_i n_i\right)}{qn\left(1-q\right)} - \frac{Cov\left(\frac{q_i n_i}{n_i - 1}, n_i\right)}{n\left(1-q\right)} - \frac{E\left(\frac{q_i n_i}{n_i - 1}\right)}{nq} \tag{3.11}$$

This result can be most easily interpreted in the situation where all groups are of equal size, so that $E\left(\frac{q_i n_i}{n_i - 1}\right) = \frac{qn}{n-1}$, $Cov\left(\frac{q_i n_i}{n_i - 1}, q_i n_i\right) = \frac{n^2}{n-1}Var(q_i)$ and $Cov\left(\frac{q_i n_i}{n_i - 1}, n_i\right) = 0$. In this case, expressions (3.10) and (3.11) simplify respectively to give:

$$q' - q = \frac{\left(n\,Var\left(q_i\right) - q(1-q)\right)b - q\left(n-1\right)\left(1-q\right)c}{\left(n-1\right)\left(f + q(b-c)\right)} \tag{3.12}$$

$$\frac{c}{b} < \frac{n\,Var\left(q_i\right)}{\left(n-1\right)q\left(1-q\right)} - \frac{1}{\left(n-1\right)} \tag{3.13}$$

The intuition for this result is that altruism is able to survive if altruists are sufficiently concentrated together that they have a higher average fitness level than the selfish types. Within a particular group, selfish individuals still do better than altruistic individuals, but across the population, altruists are able to do better than selfish individuals because the altruistic groups spread more rapidly. The $Var(q_i)$ part of the above condition is the inter-group variance of the level of altruism. The $q(1-q)$ part is the intra-group variance: the variance of the random variable formed by taking a single individual from the population and assigning a value of 1 if they have phenotype H and 0 if they have phenotype L. As $n \longrightarrow \infty$, (3.13) simplifies even further to give $\frac{c}{b} < \frac{Var(q_i)}{q(1-q)}$; the variance ratio must be greater than the ratio of the cost of co-operating to the benefit bestowed upon the other individual by doing so. The lower the cost relative to the benefit, the easier it is for altruism to evolve, because the individual-level selection pressure in favour of the selfish types is weakened relative to group-level selection.

If we now take the example of the simultaneous-move punishment game, we can see that a high altruism individual *refraining* from inflicting harm and imposing the cost of 1 on the other individual at benefit $\hat{\pi}$ to herself is logically equivalent to bestowing a benefit of 1 upon the other individual at a cost of $\hat{\pi}$ to herself. The simultaneous-move punishment game therefore has almost the same payoff matrix as the standard prisoners' dilemma, with $b = 1$ and $c = \hat{\pi}$, except that person 2, if she has phenotype L, inflicts harm upon person 3 rather than person 1. We will see when we come to derive the Price equation that this is an insignificant difference. We will also see that the simultaneous-move game can be analysed as an instance of the sequential-move game but with player 2's strategy being not to condition her actions upon those of player 1. In this context, therefore, the standard prisoners' dilemma situation can essentially be viewed as *one* of the possible equilibrium outcomes of the sequential punishment model, and thus as a subcase of this more general model.

3.7 *The Sequential-Move Game*

Before deriving the Price equation for the sequential-move game, we need to further consider the possible subgame-perfect equilibria in this game, and justify why we pay particular attention to certain of these. Player 1's moves are restricted to either inflicting harm upon player 2 or not inflicting harm at all.[17] This means that player 1 has only 2 available strategies. If player 2 has the high altruism phenotype then she only has one credible strategy available, which is not to inflict harm. If player 2 has low altruism then, after the elimination of strictly dominated strategies, she has only 4 possible strategies that could be played in a subgame-perfect Nash equilibrium.[18] These restrictions enable us to fully characterize all of the equilibria of the embedded subgame. The discussion and payoff matrices below describe the subgame-perfect Nash equilibria for the four combinations of phenotype: (H,H), (L,H), (H,L) and (L,L)[19], where L corresponds to $\theta < \hat{\pi}$ and H corresponds to $\theta \geq \hat{\pi}$.

Taking first the two cases when player 2 has high altruism, it is clear that here player 2 will choose not to harm either player 1 or player 3. This is turn means that player 1 will face no future punishment when deciding whether or not to harm player 2, and so will do so if he has low altruism, but not if he has high altruism. Taking instead the two cases where player 2 has low altruism, it is clear that player 2 will choose to inflict harm, but she will be indifferent between inflicting harm upon player 1 and player 3. This makes the strategic possibilities more interesting.

[17] We could justify this by assuming that individuals can only harm those adjacent to them.

[18] This is because player 2 can either harm player 1 or harm player 3 in response to each of player 1's possible moves. She cannot credibly threaten to refrain from inflicting harm, or to harm herself.

[19] (H,L) and (L,H) are distinct because players 1 and 2 do not have symmetric moves or information sets.

	Punish (A)	**Don't Punish (B)**
A → 1 B → 1	$\underline{\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - 1}$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\theta_1\hat{\pi} - 1$ $\underline{\hat{\pi} - \theta_2}$
A → 1 B → 3	$\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - 1$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\underline{\theta_1\hat{\pi} - \theta_1}$ $\underline{\hat{\pi} - \theta_2}$
A → 3 B → 1	$\underline{\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - \theta_1}$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\theta_1\hat{\pi} - 1$ $\underline{\hat{\pi} - \theta_2}$
A → 3 B → 3	$\underline{\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - \theta_1}$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\theta_1\hat{\pi} - \theta_1$ $\underline{\hat{\pi} - \theta_2}$

Figure 3.5: Player 1 has phenotype L and player 2 has phenotype L

The first payoff matrix in figure 3.5 illustrates the case where both individuals have low altruism. The underlined payoffs indicate potential best responses for each player,[20] so that squares with both payoffs underlined represent subgame-perfect Nash equilibria.[21] In three such equilibria, both individuals inflict harm, but there is also one where player 2 (the row player) makes a credible threat to switch harm onto player 1 (the column player) if he chooses to inflict harm, thus resulting in player 1's best response being not to inflict harm. This is a reasonably plausible equilibrium because player 2 can gain by making the threat (if it is credible), but is indifferent as to who she inflicts harm upon and so never incurs a cost from carrying out the threat, thus rendering it credible.

The second payoff matrix, in figure 3.6, illustrates the case where player 1 has phenotype H and player 2 has phenotype L. Provided $\theta_1 > \frac{1+\hat{\pi}}{2}$, all the sub-game perfect Nash equilibria involve player 1

[20] When a player is indifferent between payoffs, all of the equally preferred payoffs are underlined.

[21] Note that although, in general, a best response to a best response is a necessary, but not sufficient condition for a subgame-perfect Nash equilibrium, in this simple game *all* Nash equilibria are subgame-perfect.

	Punish (A)	**Don't Punish (B)**
A → 1 B → 1	$\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - 1$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\underline{\theta_1\hat{\pi} - 1}$ $\underline{\hat{\pi} - \theta_2}$
A → 1 B → 3	$\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - 1$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\underline{\theta_1\hat{\pi} - \theta_1}$ $\underline{\hat{\pi} - \theta_2}$
A → 3 B → 1	$\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - \theta_1$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\underline{\theta_1\hat{\pi} - 1}$ $\underline{\hat{\pi} - \theta_2}$
A → 3 B → 3	$\hat{\pi} - \theta_1 + \theta_1\hat{\pi} - \theta_1$ $\underline{\theta_2\hat{\pi} - 1 + \hat{\pi} - \theta_2}$	$\underline{\theta_1\hat{\pi} - \theta_1}$ $\underline{\hat{\pi} - \theta_2}$

Figure 3.6: Player 1 has phenotype H and player 2 has phenotype L, assuming $\theta_1 > \frac{1+\hat{\pi}}{2}$.

not inflicting harm and player 2 inflicting harm. If $\frac{1+\hat{\pi}}{2} \geq \theta_1 \geq \hat{\pi}$ then there can be a Nash equilibrium where player 2 threatens to harm player 1 *unless* he inflicts harm, and so player 1 does indeed inflict harm even though he loses utility from doing so as a standalone act. However, this equilibrium is not very believable as it involves player 2 making a threat which it is not in her interest to make. She would prefer player 1 not to inflict harm, and so it is not intuitively plausible that she would make such a threat.

Although it is not likely that player 2 would make a threat which influences player 1's behaviour in a manner which is against her own interest, it is more plausible that for some reason player 2 might not be able to effectively make the threat to induce player 1 not to inflict harm when he has low altruism. If player 1 does not believe that player 2 will carry out the threat, then, once it comes to enforcing it, player 2 is indifferent as to whether she does in fact do so, since "bygones are bygones". If the punishment equilibrium breaks down, then the sequential-move game essentially collapses into the simultaneous-move game. This can occur if the subgame-perfect equilibrium that is played involves player 2 playing either the strategy $A \rightarrow 1, B \rightarrow 1$ or the strategy $A \rightarrow 3, B \rightarrow 3$. So, in this sense, the sequential-move move model contains the simultaneous-move model as a special case.

3.8 *Deriving the Price Equations*

We can now derive the expected payoffs of high altruism and low altruism types in the sequential-move game. We assume that each individual has a $\frac{1}{3}$ chance of being player 1, 2 or 3 respectively in each interaction. We will first assume that player 2 plays either strategy $A \rightarrow 1, B \rightarrow 1$ or strategy $A \rightarrow 3, B \rightarrow 3$. In both cases, we will see that the resultant Price equation is basically the same as the standard prisoners' dilemma. We will then move on to the subgame-perfect equilibrium where strategy $A \rightarrow 1, B \rightarrow 3$ is played, and where the resultant Price equation places a more stringent condition upon the variance ratio, thus reducing the potency of the group selection mechanism.

Suppose first of all that a selfish player 2 (a player 2 with a low altruism phenotype, L) chooses to play the strategy $A \rightarrow 1, B \rightarrow 1$, so that she always inflicts harm upon player 1 regardless of whether player 1 inflicts harm upon her or not[22]. A selfish player 1 will therefore definitely choose to harm player 2, because he will be punished anyway and so will optimally wish to take his opportunity to inflict harm for a gain in his social utility. This is a subgame-perfect Nash equilibrium because once player 1 has made his choice, player 2 will be indifferent over whether she inflicts harm upon player 1 or player 3.

3.8.1 Selfish player 2 plays strategy $A \rightarrow 1, B \rightarrow 1$

We derive the Price equation condition by finding the expected felicity payoff of an altruistic individual with phenotype H and a selfish individual with phenotype L. Altruistic individuals have a $\frac{1}{3}$ chance of being player 1, 2 or 3 respectively. If they are player 1, then if player 2 is selfish (with probability $\frac{(1-q_i)n_i}{n_i-1}$), they will receive a felicity payoff of -1. Otherwise they will receive a felicity payoff of 0. If they are player 2, they also will receive a felicity payoff of -1, if player 1 is selfish (probability $\frac{(1-q_i)n_i}{n_i-1}$), and 0 otherwise. If they are player 3, they will always receive a felicity payoff

[22] As we have already seen, since there is no future an altruistic player 2 with phenotype L will never inflict harm.

of 0, because they are never punished. The expected felicity payoff for an altruist will thus be:

$$U_i{}^H = f - \frac{2}{3}\,\frac{(1-q_i)\,n_i}{n_i-1} \tag{3.14}$$

If players are selfish, then if they turn out to be player 1, they will definitely choose to harm player 2, who will harm them in turn if they too are selfish. The expected payoff if they are player 1 would therefore be $\hat{\pi} - \frac{(1-q_i)n_i-1}{n_i-1}$. If they turn out to be player 2, they will again definitely inflict harm, and player 1 will harm them if they are selfish. Again, the expected payoff would be $\hat{\pi} - \frac{(1-q_i)n_i-1}{n_i-1}$. As before, if they turn out to be player 3, their expected payoff is definitely 0. So:

$$U_i{}^L = f + \frac{2}{3}\,\hat{\pi} - \frac{2}{3}\,\frac{(1-q_i)\,n_i - 1}{n_i-1} \tag{3.15}$$

The new proportion of altruists in the population after one stage of interaction will therefore be:

$$q' = \frac{\sum_{i=1}^{m}\left(f - \frac{2}{3}\,\frac{(1-q_i)n_i}{n_i-1}\right) q_i n_i}{\sum_{i=1}^{m}\left(\left(f - \frac{2}{3}\,\frac{(1-q_i)n_i}{n_i-1}\right) q_i n_i + \left(f + \frac{2}{3}\,\hat{\pi} - \frac{2}{3}\,\frac{(1-q_i)n_i-1}{n_i-1}\right)(1-q_i)\,n_i\right)} \tag{3.16}$$

Multiplying out, dividing the numerator and denominator by n and collecting like terms gives us:

$$q' = \frac{\left(\sum_{i=1}^{m} \frac{3}{2}\,\frac{q_i n_i f}{n} - \sum_{i=1}^{m} \frac{q_i n_i{}^2}{n(n_i-1)} + \sum_{i=1}^{m} \frac{q_i{}^2 n_i{}^2}{n(n_i-1)}\right)}{\left(\sum_{i=1}^{m} \frac{3}{2}\,\frac{f n_i}{n} + \sum_{i=1}^{m} \frac{q_i n_i{}^2}{n(n_i-1)} + \sum_{i=1}^{m} \frac{\hat{\pi}\,n_i}{n} - \sum_{i=1}^{m} \frac{\hat{\pi}\,n_i q_i}{n} - \sum_{i=1}^{m} \frac{n_i{}^2}{n(n_i-1)} + \sum_{i=1}^{m} \frac{n_i}{n(n_i-1)} - \sum_{i=1}^{m} \frac{q_i n_i}{n(n_i-1)}\right)} \tag{3.17}$$

We can now apply (3.8) and (3.9) to derive the following expression for the change in the proportion of altruists in the overall population:

$$q' - q = \frac{2\left(-(1+q)\,Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) + Cov\left(\frac{q_i n_i}{n_i-1}, q_i n_i\right) - (n-q)E\left(\frac{q_i n_i}{n_i-1}\right) + q(n-1)E\left(\frac{n_i}{n_i-1}\right) + q\left(Cov\left(\frac{n_i}{n_i-1}, n_i\right) - \hat{\pi}\,n(1-q)\right)\right)}{\left(3\,fn + 2\,Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) + 2(n-1)\left(E\left(\frac{q_i n_i}{n_i-1}\right) - E\left(\frac{n_i}{n_i-1}\right)\right) - 2\,Cov\left(\frac{n_i}{n_i-1}, n_i\right) + 2\,\hat{\pi}\,n(1-q)\right)} \tag{3.18}$$

Provided f is high enough so that both types always get a positive payoff, the denominator of (3.18) will be positive, and $q' - q > 0$ if and only if the following condition is fulfilled:

$$\hat{\pi} < -\frac{(1+q)\,Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right)}{qn(1-q)} + \frac{Cov\left(\frac{q_i n_i}{n_i-1}, q_i n_i\right)}{qn(1-q)} - \frac{(n-q)E\left(\frac{q_i n_i}{n_i-1}\right)}{qn(1-q)} + \frac{(n-1)E\left(\frac{n_i}{n_i-1}\right)}{n(1-q)} + \frac{Cov\left(\frac{n_i}{n_i-1}, n_i\right)}{n(1-q)} \tag{3.19}$$

If all groups are of equal size, the conditions (3.18) and (3.19) become, respectively:

$$q' - q = \frac{2\left(n\,Var\left(q_i\right) - q\left(n-1\right)\left(1-q\right)\hat{\pi} - q(1-q)\right)}{(n-1)\left(3\,f - 2(1-q)(1-\hat{\pi})\right)} \tag{3.20}$$

$$\hat{\pi} < \frac{n\,Var\left(q_i\right)}{q\left(n-1\right)\left(1-q\right)} - \frac{1}{(n-1)} \tag{3.21}$$

If all groups are the same size, the condition on the variance ratio is therefore identical to the prisoners' dilemma, since (3.21) is identical to (3.13) with $\frac{c}{b} = \hat{\pi}$.[23]

[23] There is a slight difference between the two models when groups are of different sizes, due to the differing position of player 3 in different sized groups.

3.8.2 Selfish player 2 plays strategy $A \to 3, B \to 3$

This case will be very similar to the previous one, except that it is player 3 rather than player 1 who always receives the harm inflicted by a selfish player 2. The expected utility payoffs will therefore be the same as above, as will the Price equation.

3.8.3 Selfish player 2 plays strategy $A \to 1, B \to 3$

We will now assume that player 2 always plays strategy $A \to 1, B \to 3$, so that player 1 is always induced not to inflict harm if player 2 is of type L. Our first step will be to derive the expected felicity payoff for individuals with high and low altruism.

Taking first the expected felicity payoff of an altruistic individual, they have a $\frac{1}{3}$ chance of being player 1 in their interaction. In this case, whether or not player 2 is altruistic, the individual will not inflict harm, and so will receive a payoff of 0. If, on the other hand, they turn out to be player 2 (probability $\frac{1}{3}$), they will be punished by player 1 if player 1 is selfish (probability $\frac{(1-q_i)n_i}{n_i-1}$ and suffering a loss of 1), because they can make no credible threat to punish a selfish player 1 for doing this. The third possibility is that they will be player 3, in which case they will be punished by player 2 if player 2 turns out to be selfish (probability $\frac{(1-q_i)n_i}{n_i-1}$ and suffering a loss of 1). This is because even if player 1 turns out to be selfish, he will never choose to harm player 2 due to his fear of being punished by having the harm inflicted by player 2 focused onto him. Hence, a selfish player 2 will always harm player 3. So, the expected felicity payoff of an altruistic individual in this model is:

$$U_i{}^H = f - \frac{2}{3}\frac{(1-q_i)\,n_i}{n_i-1} \tag{3.22}$$

Now we take the case of selfish individuals. If they turn out to be player 1, they will choose to harm player 2 if and only if player 2 is altruistic (the unconditional probability of this scenario is $\frac{1}{3}\frac{q_i n_i}{n_i-1}$ and the felicity payoff would be $\hat{\pi}$).[24] If selfish individuals turn out to be player 2 (probability $\frac{1}{3}$), then they will definitely harm either player 1 or player 3, gaining a felicity payoff of $\hat{\pi}$. If they turn out to be player 3, they are in the same situation as they would be if they were altruistic, except that the probability that player 2 is selfish and inflicts harm upon them is now $\frac{(1-q_i)n_i-1}{n_i-1}$. The expected utility payoff of an altruistic individual will therefore be:

$$U_i{}^L = f + \frac{1}{3}\frac{\hat{\pi}\,q_i n_i}{n_i-1} + \frac{1}{3}\hat{\pi} - \frac{1}{3}\frac{(1-q_i)\,n_i-1}{n_i-1} \tag{3.23}$$

[24] The benefit received by inflicting harm inefficiently is equivalent to the cost that must be incurred in order to behave efficiently in the prisoners' dilemma model.

From the above, we can see that, once interactions have occurred and breeding has taken place, the new proportion of high altruism individuals in the population will be given by:

$$q' = \frac{\sum_{i=1}^{m}\left(f - \frac{2}{3}\,\frac{(1-q_i)n_i}{n_i-1}\right) q_i n_i}{\sum_{i=1}^{m}\left(\left(f - \frac{2}{3}\,\frac{(1-q_i)n_i}{n_i-1}\right) q_i n_i + \left(f + \frac{1}{3}\,\frac{\hat{\pi}\, n_i q_i}{n_i-1} + \frac{1}{3}\,\hat{\pi} - \frac{1}{3}\,\frac{(1-q_i)n_i-1}{n_i-1}\right)(1-q_i)\, n_i\right)} \tag{3.24}$$

Multiplying out, dividing the numerator and denominator by n and collecting like terms yields:

$$q' = \frac{\left(\sum_{i=1}^{m} \frac{3}{2}\,\frac{q_i n_i f}{n} - \sum_{i=1}^{m} \frac{q_i {n_i}^2}{n(n_i-1)} + \sum_{i=1}^{m} \frac{{q_i}^2 {n_i}^2}{n(n_i-1)}\right)}{\frac{1}{2}\left(3\sum_{i=1}^{m} \frac{f n_i}{n} + \sum_{i=1}^{m} \frac{{q_i}^2 {n_i}^2 (1-\hat{\pi})}{n(n_i-1)} + \sum_{i=1}^{m} \frac{\hat{\pi}\, q_i {n_i}^2}{n(n_i-1)} + \sum_{i=1}^{m} \frac{\hat{\pi}\, n_i}{n} - \sum_{i=1}^{m} \frac{\hat{\pi}\, n_i q_i}{n} - \sum_{i=1}^{m} \frac{{n_i}^2}{n(n_i-1)} + \sum_{i=1}^{m} \frac{n_i}{n(n_i-1)} - \sum_{i=1}^{m} \frac{q_i n_i}{n(n_i-1)}\right)} \tag{3.25}$$

Expressions (3.8) and (3.9) can now be applied to derive the following expression for the change in the proportion of altruists in the overall population:

$$q' - q = \frac{(2-q(1-\hat{\pi}))\, Cov\left(\frac{q_i n_i}{n_i-1}, q_i n_i\right) - (2+\hat{\pi}\, q)\, Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) - \left(\left((1-q)(q\hat{\pi}+2)+q^2\right)n - q\right) E\left(\frac{q_i n_i}{n_i-1}\right) + (n-1)qE\left(\frac{n_i}{n_i-1}\right) - q(1-q)\hat{\pi}\, n + q\, Cov\left(\frac{n_i}{n_i-1}, n_i\right)}{\left(3\, fn + \hat{\pi}\, Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) + (1-\hat{\pi})\, Cov\left(\frac{q_i n_i}{n_i-1}, q_i n_i\right) - (1-((1-q)\hat{\pi}+q)n)\, E\left(\frac{q_i n_i}{n_i-1}\right) - (n-1)E\left(\frac{n_i}{n_i-1}\right) + (1-q)\hat{\pi}\, n - Cov\left(\frac{n_i}{n_i-1}, n_i\right)\right)} \tag{3.26}$$

Assuming f is high enough to make the denominator of the RHS of (3.26) positive, $q' - q$ will be positive if and only if the following condition holds:

$$\hat{\pi} < \frac{(2-q)\, Cov\left(\frac{q_i n_i}{n_i-1}, q_i n_i\right) - 2\, Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) - \left(n(1-q)^2 + n - q\right) E\left(\frac{q_i n_i}{n_i-1}\right) + q(n-1)E\left(\frac{n_i}{n_i-1}\right) + q\, Cov\left(\frac{n_i}{n_i-1}, n_i\right)}{q\left(Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) - Cov\left(\frac{q_i n_i}{n_i-1}, q_i n_i\right) + n(1-q)\left(E\left(\frac{q_i n_i}{n_i-1}\right) + 1\right)\right)} \tag{3.27}$$

When all groups are of equal size, the relevant conditions become:

$$q' - q = \frac{(qn\hat{\pi} + n(2-q))\ Var\,(q_i) - q\,(1-q)\,(n(q+1)-1)\,\hat{\pi} - q\,(1-q)\,(n+1-nq)}{3\,(n-1)\,f + n\,(1-\hat{\pi})\ Var\,(q_i) - (1-q)\,(nq+n-1)\,(1-\hat{\pi})} \tag{3.28}$$

$$\hat{\pi} < \frac{q\,(1-q)\,(n+1-nq) - n\,(2-q)\ Var\,(q_i)}{q\,(\,Var\,(q_i)\, n - (1-q)\,(nq+n-1))} \tag{3.29}$$

Condition (3.27) will be shown in Theorem 3.I to be more stringent than the equivalent condition (3.19) for the simultaneous-move game. This means that it is unambiguously more difficult for group selection to operate in the sequential-move model. The proof works by making a number of simplifying substitutions so that the RHS of (3.27) can be directly compared to the RHS of (3.19). It can then be shown that, if the RHS of (3.27) is positive then the RHS of (3.19) must also be positive and strictly greater. In other words, if group selection can survive given some values of $\hat{\pi}$ in the sequential-move model, it must be able to survive for a wider range of $\hat{\pi}$ values in the simultaneous-move model.

Theorem 3.I

If $\hat{\pi}'_{seq} > 0$, then $\hat{\pi}'_{sim} > \hat{\pi}'_{seq}$.

Proof: Let $\hat{\pi}'_{sim}$ be the RHS of (3.19) and $\hat{\pi}'_{seq}$ be the RHS of (3.27). The following substitutions can be used to rewrite (3.19) and (3.27) in a more easily comparable form:

$$\alpha_i = Cov\left(\frac{n_i}{n_i-1}, n_i\right) + nE\left(\frac{n_i}{n_i-1}\right) - Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) - nE\left(\frac{q_i n_i}{n_i-1}\right) \tag{3.30}$$

$$\beta_i = E\left(\frac{n_i}{n_i-1}\right) - E\left(\frac{q_i n_i}{n_i-1}\right) \tag{3.31}$$

$$\gamma_i = Cov\left(\frac{q_i n_i}{n_i-1}, n_i\right) + nE\left(\frac{q_i n_i}{n_i-1}\right) - Cov\left(\frac{q_i n_i}{n_i-1}, q_i n_i\right) - qnE\left(\frac{q_i n_i}{n_i-1}\right) \tag{3.32}$$

Note that $\alpha_i > \beta_i > 0$ and that $\alpha_i > \gamma_i > 0$. Using these substitutions, (3.19) and (3.27) become:

$$\hat{\pi}'_{sim} = \frac{(\alpha_i - \beta_i)\, q - \gamma_i}{qn\,(1-q)} \tag{3.33}$$

$$\hat{\pi}'_{seq} = \frac{(\alpha_i - \beta_i)\, q - (2-q)\,\gamma_i}{(\gamma_i + n\,(1-q))\, q} \tag{3.34}$$

It can now be seen clearly by observation that if $\hat{\pi}'_{seq} > 0$, then $\hat{\pi}'_{sim} > \hat{\pi}'_{seq}$.

3.9 Illustration - "Haystacks" Model

The "Haystacks" model is a well-known scenario in evolutionary biology which provides one possible population structure which could produce sufficient inter-group variance to allow altruism to survive via group selection. The original analogy was a population of mice which splits into haystacks, following which interaction takes place entirely within each haystack for a period of time. Periodically, however, the hay is taken away and the mice are forced out, so that the meta-population once again intermingles. If altruistic mice are sufficiently concentrated within the haystacks, they will breed more rapidly on average than selfish mice.

We are not seeking here to demonstrate the conditions under which such a population structure will enable group selection to take place. There has been controversy about the nature of the mechanism by which the haystacks are formed required to enable group selection to operate effectively. It has been shown that there must either be assortative group formation or more than one period of isolation in order for altruism to survive (Bergstrom, 2002). A thorough investigation has now been undertaken

(Cooper & Wallace, 2004). Cooper and Wallace have shown that altruism can indeed evolve by group selection, even with finite groups and when the assortment mechanism is completely random, provided the ratio of benefit to cost is high enough and the number of periods of isolation is within an intermediate “Goldilocks” band.

We will use different lengths of the isolation period to illustrate the phenomenon that the sequential-move model makes it more difficult for altruism to survive by, as we have already seen analytically, making the Price equation condition more stringent. The haystacks structure acts as a kind of amplification device for the inter-group variance. The longer groups are isolated, the longer the altruists have to benefit one another. However, counteracting this is the fact that the longer the groups are isolated, the better the selfish individuals are doing at the expense of the altruists within each haystack. This means that there is an optimal amount of time for the haystacks isolation, in terms of maximizing the success of the altruists. Isolating the haystacks for a longer or shorter period than this results in a lower average proportion of altruists evolving.

The haystacks model can only operate if there is some initial inter-group variance to be amplified. This is usually achieved by introducing randomness into the assortment process by which groups are formed. In the simulations that follow, we assume that individuals in the overall population are sorted into groups of size 6. We use the hypergeometric distribution to approximate this process. Simulations are “smoothed”, in the sense that the fraction of altruists is treated as a continuous variable, even though the Price equations are based on finite group sizes. This has been found to deliver a reasonably close approximation to the discrete model, and allows for much faster simulations, and thus better quality data (Cooper & Wallace, 2004). The haystacks idea is not here being used primarily to justify the possibility of group selection working, but to provide a variable, in the number of periods of isolation g, that can be used to adjust the inter-group variance and illustrate the differences between the sequential-move model and the simultaneous-move model.

Group selection can, of course, occur by other methods aside from a Haystacks population structure. An example commonly used in the social sciences is assortative interaction, where altruists are able to disproportionately interact with one another by forming groups and excluding selfish types. It is commonly argued that group selection is likely to be stronger in human cultural evolution than in human biological evolution not only because cultural phenotypes can be transmitted more rapidly (e.g.

by imitation) than biological ones (which must be passed on genetically via biological reproduction) but also because the concentration of particular cultural traits in groups of humans does not have to rely on randomness as it does in biological models of group selection. For example, groups can expel or reject interaction with non-altruists, or bring extra pressures to bear to enforce conformity. One empirical study (Soltis et al., 1995) found sufficient empirical evidence from anthropological studies of group formation and interaction to conclude that cultural group selection may occur in this manner over a long time scale in human society.

The simulations proceed as follows. The population begins at size 100, with $\frac{1}{3}$ of the individuals having high altruism. This is split into groups of 6, approximated by a hypergeometric distribution for the proportion of groups with each different possible composition, then multiplied by $\frac{100}{6}$ to give the number of each type of group.[25] Each group then evolves in isolation for g periods. The members of the group are formed each period into triplets to play the sequential punishment game, assuming that its simultaneous-move and sequential-move equilibria are played respectively for the simulations with and without the use of punishment. At the end of each period, mutations occur where a fraction ϵ of each type change into the other type. The new population composition is then generated as a weighted average of the different group types after g periods. The number of individuals in the overall population is then normalised back to 100, but preserving the new proportion of altruists q'. (This is done in order to prevent the population from exploding to infinity, to aid the running of the simulations.)

Figures 3.7 through 3.11 overleaf illustrate the outcomes from the simulation over 500 generations, with $\pi = 0.075$ and $\epsilon = \frac{1}{500}$ (time being measured along the x-axis), in the two models, with isolation times of $g = 3$, $g = 6$, $g = 52$, $g = 117$ and $g = 204$ periods respectively.[26] The black line illustrates the simultaneous-move model and the grey line the sequential-move model. The top graph shows the proportion of altruists, and the bottom graph shows the average value of the social welfare function.[27]

In figure 3.7, $g = 3$ is not high enough for altruism to survive in the long run in either model. Since the sequential-move model results in a socially superior static outcome, social welfare can be seen to be higher with the sequential move model (the grey line in the bottom diagram) than with the simultaneous-move model (the black line in the bottom diagram).

[25] Recall that the simulations are smoothed, so there is no reason why the population size need be an integer multiple of the group size.

[26] The simulations were written in Ox. Source code is provided in the appendix.

[27] Calculated as the expected felicity payoff for each individual, and normalized so that both individuals choosing to inflict harm results in a felicity of 0.

In figure 3.8, $g = 6$ is high enough for altruism to survive in the simultaneous move model, but not the sequential-move model. Once sufficient time has elapsed for the proportion of altruists in the population to become high enough, social welfare in the simultaneous model ends up higher than social welfare in the sequential model.

Figure 3.9 shows a situation where $g = 52$ is in the range necessary for altruism to survive in both types of model. However, the analytic result from Theorem 3.I still results in the proportion of altruists oscillating around a lower average in the sequential model. This can be seen to lead to lower average social welfare in the sequential model.

Figure 3.10 shows a situation where $g = 117$ is too high for altruism to survive in the sequential model. It is still able to survive in the simultaneous model, where average social welfare is higher. In figure 3.11, however, $g = 204$ is sufficiently high that although altruism still survives in the simultaneous model, it oscillates so much that social welfare is actually higher on average in the sequential model.

Figures 3.12 and 3.13 show the average values of the proportion of altruists, q and the social welfare function over 5000 generations for different values of g on the x-axis, given two different possible values for $\hat{\pi}$. The dashed grey line in the top diagram in each figure shows the initial proportion of altruists $q = \frac{1}{3}$. The important features to note are, firstly, that there is a wider range where altruism survives in the simultaneous model than in the sequential model and, secondly, that there is a range of values of g where altruism survives in the simultaneous model and not in the sequential model, but average social welfare is nonetheless higher in the sequential model. This shows that the normative consequences of the weakening of group selection are ambiguous, despite the unambiguous result that it is harder for altruism to evolve in the sequential model.

Figures 3.14 and 3.15 show, for two different values of ϵ, the region in (g, π) space where the simultaneous model results in a higher average value of the social welfare function in black and the region where the sequential model results in a higher average value in white. This shows most clearly the result that the use of the social control mechanism of person 2's ability to conditionally punish person 1 is a mixed blessing. In some circumstances, in which the group selection mechanism would have been too weak to operate, it will improve the social efficiency of the evolutionary outcome. However, in other circumstances, where the group selection mechanism would have been strong enough (corresponding to the black area of the diagram), the use of conditional punishment can weaken the group selection mechanism and thus, by reducing the number of altruists in the evolutionary equilibrium, actually result in society being worse off than in the "anarchic" equilibrium, where punishment is not used.

3.10 *Conclusion*

This paper has shown that the use of punishment is a "double-edged sword" for the evolution of altruism in that it may help selfish types to do better evolutionarily, because they are more willing to make use of opportunities to harm others at benefit to themselves. In the traditional literature on altruistic punishment, altruistic punishers are modelled as a specific behavioural phenotype. This is unsatisfactory from the viewpoint of economic theory because it begs the question of how different available punishment technologies might interact with such evolved preferences. The indirect evolution methodology provides a useful way to approach this question, because it allows a separation between altruistic preferences and altruistic behaviour. The central message is that the ability of humans to punish one another may weaken the selection pressure in favour of altruistic preferences, with potentially negative dynamic welfare implications.

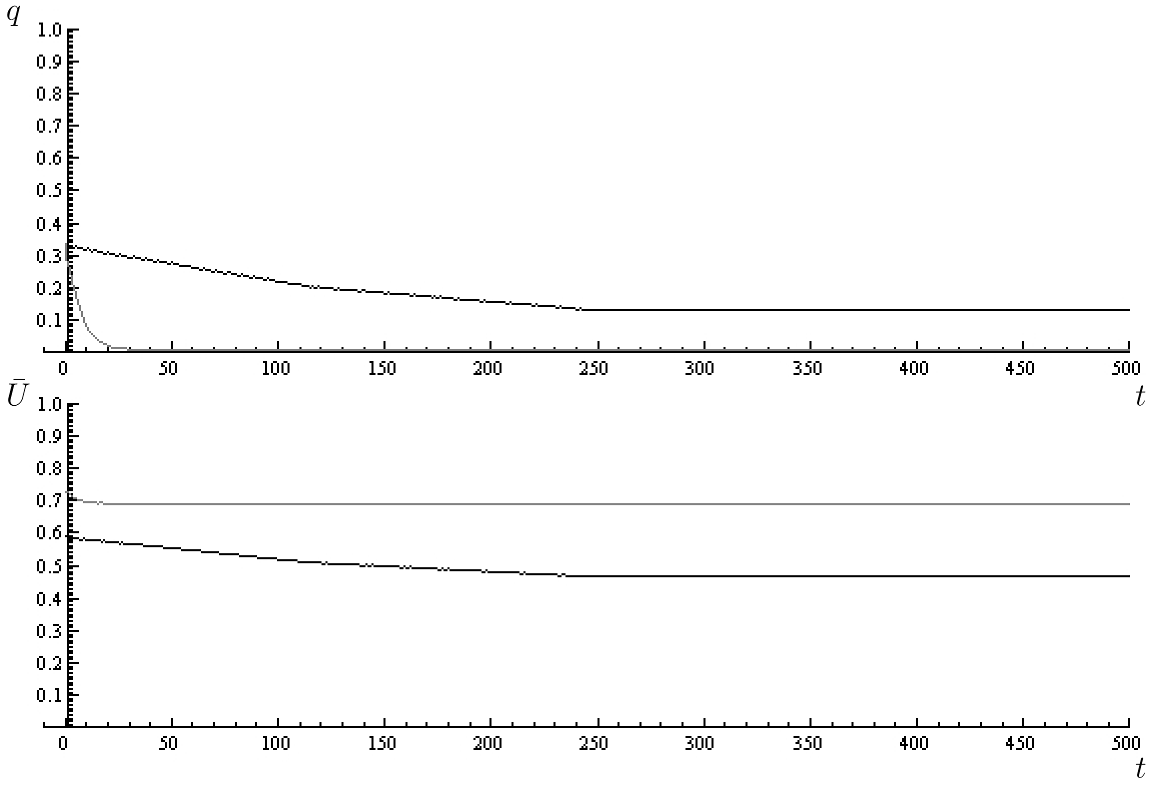

Figure 3.7: 3 periods of isolation, $\pi = 0.075$

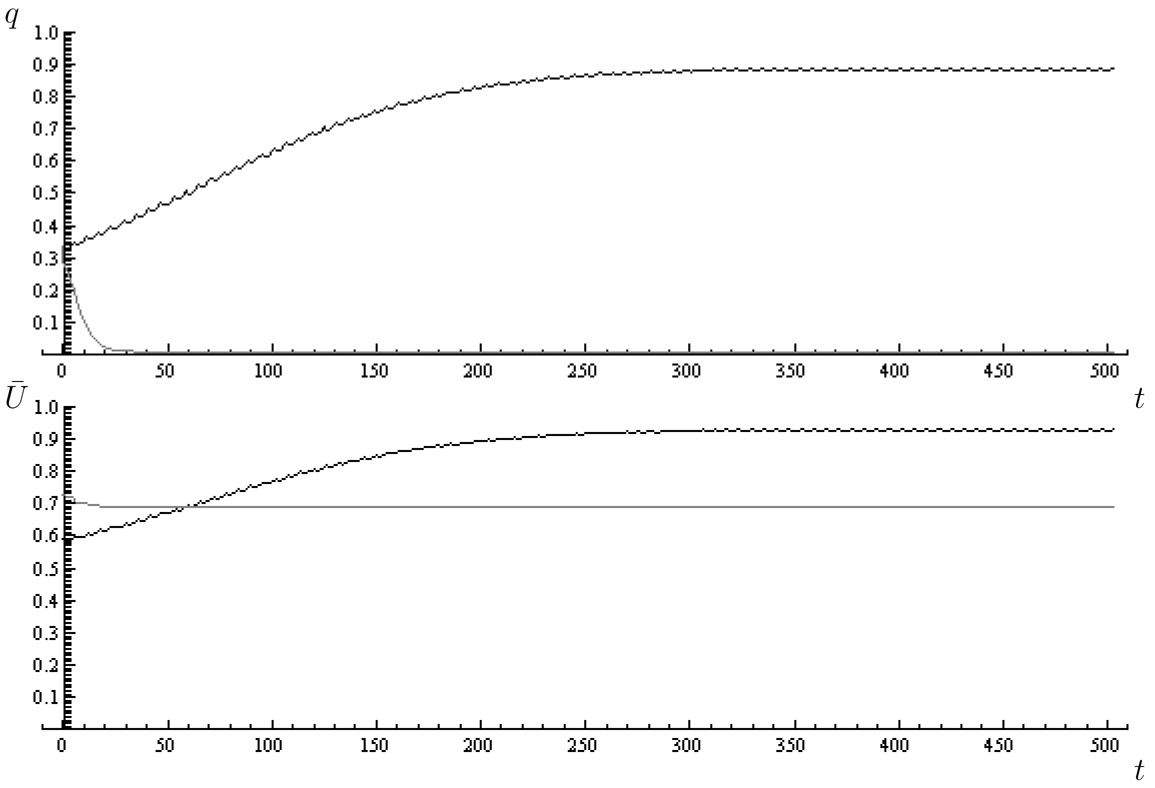

Figure 3.8: 6 periods of isolation, $\pi = 0.075$

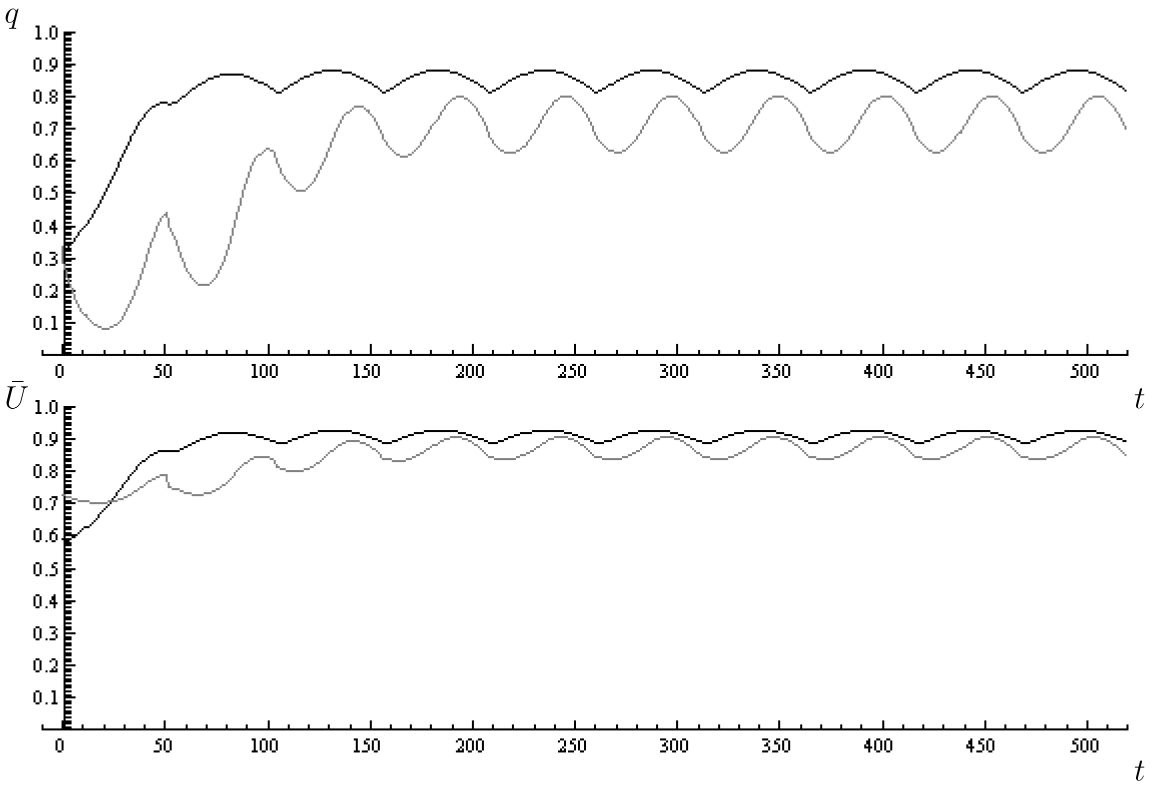

Figure 3.9: 52 periods of isolation, $\pi = 0.075$

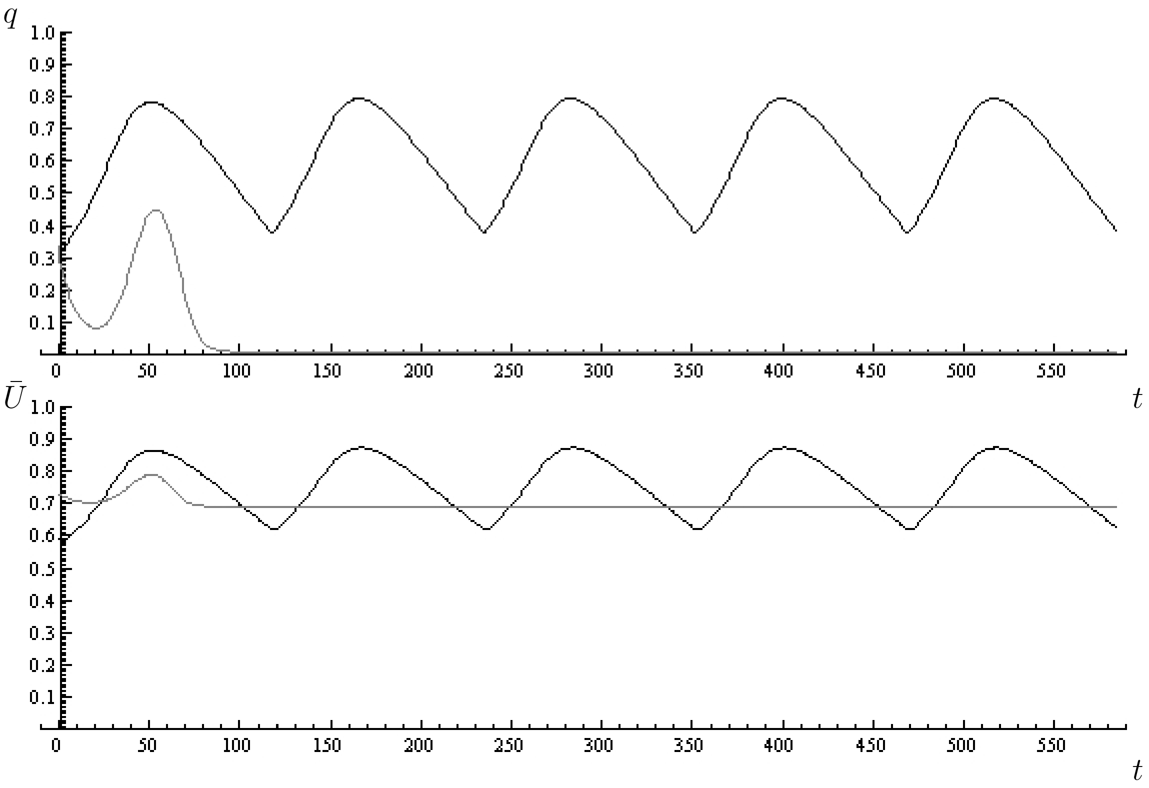

Figure 3.10: 117 periods of isolation, $\pi = 0.075$

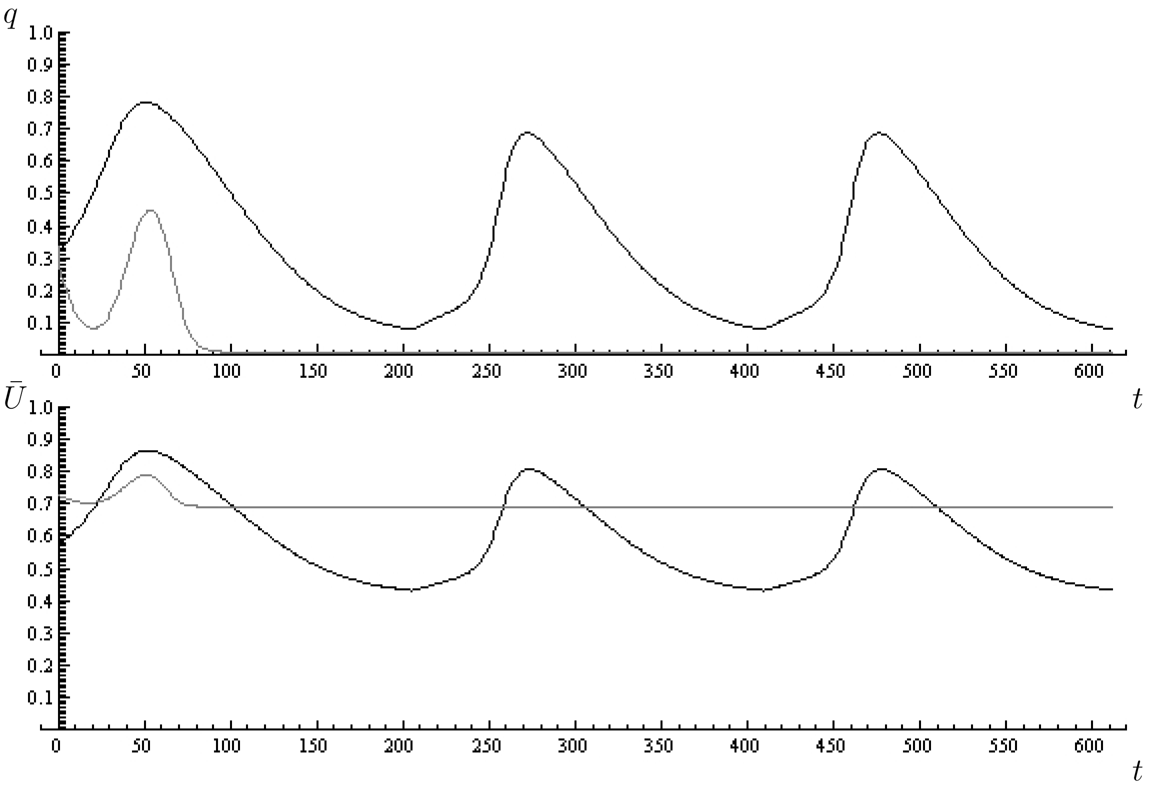

Figure 3.11: 204 periods of isolation, $\pi = 0.075$

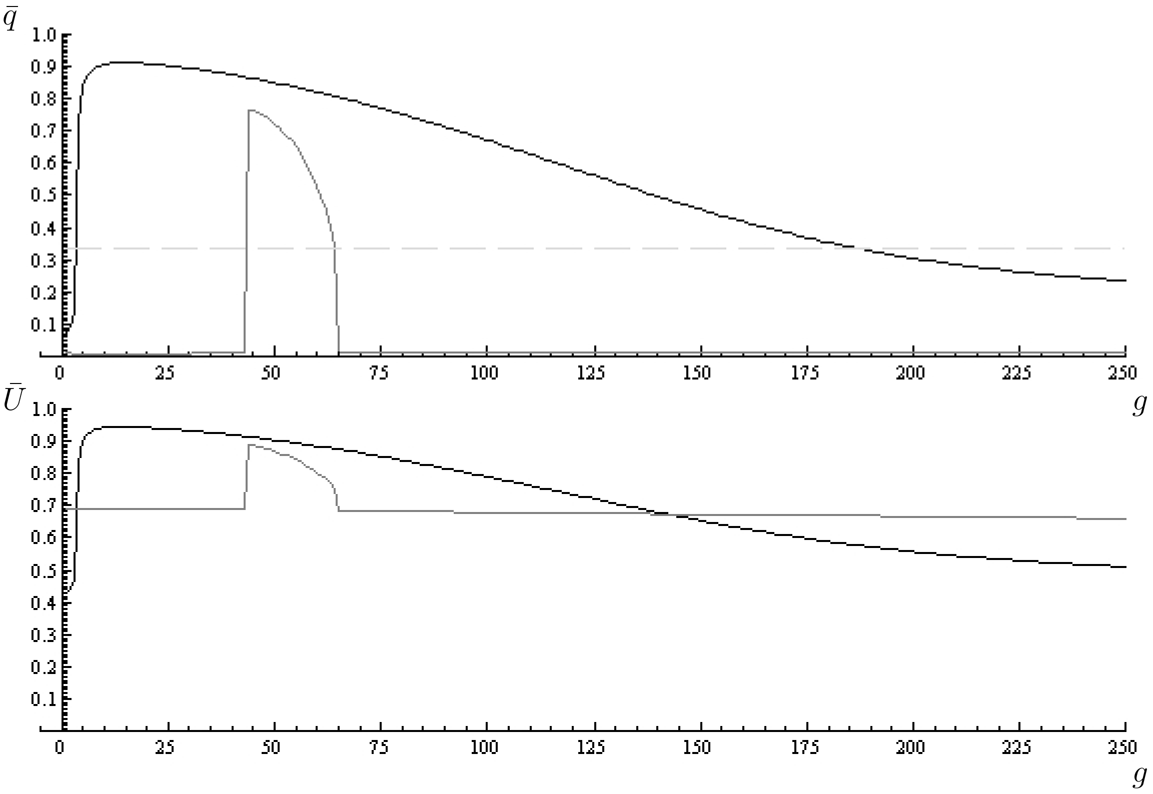

Figure 3.12: Average level of altruism and social welfare, $\pi = 0.075$

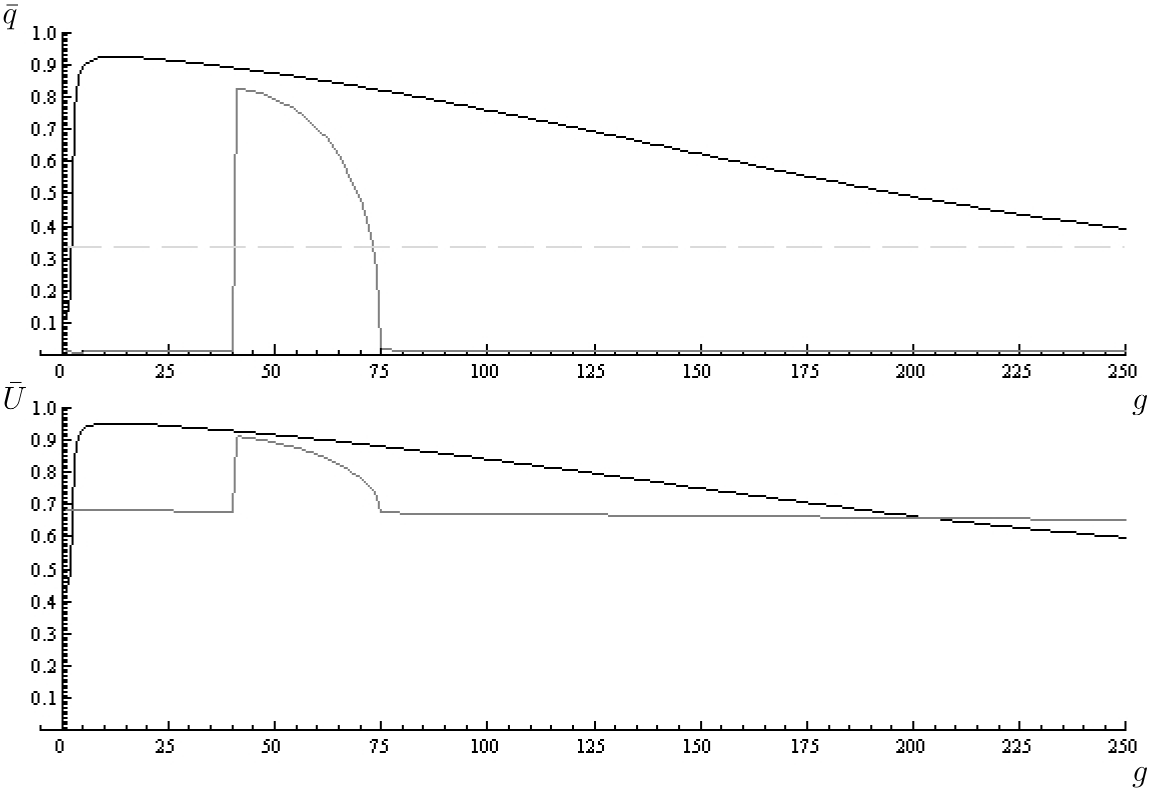

Figure 3.13: Average level of altruism and social welfare, $\pi = 0.05$

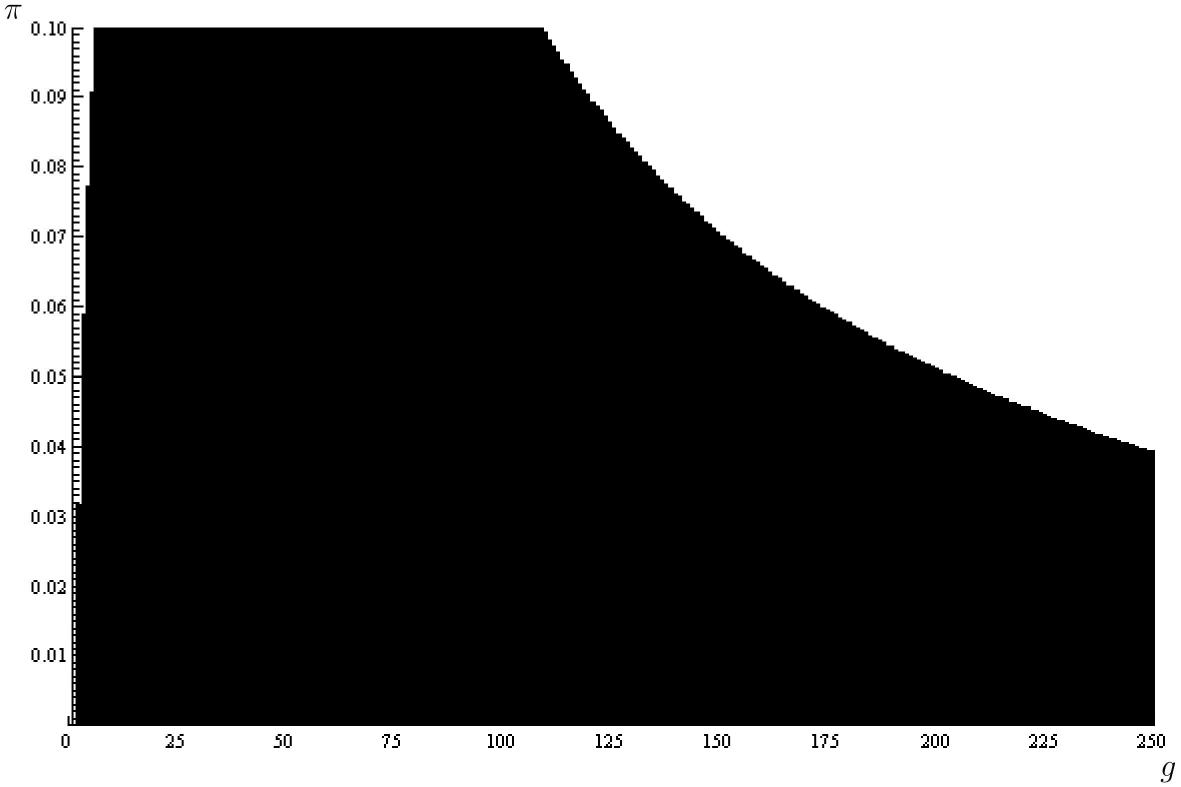

Figure 3.14: Comparison of simultaneous-move and sequential-move game, $\epsilon = \frac{1}{500}$

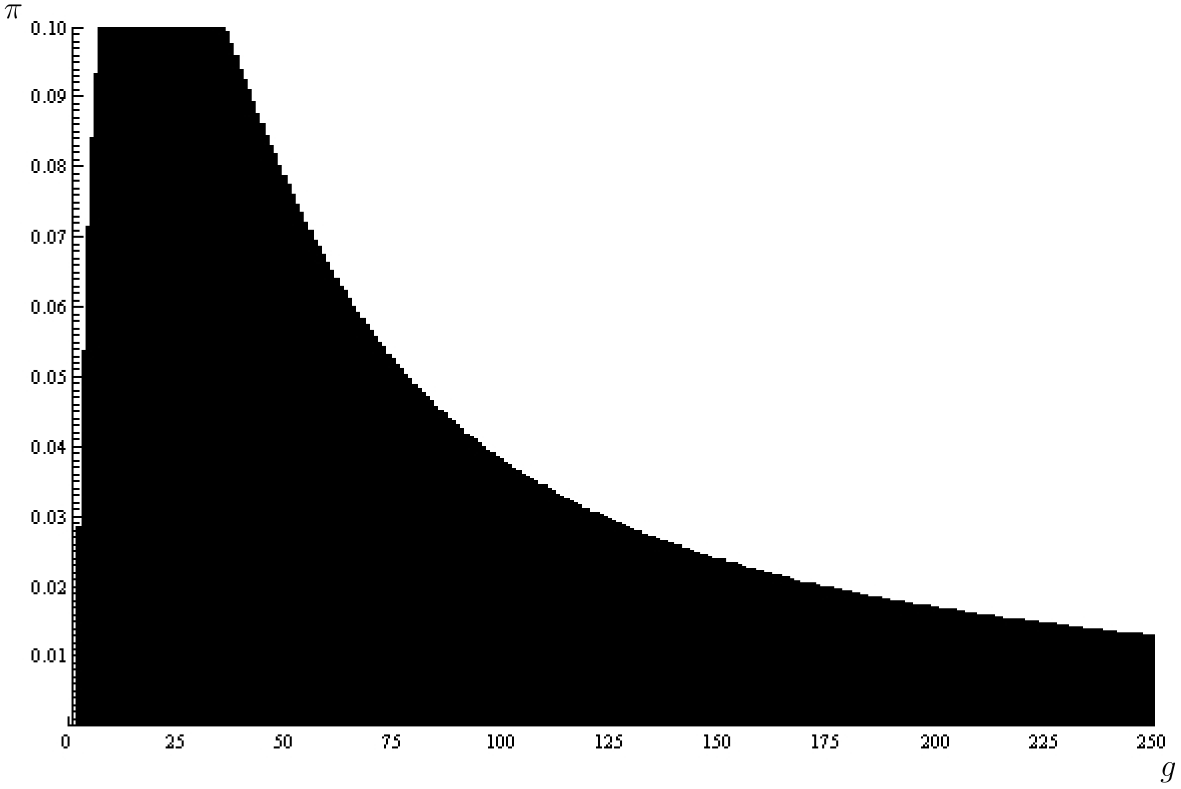

Figure 3.15: Comparison of simultaneous-move and sequential-move game, $\epsilon = \frac{1}{100}$

Appendix

This is the source code to generate figure 3.13:

```
#include <oxstd.h>
#include <oxdraw.h>
#include <oxprob.h>

decl f=2;
decl pi=0.05;
decl qs=1/3;
decl gl=250;
decl totgen=5000;
decl n=6;
decl m=100;
decl e=1/500;

hypergeom(N,p,n,x)
{
return binomial(N*p,x)*binomial(N*(1-p),n-x)/binomial(N,n);
}

class Group
{
decl n;
decl nh;
decl nq;
decl swf;
Group();
reset(nn,nnh,nnq);
nextgen();
nextgenseq();
NH();
N();
SWF();
}

Group::Group()
{
}
```

```
Group::reset(nn,nnh,nnq)
{
n=nn;
nh=nnh;
nq=nnq;
}

Group::nextgen()
{
decl nn,nnh;
nnh=nh*(f-2/3*(n-nh)/(n-1));
nn=n*f-2/3*nh*(n-nh)/(n-1)+2/3*(n-nh)*(pi-(n-nh-1)/(n-1));
swf=nn-f*n+n;
decl q=nnh/nn;
q=(1-e)*q+e*(1-q);
n=nn;
nh=q*nn;
}

Group::nextgenseq()
{
decl nn,nnh;
nnh=nh*(f-2/3*(n-nh)/(n-1));
nn=n*f-2/3*nh*(n-nh)/(n-1)+1/3*(n-nh)*(pi*nh/(n-1)+pi-(n-nh-1)/(n-1));
swf=nn-f*n+n;
decl q=nnh/nn;
q=(1-e)*q+e*(1-q);
n=nn;
nh=q*nn;
}

Group::NH()
{
return nh*nq;
}

Group::N()
{
```

```
return n*nq;
}

Group::SWF()
{
return swf*nq;
}

main()
{
SetDrawWindow("Simulation␣Output");
DrawAxis(0,0,0,0,1,0.1,0.1,0.01,0);
DrawAxis(1,0,0,0,1,0.1,0.1,0.01,0);
decl i;
decl j;
decl qa;
decl group=new array[n+1];
decl gen;
decl mg;
for(mg=0;mg<n+1;mg++)
{
group[mg]=new Group();
}
decl g;
decl q;
decl NH;
decl N;
decl gb;
decl oldx;
decl oldy;
decl oldSWFx;
decl oldSWFy;
decl tSWF;
decl pass;
decl oldN;
decl SWF;

oldx=0;
oldy=0;
```

```
oldSWFx=0;
oldSWFy=0;
for(j=1;j<gl+1;j++)
{
gb=j;
q=qs;
qa=0;
tSWF=0;
for(gen=0;gen*gb<totgen;gen++)
{
N=m*n;
for(mg=0;mg<n+1;mg++)
{
group[mg].reset(n,mg,N/n*hypergeom(N,q,n,mg));
}
for(g=0;g<gb;g++)
{
oldN=N;
N=0;
NH=0;
SWF=0;
for(mg=0;mg<n+1;mg++)
{
group[mg].nextgen();
N=N+group[mg].N();
NH=NH+group[mg].NH();
SWF=SWF+group[mg].SWF();
}
SWF=SWF/oldN;
tSWF=tSWF+SWF;
q=NH/N;
qa=qa+q;
}
}
qa=qa/(gen*gb+g);
tSWF=tSWF/(gen*gb+g);
if(oldx==0&&oldy==0) {oldx=j; oldy=qa;}
DrawLine(0,oldx,oldy,j,qa,1);
oldx=j;
```

```
oldy=qa;
if(oldSWFx==0&&oldSWFy==0)
{oldSWFx=j; oldSWFy=tSWF; DrawLine(1,oldSWFx,oldSWFy,j,tSWF,1);}
else
{DrawLine(1,oldSWFx,oldSWFy,j,tSWF,1); oldSWFx=j; oldSWFy=tSWF;}
}
oldx=0;
oldy=0;
oldSWFx=0;
oldSWFy=0;
for(j=1;j<gl+1;j=j+1)
{
gb=j;
q=qs;
qa=0;
tSWF=0;
for(gen=0;gen*gb<totgen;gen++)
{
N=m*n;
for(mg=0;mg<n+1;mg++)
{
group[mg].reset(n,mg,N/n*hypergeom(N,q,n,mg));
}
for(g=0;g<gb;g++)
{
oldN=N;
N=0;
NH=0;
SWF=0;
for( mg=0;mg<n+1;mg++)
{
group[mg].nextgenseq();
N=N+group[mg].N();
NH=NH+group[mg].NH();
SWF=SWF+group[mg].SWF();
}
SWF=SWF/oldN;
tSWF=tSWF+SWF;
q=NH/N;
qa=qa+q;
}
}
qa=qa/(gen*gb+g);
tSWF=tSWF/(gen*gb+g);
if(oldx==0&&oldy==0) {oldx=j; oldy=qa;}
DrawLine(0,oldx,oldy,j,qa,14);
oldx=j;
oldy=qa;
if(oldSWFx==0&&oldSWFy==0)
{oldSWFx=j; oldSWFy=tSWF; DrawLine(1,oldSWFx,oldSWFy,j,tSWF,14);}
else
{DrawLine(1,oldSWFx,oldSWFy,j,tSWF,14); oldSWFx=j; oldSWFy=tSWF;}
}
DrawLine(0,1,qs,gl,qs,10);
DrawText(0, ”Periods␣of␣Isolation”, 0, 0, −1, −1, TEXT_XLABEL);
DrawText(0, ”Average␣Proportion␣of␣Altruists”, 0, 0, −1, −1, TEXT_YLABEL);
DrawText(1, ”Periods␣of␣Isolation”, 0, 0, −1, −1, TEXT_XLABEL);
DrawText(1, ”Average␣Social␣Welfare”, 0, 0, −1, −1, TEXT_YLABEL);
ShowDrawWindow();
}
```

Conclusion

This thesis has explored the interaction between altruistic preferences and punishment systems using a new model, the sequential punishment model. This has similarities to existing infinite horizon sequential games such as those analysed by Samuelson (Samuelson, 1958) and Hammond (Hammond, 1975). The intended style of model is to encapsulate something general about the world in a simple and abstract manner, in the idiom of Samuelson's overlapping-generations pension game (Samuelson, 1958) or Diamond's model of fiat money in a "coconut swapping" economy (Diamond, 1984).

The structure of the model in is straightforward. A series of individuals receive sequential opportunities to inflict harm upon others. They can choose either to harm themselves, another individual, or to refrain from inflicting harm. If they choose to inflict harm, they impose a cost of one unit of felicity upon the individual they harm, but receive a felicity benefit π_t whose value is randomly generated from the same distribution each period, but known to *all* individuals *before* each player makes their move. Each period, a new distinct individual receives a harm opportunity, so that each individual only moves once during the entire game. Individuals are indifferent as to whom they harm, and so the ability to focus harm onto the most recent defector allows punishment strategies to be specified in a fairly economical manner. Although the model is simple, fully analysing all of the possible subgame-perfect Nash equilibria in the infinite-move version of the sequential punishment game is nonetheless an involved process, requiring the use of Abreu's framework of optimal punishment paths, originally developed for infinitely-repeated simultaneous stage games (Abreu, 1988).

The sequential punishment model is intended to capture the fundamentally vicarious nature of human social existence, in the sense that social interactions consist of repeated opportunities to impose negative externalities upon other individuals, with a gain to onesself. In this sense it is meant to analyse the *emergence* of incentive systems (of which the market is of course the most common, but by no means the only, example studied by economists), rather than assuming that they are already in place. The second key feature of the sequential punishment model is that defectors from socially agreed equilibria can be punished for their misbehaviour. Finally, the presence of bona fide altruistic preferences is central to the model. Altruism and punishment systems therefore provide *substitute* methods of achieving a more socially efficient outcome.

The fundamental argument of the thesis is that these substitute methods can interfere with one another in a perverse manner. Once punishment can be used, it is no longer the case that more altruism is always better for society. This was the argument in "The Socially Optimal Level of Altruism" (Chapter 2). In particular, when altruism becomes close to perfect in the infinite-move sequential punishment game, then socially efficient equilibria which are supportable at lower levels of altruism are rendered unsupportable. This result was proved for a fixed cost, and for a general continuous random distribution of the benefit, provided that the benefit and cost of punishment opportunities are always positive, and that the benefit from punishment is always less than the cost (i.e. $0 \leq \hat{\pi} < \pi_t \leq 1$).

The intuition for this result can best be seen using the framework of willingness, severity and temptation effects. With greater altruism, individuals are less *willing* to punish, a given punishment becomes less *severe* and individuals are less *tempted* to misbehave. With perfect altruism (where the coefficient of altruism $\theta = 1$), there is no temptation to misbehave at all. However, there is also no credible threat of punishment, because threatening to "focus" the harm inflicted by other individuals onto a fully altruistic defector does not affect their social utility. Since punishment is ineffective as a deterrent, and fully altruistic individuals will also not want to inflict harm, no punishment will be possible. Therefore, if the coefficient of altruism θ is reduced slightly below 1, the temptation effect (leading to greater temptation to defect and therefore making it more difficult to support an efficient outcome) will dominate. Thus, high θ but less than 1 will destroy equilibria which were available at $\theta = 1$. If the discount factor δ is sufficiently high, however, then further reductions in θ bring the willingness and severity effects into play, both of which can lead to an efficient equilibrium becoming supportable once again. Therefore the analysis of the sequential punishment model allows us to conclude that intermediate levels of altruism are the most socially desirable.

One loose end that has not been fully tied down is whether optimal paths are always quasi-flat. Ideally, one would like to be able to prove this. However, the result that the optimal *semi-constrained* path must be quasi-flat was sufficient for proving the main result of the paper, that too much altruism will prevent the socially efficient outcome. It would also be interesting in future to analyse the case where individuals can be martyrs, with $\theta \geq 1$.

Although the willingness, severity and temptation effects had a very specific application in analysing the equilibria of the sequential punishment model, they should also have important broader applications. They would be highly relevant to other areas of economics such as optimal taxation theory, the application of economic theory to law and criminology and to further modelling of optimal moral codes in different and extended contexts (see concluding paragraph below). In particular, the severity effect will emerge when any system of fines in used to punish offenders, provided some or all of the revenue is redistributed. The argument from the sequential punishment model that altruism "dents" incentive schemes should thus have wide implications.

A related, but distinct, result - that the use of punishment systems weakens the evolution of altruistic preferences via the group selection mechanism, with ambiguous normative consequences - was presented in "Punishment and the Potency of Group Selection" (Chapter 3). This was proved for a simplified 3-player, 2-move version of the sequential punishment model. The result is general in the sense that it applies for any kind of dispersal mechanism, not just the "haystacks" model used for illustrative purposes. It is very specific, on the other hand, in that the result is limited to the very specific game utilised. One feels that more general results should be available in this area.

Finally, the most serious limitation of the sequential punishment model given the purpose for which it was developed (the normative analysis of altruistic preferences) is its reliance on the assumption of perfect information. One feels that another important explanation for why imperfect altruism might be optimal is the issue of asymmetric information - individuals do not know enough about the world to allow a perfect utilitarian moral code to be workable. Since individuals know more about themselves and others in their immediate vicinity, it might be optimal for them to weight the welfare of such proximate agents more heavily than those who are more remote. A model which is able to capture this insight, either as an extension, or in parallel, to the sequential punishment model, would be a desirable avenue for further research. A key component of such a model would need to be the use of government sanctions (e.g. fines) to punish certain actions prohibited by the authorities to such partially altruistic individuals. The analytic framework of willingness, severity and temptation effects developed in this thesis will therefore continue to be highly relevant.

References

ABEL, ANDREW B. AND WARSHAWSKY, MARK (1988). **"Specification of the Joy of Giving: Insights from Altruism"**. *The Review of Economics and Statistics*, 70(1), 145–149.

ABREU, DILIP (1986). **"Extremal Equilibria of Oligopolistic Supergames"**. *Journal of Economic Theory*, 39, 191–225.

ABREU, DILIP (1988). **"On the Theory of Infinitely Repeated Games with Discounting"**. *Econometrica*, 56(2), 383–396.

ADAMS, JAMES D. (1980). **"Personal Wealth Transfers"**. *The Quarterly Journal of Economics*, 95(1), 159–179.

AGEE, MARK D. AND CROCKER, THOMAS D. (1996). **"Parents' Discount Rates for Child Quality"**. *Southern Economic Journal*, 63(1), 36–50.

ALESINA, ALBERTO AND GLAESER, EDWARD AND SACERDOTE, BRUCE (2001). **"Why Doesn't the United States Have a European-Style Welfare State?"**. *Brookings Papers on Economic Activity*, 2001(2), 187–254.

ALTIG, DAVID AND AUERBACH, ALAN J. AND KOTLIKOFF, LAURENCE J. AND SMETTERS, KENT A. AND WALLISER, JAN (2001). **"Simulating Fundamental Tax Reform in the United States"**. *The American Economic Review*, 91(3), 574–595.

ALTONJI, JOSEPH G. AND HAYASHI, FUMIO AND KOTLIKOFF, LAURENCE J. (1992). **"Is the Extended Family Altruistically Linked? Direct Tests Using Micro Data"**. *The American Economic Review*, 82(5), 1177–1198.

Altonji, Joseph G. and Hayashi, Fumio and Kotlikoff, Laurence J. (1997). **"Parental Altruism and Inter Vivos Transfers: Theory and Evidence"**. *The Journal of Political Economy*, 105(6), 1121–1166.

Andreoni, James (1989). **"Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence"**. *The Journal of Political Economy*, 97(6), 1447–1458.

Andreoni, James (1990). **"Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving"**. *The Economic Journal*, 100(401), 464–477.

Andreoni, James (1995a). **"Cooperation in Public-Goods Experiments: Kindness or Confusion?"**. *The American Economic Review*, 85(4), 891–904.

Andreoni, James (1995b). **"Warm-Glow Versus Cold-Prickle: The Effects of Positive and Negative Framing on Cooperation in Experiments"**. *The Quarterly Journal of Economics*, 110(1), 1–21.

Andreoni, James and Castillo, Marco and Petrie, Ragan (2003). **"What Do Bargainers' Preferences Look Like? Experiments with a Convex Ultimatum Game"**. *The American Economic Review*, 93(3), 672–685.

Andreoni, James and Miller, John (2002). **"Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism"**. *Econometrica*, 70(2), 737–753.

Andreoni, James and Miller, John H. (1993). **"Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence"**. *The Economic Journal*, 103(418), 570–585.

Arrow, Kenneth J. and Debreu, Gerard (1954). **"Existence of an Equilibrium for a Competitive Economy"**. *Econometrica*, 22(3), 265–290.

Aumann, Robert J. and Shapley, Lloyd S. (1992). **"Long Term Competition - A Game Theoretic Analysis"**. UCLA Economics Working Papers 676, UCLA Department of Economics.

Barrett, Scott (1994). **"Self-Enforcing International Environmental Agreements"**. *Oxford Economic Papers*, 46, pp. 878–894.

Barro, Robert J. (1974). **"Are Government Bonds Net Wealth?"**. *The Journal of Political Economy*, 82(6), 1095–1117.

Baumol, William J. (1975). **Business Responsibility and Economic Behaviour**. In E. S. Phelps (Ed.), *Altruism, Morality and Economic Theory* (pp. 45–57). Russell Sage Foundation, New York.

Becker, Gary S. (1974). **"A Theory of Social Interactions"**. *Journal of Political Economy*, 82, 1063–1093.

Becker, Gary S. (1976). **"Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology"**. *Journal of Economic Literature*, 14(3), 817–826.

Becker, Gary S. (1981). **"Altruism in the Family and Selfishness in the Market Place"**. *Economica*, 48(189), 1–15.

Behrman, Jere R. and Pollak, Robert A. and Taubman, Paul (1986). **"Do Parents Favor Boys?"**. *International Economic Review*, 27(1), 33–54.

Behrman, Jere R. and Taubman, Paul (1986). **"Birth Order, Schooling, and Earnings"**. *Journal of Labor Economics*, 4(3), S121–S145.

Benoit, Jean-Pierre and Krishna, Vijay (1993). **"Renegotiation in Finitely Repeated Games"**. *Econometrica*, 61(2), 303–23.

Bergstrom, Theodore C. (1989). **"A Fresh Look at the Rotten Kid Theorem–and Other Household Mysteries"**. *The Journal of Political Economy*, 97(5), 1138–1159.

Bergstrom, Theodore C. (2002). **"Evolution of Social Behavior: Individual and Group Selection"**. *Journal of Economic Perspectives*, 16(2), 67–88.

Bernheim, B. Douglas and Bagwell, Kyle (1988). **"Is Everything Neutral?"**. *The Journal of Political Economy*, 96(2), 308–338.

Bernheim, Douglas B. and Stark, Oded (1988). **"Altruism within the Family Reconsidered: Do Nice Guys Finish Last?"**. *The American Economic Review*, 78(5), 1034–1045.

Blackmore, Susan J. (1999). **The Meme Machine**. Oxford University Press, Oxford.

Boyd, Robert and Gintis, Herbert and Bowles, Samuel and Richerson, Peter J. (2003). **"The Evolution of Altruistic Punishment"**. *Proceedings of the National Academy of Sciences*, 100(6), 3531–3535.

Boyd, R. and Richerson, Peter J. (1982). **"Cultural Transmission and the Evolution of Cooperative Behavior"**. *Human Ecology*, 10(3), 325–351.

Browning, Martin and Chiappori, Pierre-André (1998). **"Efficient Intra-Household Allocations: A General Characterization and Empirical Tests"**. *Econometrica*, 66(6), 1241–1278.

Bruce, Neil and Waldman, Michael (1990). **"The Rotten-Kid Theorem Meets the Samaritan's Dilemma"**. *The Quarterly Journal of Economics*, 105(1), 155–165.

Calmfors, Lars and Driffill, John (1988). **"Bargaining Structure, Corporatism and Macroeconomic Performance"**. *Economic Policy*, 6, 13–61.

Camerer, Colin and Thaler, Richard H. (1995). **"Anomalies: Ultimatums, Dictators and Manners"**. *The Journal of Economic Perspectives*, 9(2), 209–219.

Case, Anne and Lin, I. Fen and McLanahan, Sara (1999). **"Household Resource Allocation in Stepfamilies: Darwin Reflects on the Plight of Cinderella"**. *The American Economic Review*, 89(2), 234–238.

Chiappori, Pierre-André (1992). **"Collective Labor Supply and Welfare"**. *The Journal of Political Economy*, 100(3), 437–467.

Clark, Jeremy (1998). **"Fairness in Public Good Provision: An Investigation of Preferences for Equality and Proportionality"**. *The Canadian Journal of Economics*, 31(3), 708–729.

Cohen, Gerry A. (1978). **Karl Marx's Theory of History: A Defence**. Clarendon Press, Oxford.

Collard, David (1978). **Altruism and Economy**. Martin Robertson Ltd, Oxford.

Cooper, Ben and Wallace, Chris (2004). **"Group Selection and the Evolution of Altruism"**. *Oxford Economic Papers*, 56, 307–330.

Cox, Donald (1987). **"Motives for Private Income Transfers"**. *The Journal of Political Economy*, 95(3), 508–546.

Cox, Donald and Rank, Mark R. (1992). **"Inter-Vivos Transfers and Intergenerational Exchange"**. *The Review of Economics and Statistics*, 74(2), 305–314.

Cremer, Jacques (1986). **"Cooperation in Ongoing Organizations"**. *The Quarterly Journal of Economics*, 101(1), 33–49.

d'Aspremont, Claude and Gevers, Louis (1977). **"Equity and the Informational Basis of Collective Choice"**. *Review of Economic Studies*, 44(2), 199–209.

Davies, James B. and Zhang, Junsen (1995). **"Gender Bias, Investments in Children, and Bequests"**. *International Economic Review*, 36(3), 795–818.

Dawes, Robyn M. and Thaler, Richard H. (1988). **"Anomalies: Cooperation"**. *The Journal of Economic Perspectives*, 2(3), 187–197.

Dawkins, Richard (1976). **The Selfish Gene**. Oxford University Press, Oxford.

Diamond, Peter (1984). **"Money in Search Equilibrium"**. *Econometrica*, 52(1), 1–20.

Diamond, Peter A. and Stiglitz, Joseph E. (1974). **"Increases in Risk and in Risk Aversion"**. *Journal of Economic Theory*, 8(3), 337–360.

Dowrick, Steve and Dunlop, Yvonne and Quiggin, John (1998). **"The Cost of Life Expectancy and the Implicit Social Valuation of Life"**. *The Scandinavian Journal of Economics*, 100(4), 673–691.

Farrell, Joseph T. and Maskin, Eric S. (1989). **"Renegotiation in Repeated Games"**. *Games and Economic Behaviour*, 1(1), 327–360.

Fehr, Ernst and Fischbacher, Urs (2003). **"The Nature of Human Altruism"**. *Nature*, 425, 785–791.

Fehr, Ernst and Gächter, Simon (2000a). **"Cooperation and Punishment in Public Goods Experiments"**. *The American Economic Review*, 90(4), 980–994.

Fehr, Ernst and Gächter, Simon (2000b). **"Fairness and Retaliation: The Economics of Reciprocity"**. *The Journal of Economic Perspectives*, 14(3), 159–181.

Fehr, Ernst and Gächter, Simon (2002a). **"Altruistic Punishment in Humans"**. *Nature*, 415, 137–140.

Fehr, Ernst and Gächter, Simon (2002b). **"Strong Reciprocity, Human Cooperation and the Enforcement of Social Norms"**. *Human Nature*, 13, 1–25.

Frank, Robert H. and Gilovich, Thomas D. and Regan, Dennis T. (1996). **"Do Economists Make Bad Citizens?"**. *The Journal of Economic Perspectives*, 10(1), 187–192.

Freeman, Katherine B. (1984). **"The Significance of Motivational Variables in International Public Welfare Expenditures"**. *Economic Development and Cultural Change*, 32(4), 725–748.

Freeman, Richard B. (1997). **"Working for Nothing: The Supply of Volunteer Labor"**. *Journal of Labor Economics*, 15(1), S140–S166.

Fudenberg, Drew and Maskin, Eric (1986). **"The Folk Theorem in Repeated Games with Discounting or with Incomplete Information"**. *Econometrica*, 54(3), 533–54.

Gevers, Louis (1979). **"On Interpersonal Comparability and Social Welfare Orderings"**. *Econometrica*, 47(1), 75–89.

Güth, W. and Yaari, M. (1992). **An Evolutionary Approach to Explain Reciprocal Behavior in a Simple Strategic Game**. In U. Witt (Ed.), *Explaining Process and Change: Approaches to Evolutionary Economics* (pp. 23–34). University of Michigan Press.

HAMILTON, W. D. (1963). **"The Evolution of Altruistic Behavior"**. *The American Naturalist*, 97(896), 354–356.

HAMILTON, W D (1972). **"Altruism and Related Phenomena, Mainly in Social Insects"**. *Annual Review of Ecology and Systematics*, 3(1), 193–232.

HAMMOND, PETER (1975). **Charity: Altruism or Cooperative Egoism?** In E. S. Phelps (Ed.), *Altruism, Morality and Economic Theory* (pp. 115–131). Russell Sage Foundation, New York.

HANEMANN, W. MICHAEL (1994). **"Valuing the Environment Through Contingent Valuation"**. *The Journal of Economic Perspectives*, 8(4), 19–43.

HARSANYI, JOHN C. (1986). **Utilitarian Morality in a World of Very Half-hearted Altruists**. In R. M. Heller, Walter P. Starr & D. A. Starrett (Eds.), *Social Choice and Public Decision Making* (pp. 57–73). Cambridge University Press.

HAYASHI, FUMIO (1995). **"Is the Japanese Extended Family Altruistically Linked? A Test Based on Engel Curves"**. *The Journal of Political Economy*, 103(3), 661–674.

HAYEK, FRIEDRICH A. (1960). **The Constitution of Liberty**. Routledge and Kegan Paul Ltd, London.

HAYEK, FRIEDRICH A. (1988). **The Fatal Conceit: The Errors of Socialism**. Routledge.

HECKATHORN, DOUGLAS D. (1990). **"Collective Sanctions and Compliance Norms: A Formal Theory of Group-Mediated Social Control"**. *American Sociological Review*, 55(3), 366–384.

HIRSHLEIFER, JACK (1977). **"Economics from a Biological Viewpoint"**. *The Journal of Law and Economics*, 20(1), 1.

HOFFMAN, ELIZABETH AND SPITZER, MATTHEW L. (1985). **"Entitlements, Rights, and Fairness: An Experimental Examination of Subjects' Concepts of Distributive Justice"**. *The Journal of Legal Studies*, 14(2), 259–297.

Hollander, Heinz (1990). **"A Social Exchange Approach to Voluntary Cooperation"**. *The American Economic Review*, 80(5), 1157–1167.

Holmes, Thomas P. (1990). **"Self-Interest, Altruism, and Health-Risk Reduction: An Economic Analysis of Voting Behavior"**. *Land Economics*, 66(2), 140–149.

Huck, Steffen and Oechssler, Jorg (1999). **"The Indirect Evolutionary Approach to Explaining Fair Allocations"**. *Games and Economic Behavior*, 28(1), 13–24.

Johansson, Olof (1997). **"Optimal Pigovian Taxes under Altruism"**. *Land Economics*, 73(3), 297–308.

John, A. Andrew and Pecchenino, Rowena A. (1997). **"International and Intergenerational Environmental Externalities"**. *The Scandinavian Journal of Economics*, 99(3), 371–387.

Jones-Lee, Michael. W. (1992). **"Paternalistic Altruism and the Value of Statistical Life"**. *The Economic Journal*, 102(410), 80–90.

Jouvet, Pierre-Andre and Michel, Philippe and Vidal, Jean-Pierre (2000). **"Intergenerational Altruism and the Environment"**. *The Scandinavian Journal of Economics*, 102(1), 135–150.

Kandori, Michihiro (1992). **"Repeated Games Played by Overlapping Generations of Players"**. *The Review of Economic Studies*, 59(1), 81–92.

Kaplow, Louis (1998). **"Tax Policy and Gifts"**. *The American Economic Review*, 88(2), 283–288.

Kotlikoff, Laurence J. (1988). **"Intergenerational Transfers and Savings"**. *The Journal of Economic Perspectives*, 2(2), 41–58.

Kotlikoff, Laurence J. and Persson, Torsten and Svensson, Lars E. O. (1988). **"Social Contracts as Assets: A Possible Solution to the Time-Consistency Problem"**. *The American Economic Review*, 78(4), 662–677.

KOTLIKOFF, LAURENCE J. AND SPIVAK, AVIA (1981). **"The Family as an Incomplete Annuities Market"**. *The Journal of Political Economy*, 89(2), 372–391.

KUEHLWEIN, MICHAEL (1993). **"Life-Cycle and Altruistic Theories of Saving with Lifetime Uncertainty"**. *The Review of Economics and Statistics*, 75(1), 38–47.

KUHN, THOMAS S. (1970). **The Structure of Scientific Revolutions**. University of Chicago Press, Chicago.

KUHN, THOMAS S. (1977). **Objectivity, Value Judgement and Theory Choice**. In *The Essential Tension*. University of Chicago Press, Chicago.

KURZ, MORDECAI (1984). **"Capital Accumulation and the Characteristics of Private Intergenerational Transfers"**. *Economica*, 51(201), 1–22.

LAGRANGE, JOSEPH-LOUIS (1806). **Leçons Sur le Calcul des Fonctions**. Chez Courgier, Paris.

LAITNER, JOHN AND JUSTER, F. THOMAS (1996). **"New Evidence on Altruism: A Study of TIAA-CREF Retirees"**. *The American Economic Review*, 86(4), 893–908.

LAKATOS, IMRE (1970). **Falsification and the Methodology of Social Science Research Programmes**. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge*. Cambridge University Press, Cambridge.

LAMBSON, VAL EUGENE (1987). **"Optimal Penal Codes in Price-Setting Supergames with Capacity Constraints"**. *Review of Economic Studies*, 54(3), 385–97.

LUCAS, ROBERT JR (1976). **"Econometric Policy Evaluation: A Critique"**. *Carnegie-Rochester Conference Series on Public Policy*, 1(1), 19–46.

MANDEVILLE, B. (1732). **The Fable of the Bees or Private Vices, Publick Benefits**. Liberty Fund, Indianapolis.

MCGRANAHAN, LESLIE MOSCOW (2000). **"Charity and the Bequest Motive: Evidence from Seventeenth-Century Wills"**. *The Journal of Political Economy*, 108(6), 1270–1291.

McKean, Roland N. (1975). **Economics of Trust, Altruism and Corporate Responsibility**. In E. S. Phelps (Ed.), *Altruism, Morality and Economic Theory* (pp. 29–44). Russell Sage Foundation, New York.

McKelvey, Richard D. and Palfrey, Thomas R. (1992). **“An Experimental Study of the Centipede Game”**. *Econometrica*, 60(4), 803–836.

Meade, James E. (1973). **Theory of Economic Externalities: The Control of Environmental Pollution and Similar Social Costs**. Sijthoff, Leiden.

Merton, Robert K. (1968). **Social Theory and Social Structure**. Free Press, New York.

Messner, Matthias and Polborn, Mattias K. (2003). **“Cooperation in Stochastic OLG games”**. *Journal of Economic Theory*, 108(1), 152–168.

Olson, Mancur (1965). **The Logic of Collective Action**. Harvard Economic Studies, Harvard.

Pattanaik, Prasanta K. (1971). **Voting and Collective Choice**. Cambridge University Press, Cambridge.

Popp, David (2001). **“Altruism and the Demand for Environmental Quality”**. *Land Economics*, 77(3), 339–349.

Popper, Karl (1959). **The Logic of Scientific Discovery**. Hutchinson Press, London.

Price, George R. (1970). **“Selection and Covariance”**. *Nature*, 227, 520–521.

Quiggin, John (1998). **“Individual and Household Willingness to Pay for Public Goods”**. *American Journal of Agricultural Economics*, 80(1), 58–63.

Rangazas, Peter C. (1996). **“Fiscal Policy and Endogenous Growth in a Bequest-Constrained Economy”**. *Oxford Economic Papers*, 48(1), 52–74.

Rawls, John (1999). **A Theory of Justice**. Oxford University Press, Oxford.

Rees, Ray (1993). **“Tacit Collusion”**. *Oxford Review of Economic Policy*, 9(2), 27–40.

Roberts, Kevin W S (1980). **“Interpersonal Comparability and Social Choice Theory”**. *Review of Economic Studies*, 47(2), 421–39.

Rotemberg, Julio J. (1994). **“Human Relations in the Workplace”**. *The Journal of Political Economy*, 102(4), 684–717.

Rothschild, Michael and Stiglitz, Joseph E. (1970). **“Increasing Risk: I. A Definition”**. *Journal of Economic Theory*, 2(3), 225 – 243.

Rubinstein, Ariel (1979). **“Equilibrium in Supergames with the Overtaking Criterion”**. *Journal of Economic Theory*, 21(1), 1–9.

Ruffin, Roy J. (1972). **“Pollution in a Crusoe Economy”**. *The Canadian Journal of Economics*, 5(1), 110–118.

Samuelson, Paul A. (1958). **“An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money”**. *Journal of Political Economy*, 66, 467.

Seater, John J. (1993). **“Ricardian Equivalence”**. *Journal of Economic Literature*, 31(1), 142–190.

Sen, Amartya (1974). **“Informational bases of alternative welfare approaches : Aggregation and income distribution”**. *Journal of Public Economics*, 3(4), 387–403.

Simon, Carl P. and Blume, Lawrence (1994). **Mathematics for Economists**. W. W. Norton and Company, New York.

Simon, Herbert A. (1993). **“Altruism and Economics”**. *The American Economic Review*, 83(2), 156–161.

Sloan, Frank A. and Zhang, Harold H. and Wang, Jingshu (2002). **“Upstream Intergenerational Transfers”**. *Southern Economic Journal*, 69(2), 363–380.

Smith, Adam [1776] (1976). **An Inquiry into the Nature and Causes of the Wealth of Nations**. Oxford University Press, Oxford.

Sober, E. and Wilson, D. S. (1994). **"Reintroducing Group Selection to the Human Behavioral Sciences"**. *Behavioral and Brain Sciences*, 17(4), 585–654+.

Sober, Elliot and Wilson, David S. (1999). **Unto Others**. Harvard University Press, Cambridge, Massachussets.

Soltis, Joseph and Boyd, Robert and Richerson, Peter J. (1995). **"Can Group-Functional Behaviors Evolve by Cultural Group Selection?: An Empirical Test"**. *Current Anthropology*, 36(3), 473–494.

Stark, Oded (1989). **"Altruism and the Quality of Life"**. *The American Economic Review*, 79(2), 86–90.

Stark, Oded (1995). **Altruism and Beyond**. Cambridge University Press, Cambridge.

Stark, Oded and Falk, Ita (1998). **"Transfers, Empathy Formation, and Reverse Transfers"**. *The American Economic Review*, 88(2), 271–276.

Stevens, Thomas H. and More, Thomas A. and Glass, Ronald J. (1994). **"Interpretation and Temporal Stability of CV Bids for Wildlife Existence: A Panel Study"**. *Land Economics*, 70(3), 355–363.

Sugden, Robert (1982). **"On the Economics of Philanthropy"**. *The Economic Journal*, 92(366), 341–350.

Sugden, Robert (1984). **"Reciprocity: The Supply of Public Goods Through Voluntary Contributions"**. *The Economic Journal*, 94(376), 772–787.

Tcha, Moonjoong (1996). **"Altruism and Migration: Evidence from Korea and the United States"**. *Economic Development and Cultural Change*, 44(4), 859–878.

Turner, Matthew A. (1997). **"Parental Altruism and Common Property Regulation"**. *The Canadian Journal of Economics*, 30(4a), 809–821.

Weaver, Robert D. (1996). **"Prosocial Behavior: Private Contributions to Agriculture's Impact on the Environment"**. *Land Economics*, 72(2), 231–247.

Weiss, Yoram and Willis, Robert J. (1993). **“Transfers among Divorced Couples: Evidence and Interpretation”**. *Journal of Labor Economics*, 11(4), 629–679.

Wilhelm, Mark O. (1996). **“Bequest Behavior and the Effect of Heirs’ Earnings: Testing the Altruistic Model of Bequests”**. *The American Economic Review*, 86(4), 874–892.

Willis, Robert J. (1999). **“A Theory of Out-of-Wedlock Childbearing”**. *The Journal of Political Economy*, 107(6), S33–S64.