manuscript No.
(will be inserted by the editor)

Socially Optimal Altruism in a Game of Sequential Punishment

Richard Povey

30th April 2020

Abstract It is commonly recognized that altruism may have socially detrimental effects. This paper seeks further insights by analyzing the interaction between an abstract punishment "technology" and the intensity of altruistic motivation in an elementary game of sequential punishment. Sufficiently high levels of altruism are shown to generally result in a socially suboptimal outcome. Although the central model is stylized, the key driving effects should appear in more specific examples. 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 ("non-wasteful" punishments are less severe for more altruistic individuals).

Keywords altruism · incentives · efficiency · punishment · welfare

1 *Introduction*

> *[T]he 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)*

> *[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 and Stark, 1988)*

1.1 A Parable

Many problems of economic and social policy share the common feature that a mechanism must be found to give individual agents the incentive to act in a manner which is beneficial for society as a whole. Such incentives can be *intrinsic* to the

Hertford College & St Hilda's College, University of Oxford: Hertford College, Catte Street, Oxford, OX1 3BW. ; St Hilda's College, Cowley Place, Oxford, OX4 1DY. E-mail: *richard.povey@hertford.ox.ac.uk*, *richard.povey@st-hildas.ox.ac.uk*

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 complex human societies may in fact be a socially optimal "policy mix", rather than a mere second-best correction for the inadequate intrinsic motivation of individuals to act in a socially efficient manner.[1]

Altruistic behaviour has been fruitfully modelled in economic theory using infinitely-repeated stage games (Fudenberg and Maskin, 1986; Abreu, 1986), and infinite dynamic sequential games such as models with overlapping generations (Samuelson, 1958; Hammond, 1975; Cremer, 1986). This paper uses a similar stylized canonical framework to approach its central question. In common with the infinitely-repeated stage game model, agents in the model are infinitely-lived and discount the future. In common with overlapping generations models, players are "born"[2] and move sequentially, each player only getting to move once during the entire game.[3] The sequential punishment game 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 game.

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. (This cost of a bottle landing in ones' garden is normalized to 1 throughout the paper.) 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 of reality 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 personal gain, in myriad ways.

[1] For a discussion of other ways in which extrinsic motivations may undermine intrinsic ones, see Bénabou and Tirole (2003).

[2] This significantly simplifies welfare metrics.

[3] This may seem unrealistic, but we can argue that we only need to split a finite set of players into an infinite set of "egos" (Hammond, 1975).

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.[4] On the other hand, an alternative potential solution to the problem is altruism; if people care about others, they may refrain from harming them. 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".[5]

It is commonly recognized that the repetition of stage games such as the prisoners' dilemma, or the Cournot and Bertand oligopoly games, allows apparently altruistic behaviour to be incentivized, resulting in a Pareto-superior outcome for the players. However, such apparent altruism only reflects "enlightened self-interest". 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 an infinitely repeated game (Aumann and Shapley, 1992; Rubinstein, 1979; Fudenberg and Maskin, 1986; Wen, 2002). 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 provides a general workhorse model in which such results can be extended in order to accommodate *bona fide* altruistic motivation.[6] It approaches the standard question of sustainability of collusion but in the reverse direction: Given a certain level of impatience, how high or low can the level of altruism be in order for a socially efficient outcome to be attained? It is shown that there must generally be a "Goldilocks" level of altruism, neither too high or too low.

---

[4] We will throughout use "harm" to refer to the inflicting of a negative externality and "punish" to the specific use of such harm opportunities to construct punishment equilibria.

[5] Such a result has been hinted at in an intuitive manner in the existing literature, most directly by Bernheim and Stark:

> *[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)

Altruism has also been shown to be potentially welfare-reducing in the context of time inconsistency of interactions between altruistic agents due to moral hazard or "free-riding" (Lindbeck and Weibull, 1988) and intergenerational altruism in OLG models with environmental externalities (Asheim and Nesje, 2016). However, none of these studies examine the drawbacks to the abstract form of universal altruism modelled in the sequential punishment game explored in this paper.

[6] Hence we investigate the social welfare consequences of a simple version of "intrinsic reciprocity" (Sobel, 2005), assuming for simplicity and tractability a homogenous level of such motivation in the population of agents.

1.2 A Three-Effects Framework

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 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 miscreant is fined a certain amount and the revenue is spent on other individuals, this is a more severe punishment for a less altruistic agent because the fact that the fine revenue is spent on others mitigates the detrimental effect of the fine on the agent's utility by a smaller amount than if the agent were more altruistic.[7] 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. 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.[8]

1.3 Overview

Section 2 discusses some preliminary issues regarding the modelling of altruistic preferences. Sections 3 through 4 set up the notation for the sequential punishment game. The main body of novel results is in Sections 5 through 8, which progressively derive the central results characterizing the "Goldilocks" range of altruism levels within which a socially efficient outcome is obtainable. Section 9 provides an important final piece of the argument, in that it is necessary not only to show that too high a level of altruism will break the supportability of the socially efficient outcome, but also that the loss of social welfare in the resulting second-best world will often be non-negligible. Section 10 concludes by discussing some applications of the key theoretical results to economic, social and political institutions.

2 *Modelling Altruism*

When modelling altruism, a common approach, which will be applied in this paper, is to distinguish between *felicity*, which represents "direct" individual welfare

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

[8] Stark and Bernheim have observed that altruism can reduce the credibility of punishment if the potential punisher is perceived as a "softy" (Bernheim and Stark, 1988). They observe that this can lead greater altruism to have a negative impact (this is precisely the phenomenon we term the willingness effect) but argue that this must be analysed on a case-by-case basis. Here we provide an abstract minimal framework in which a plausible and intuitive socially optimal level (or, more precisely, range of levels) of altruism emerges, and which is shaped by the three incentive effects introduced in section 1.2 that should reproduce themselves in many more specific situations.

from consuming economic goods and *utility*, which, although personal to each individual, may depend upon the felicities of more than one individual. In order to conduct a normative analysis, a social welfare function must also be constructed, which must be a function of individual felicities or utilities (or both). The main alternative to the felicity-based approach is to model altruism as a "warm glow" where the altruistic action (e.g. giving resources to another individual) is itself a good which directly enters the individual's utility function (hence avoiding the need for the distinct felicity concept).[9]

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. Let felicity be represented by $v_1(x_1)$ and $v_2(x_2)$ and utility be denoted by $u_1(x_1,x_2)$ and $u_2(x_1,x_2)$, where x_i is the consumption bundle of agent i. We could take a number of directions in representing the two individuals' utility functions. We could have each individual's utility depend on a weighted sum of their felicity and the felicity of the other. Letting $\theta_i \geq 0$ be the *coefficient of altruism* for agent i, this can be represented as:

$$u_1(x_1,x_2) = v_1(x_1) + \theta_1 v_2(x_2)$$
$$u_2(x_1,x_2) = v_2(x_2) + \theta_2 v_1(x_1)$$

The potential problem with this formulation is that increasing their coefficient of altruism automatically increases the utility of each individual, regardless of any effect upon their behaviour. The sequential punishment game nonetheless uses this approach, because felicities rather than utilities form the basis of the normative analysis via a utilitarian (simple sum of felicities) social welfare function.[10]

An alternative approach worth considering briefly is to have each individual's utility depend on their felicity and the *utility* of other individuals:

$$u_1(x_1,x_2) = (1-\theta_1)v_1(x_1) + \theta_1 u_2(x_1,x_2)$$
$$u_2(x_1,x_2) = (1-\theta_2)v_2(x_2) + \theta_2 u_1(x_1,x_2)$$

[9] For a much more detailed survey of the many different forms that altruism has taken in the theoretical and empirical literature in economics, as well as an extensive justification for making the modelling of altruism a central research objective, see Sobel (2005).

[10] This approach 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 utility function. We wish to be able to assess whether moral preferences such as altruism are welfare-improving, and to show that this is sometimes not the case. Since altruism is endogenous to social welfare, there is a risk that the argument becomes circular. There are a number of potential approaches to providing a basis for a utilitarian (simple sum of felicities) welfare metric. As shown by Bergstrom (2006), if altruism takes a non-paternalistic form as in equation (2) and certain other conditions are fulfilled then a necessary condition for Pareto-efficiency is that the simple sum of felicities (private utilities) is maximized. However, given the infinite number of players and OLG-like structure of the sequential punishment model, the simplest way to justify the necessary and sufficient condition for social efficiency that no harm opportunities be taken (no bottles thrown in the island parable analogy) is to consider an "original position" (Rawls, 1999) where players currently "born" act as a "co-ordination device" by choosing equilibrium strategies for players who will be "born" and move in subsequent time periods. In that case, currently living players would unanimously agree upon an equilibrium in which no harm opportunities are taken in future, since any harm opportunity taken only produces a (social utility) benefit of θ multiplied by the "distance to the bin" (π) which is always less than the (social utility) cost of θ (multiplied by 1, the felicity cost of a bottle landing).

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

$$u_1(x_1,x_2) = \left(\frac{1-\theta_1}{1-\theta_1\theta_2}\right)v_1(x_1) + \left(\frac{\theta_1-\theta_1\theta_2}{1-\theta_1\theta_2}\right)v_2(x_2)$$

$$u_2(x_1,x_2) = \left(\frac{1-\theta_2}{1-\theta_1\theta_2}\right)v_2(x_2) + \left(\frac{\theta_2-\theta_1\theta_2}{1-\theta_1\theta_2}\right)v_1(x_1)$$

Even though the utilities of both individuals have been normalized so that they are an average of their own felicity and the utility of the other individual, we still get a multiplier term at the beginning of each individual's solved-out utility function. This multiplier effect can produce interesting results in some models.[11]

Note however that (as long as $\theta_1\theta_2 \neq 1$) we can still express each individual's utility as being ultimately dependent only on the felicities of all individuals. 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.

The definition of social utility functions in terms of weighted averages of felicities (and, if we wish to use this normative justification - see footnote 10 - 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 (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).

3 *The Sequential Punishment Game*

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 chooses to *accept* their harm opportunity, then player A_t suffers a *cost* in *felicity* of 1 unit. If player t *rejects* the opportunity, then there are no changes in felicity. If player t accepts, then they gain felicity equal to the *benefit*, π_t, which is drawn randomly and independently from a distribution defined by the probability density function $g(\pi)$, with support $[0,1]$. All players publicly observe the value of π_t before player t moves. We assume throughout that $g(\pi)$ is twice continuously differentiable. We will frequently use as an exemplar the case of a uniform benefit distribution so that $g(\pi) = 1$ for $0 \leq \pi \leq 1$, but results will be general for any $g(\pi)$.

Definition 1 *The* expected benefit value *will be denoted:* $\bar{\pi} = \int_0^1 \pi g(\pi) d\pi$.

[11] See for example Stark's model of marriage between altruistic females and males where males vary in their degree of altruism towards females, so that females may actually prefer a less altruistic male partner due to the multiplier effect (Bernheim and Stark, 1988).

3.1 Players' Preferences

Players act to maximise their utility function, 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 $0 \leq \delta < 1$.[13] We let 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 that $0 \leq \theta \leq 1$.[14]

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

Let T_t denote a *trigger level* for the benefit π_t above which player t chooses to inflict harm in period t (as we shall see, this depends upon the equilibrium strategies being played). Let $v_{i,t}$ be the *felicity* of player i in period t and let u_t be the expected stream of discounted utility of player t looking forward from period t:[15]

$$v_{i,t} = \left\{ \begin{array}{c} -1 \quad \text{if } T_t < \pi_t \text{ and } t \neq i \text{ and } A_t = i \\ \pi_t \quad \text{if } T_t < \pi_t \text{ and } t = i \text{ and } A_t \neq i \\ \pi_t - 1 \text{ if } T_t < \pi_t \text{ and } t = i \text{ and } A_t = i \\ 0 \qquad\qquad \text{otherwise} \end{array} \right\}$$

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

3.2 *Punishment Paths*

The sequential punishment game 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

[12] Every individual's utility function is a social welfare functional which aggregates 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 when no harm opportunities at all are taken. If we let $\theta = 1$ then we get the social welfare function, which also satisfies anonymity.

[13] 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).

[14] Note that "martyrs" with $\theta > 1$ and malevolent individuals with $\theta < 0$ are ruled out *a priori* for sake of tractability.

[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 and 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.

(1988)[16], who showed that optimal punishment can be exhaustively described using *punishment paths*. These will in general have a *carrot-and-stick* 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 game 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 game 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 game. 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 *A punishment path, denoted ψ, is a vector of trigger levels, subscripted by the point reached along the path.*[17] *Trigger levels must lie within the support for π, therefore $\forall_\psi \forall_k : \psi_k \in [0,1]$. 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 $\forall_k : \bar{\psi}_k = \bar{\psi}$.*[18] *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*

[16] 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 game analysed here is one such context. Although the sequential punishment game 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.

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

[18] We use $\bar{\psi}$ to denote both a flat punishment 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.

punishment path. Along a punishment path, it will be desirable to harm the most recent deviator as much as possible. 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 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 they are themselves threatened with punishment, in order to give them an incentive to inflict harm when it is unpleasant for them 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 game.

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). The different result in the sequential punishment game is driven by the presence of altruistic preferences, which cause "neutral observers", who are not being punished (but 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 inter-temporal 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 difference to a "neutral observer". (See Lemma 6.)

The socially efficient outcome is said to be *supportable* for given values of δ and θ if and only if there exists a punishment path such that the corresponding strategy

profile forms a subgame-perfect Nash equilibrium with a socially efficient initial path. A punishment path can be conceived as an infinite sequence of values of A_t and T_t (using notation defined in section 3.1), which is imposed following a deviation from the socially efficient initial path (upon which $\forall_t : T_t = 1$ so that no harm is ever inflicted at any time[19]). Checking for supportability involves two conditions. Firstly, the punishment path must be *sustainable*. This requires that individuals be incentivized to co-operate with the punishment path, generally by inflicting harm when they would prefer not to unless further unpleasant consequences are threatened. Secondly, given a sustainable punishment path, this path must also 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 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.*

$$U_k(\psi) = \left(\frac{1-\delta}{\delta}\right) \sum_{i=k+1}^{\infty} \left[\delta^{i-k} \int_{\psi_i}^{1} (\theta\pi - 1) g(\pi) d\pi\right] \tag{2}$$

$$V_k(\psi) = \left(\frac{1-\delta}{\delta}\right) \sum_{i=k+1}^{\infty} \left[\delta^{i-k} \int_{\psi_i}^{1} (\theta\pi - \theta) g(\pi) d\pi\right] \tag{3}$$

Note that, for a flat path, $\bar{\psi}$, these functions simplify to give (where $U : \mathbb{R} \longrightarrow \mathbb{R}$ and $V : \mathbb{R} \longrightarrow \mathbb{R}$):[20]

$$\begin{aligned} \forall_k : U_k(\bar{\psi}) = U(\bar{\psi}) &= \int_{\bar{\psi}}^{1} (\theta\pi - 1) g(\pi) d\pi \\ \forall_k : V_k(\bar{\psi}) = V(\bar{\psi}) &= \int_{\bar{\psi}}^{1} (\theta\pi - \theta) g(\pi) d\pi \end{aligned} \tag{4}$$

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.[21]

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

[19] Given the assumptions laid out in section 3, the socially efficient initial path always involves no harm opportunities ever being taken for any value of π in the support of $g(\pi)$.

[20] The functions defined in (5) through (8) 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.

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

punishment path is the one which minimises $U_0(\psi)$ subject to these constraints, which is the same as maximizing 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.

$$\lambda_k(\psi) = \left(\frac{\delta}{1-\delta}\right) V_k(\psi) - \left(\frac{\delta}{1-\delta}\right) U_0(\psi) + \psi_k - \theta \geq 0 \tag{5}$$

$$\mu_k(\psi) = \left(\frac{\delta}{1-\delta}\right) V_k(\psi) - \left(\frac{\delta}{1-\delta}\right) U_0(\psi) - \psi_k + \theta \geq 0 \tag{6}$$

$$\kappa(\psi) = -\left(\frac{\delta}{1-\delta}\right) U_0(\psi) + \theta - 1 \geq 0 \tag{7}$$

$$\phi(\psi) = -\left(\frac{\delta}{1-\delta}\right) U_0(\psi) \tag{8}$$

3.3 Intuition for the Central Result

In the one-shot version of the sequential punishment game (where $\delta = 0$), a socially efficient outcome can only be achieved if the coefficient of altruism θ is high enough. This is due solely to the temptation effect. Since there is therefore always a lower limit on θ, a higher θ (i.e. closer to 1) is always socially superior. (See Proposition 1.) In this simplest case there are no willingness or severity effects to consider, and so the result that altruism is always socially beneficial is unambiguous.

In the more general infinite-period sequential punishment game (where $0 < \delta < 1$), a socially efficient outcome can be achieved more easily due to the fact that players are able to observe other players' past actions and to choose whom they harm in order to enforce credible threats of future punishment upon players who are tempted to inflict harm in the current period when this would be socially inefficient. The ability of a player to inflict harm now 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.[22] 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 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

[22] It may seem unreasonable to assume that each harm opportunity inflicts an identical cost to the punishee, but by stringing together an infinite series of such opportunities, it is possible to construct a punishment of any desired level of severity if the future is discounted sufficiently slowly. We are in essence assuming that each period of time gives the same finite intensity of harm opportunity, and that punishments must therefore be composited from the harm opportunities of an infinite number of individuals. Although this is a stylized assumption, it is arguably realistic in the sense that many individuals must generally co-operate in order to implement real-world punishment systems.

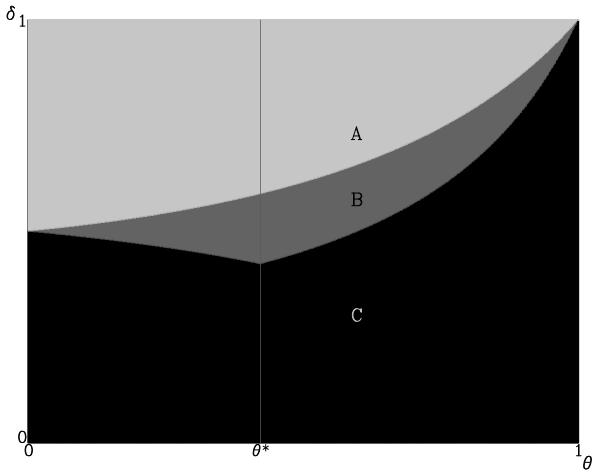

Fig. 1 Socially efficient equilibria (first-best shown in grey) and the first-best over most possible worlds coefficient of altruism θ^*

away because further 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 Proposition 3). More significantly, however, it transpires that too much altruism will, for any value of the discount factor, *prevent* the socially efficient outcome from being achieved (see Proposition 2). Most importantly of all, for low enough values of δ, it will be shown that the values of θ which enable a socially optimal outcome lie between 0 and 1, involving imperfect altruism of the benevolent form ($0 < \theta < 1$) but not perfect altruism ($\theta = 1$) (see Proposition 4).

4 *Defining Social Optima*

Figure 1 provides a preliminary illustrative schematic (using a uniform benefit distribution) for the possible subgame-perfect socially efficient equilibria which are supportable for different values of θ and δ in the sequential punishment game. The most lightly shaded area *A* shows values of (θ, δ) where, by using *"Nash-reversion" punishment* (which requires each individual in an equilibrium where a malefactor is being punished to take their harm opportunity whenever $\pi_t > \theta$, as they would with a one-shot harm opportunity in a single-move game - see section 5) a socially efficient equilibrium can be constructed.[23] The darker grey area *B* shows those values of (θ, δ)

[23] "Nash-reversion" punishment is analogous to Nash-reversion in infinitely-repeated stage games and, as with most infinitely-repeated stage games (such as the Cournot model) there are more severe punishments available via a "carrot-and-stick" approach (Fudenberg and Maskin, 1986; Abreu, 1986)

for which social efficiency can only be supported by using a punishment path more severe than that available with "Nash-reversion" punishment. This requires that the individuals doing the punishing be induced to go "beyond their comfort zone" by inflicting harm in some cases where $\pi_t \leq \theta$ where they would not wish to do so in a single-move game, due to their partial altruism. Recall that we use $\bar{\psi}$ to denote the benefit value above which harm opportunities are taken along a flat equilibrium punishment path (i.e. $T_t = \bar{\psi}$ at any time t along the punishment path). The main analytic task of this paper is to characterise the socially efficient equilibria that can be supported using an optimal flat punishment scheme[24] (conducted in section 6) before generalizing the most important results to the optimal generic punishment scheme (i.e. allowing for *non-flat* punishment paths).

The black region C shows those values of (θ, δ) for which social efficiency is not supportable, even with the use of an optimal punishment path.[25] One of the central results of this paper is that region C is "thinnest" at a *first-best over most possible worlds level of altruism*, θ^* (see definition 5). It will be shown that, under the fairly general assumptions made regarding the distribution of the benefit and the value of θ and δ, it is always the case that $0 < \theta^* < 1$ (see Proposition 4).

There are in fact a number of senses in which we shall use the concept of social efficiency and of a socially optimal level of altruism throughout the paper:

Definition 4 *A* first-best efficient *outcome refers to a subgame-perfect equilibrium where no harm is ever inflicted. Conditional on a particular level of δ, a coefficient of altruism θ is* first-best optimal *if it supports a first-best efficient equilibrium* and *this first-best outcome is robust to an infinitesimal change in θ. There may, given δ, exist no such θ value or a range of such θ values. We can denote the set of such first-best optimal values of θ given δ as $\Theta^*(\delta)$. A value of θ is* second-best optimal *for a given δ if there exists no first-best optimum, and the θ value minimises the range of values of the benefit π for which harm opportunities are taken in the domain $\theta \leq 1$.*

We use π^* to denote the benefit value below which harm opportunities are deterred from being taken in equilibrium. Thus a first-best optimum achieves $\pi^* = 1$ and a second-best optimum maximises the value of π^* at a value less than 1.

Definition 5 *The* first-best over most possible worlds level of altruism *refers to the value of θ which constitutes a first-best optimum for the* widest possible *range of δ. The lowest δ for which a first-best optimum exists is denoted δ^*. A value of θ is* knife-edge optimal *if it supports an efficient outcome but an infinitesimal increase or decrease in θ results in an inefficient outcome whilst an infinitesimal increase in δ maintains an efficient outcome.*

The first-best over most possible worlds coefficient of altruism θ^* will be knife-edge optimal for discount rate δ^*. This is the approach we shall use to prove the existence of θ^* and to derive its characteristics throughout the paper.

whereby altruistic agents implementing a punishment path are persuaded to accept the "stick" of harming another agent when they would prefer not to (due to their altruism) in exchange for the "carrot" of not becoming the agent onto whom all future punishment is "focussed".

[24] See equations (9) through (14) for a formal description.

[25] We later establish, in Proposition 7, that the optimal punishment path must in this uniform distribution case (though *not* in all cases) be flat.

5 *The Single-Move Game*

Consider first a single-move game (this can be thought of as a special case of the infinite-move game in which $\delta = 0$ so that there is no future) in which a single individual (without loss of generality we will label the current time period and the individual who moves 1) has an opportunity to harm another. The individual's altruism level must therefore be sufficiently high in order to prevent them 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.

Proposition 1 *In a single-move game, there is no first-best optimum, and the second-best optimal coefficient of altruism is 1. (If* $\delta = 0$*, then* $\lim_{\theta \to 1^-} \{\pi^*\} = 1^-$*.)*

Proof It is first-best efficient for a harm opportunity to be taken only if $\pi_1 \geq 1$. The individual receiving the harm opportunity (individual 1), meanwhile, will choose to inflict harm if $\pi_1 > \theta$, since they value 1 unit of harm done to another individual at θ. Hence $\pi^* = \theta$. The outcome can therefore only be second-best efficient as $\theta \longrightarrow 1^-$. The outcome is not first-best efficient, however, because an infinitesimal reduction in θ below 1 will result in an inefficient outcome for values of π_1 in the range $\theta < \pi_1 < 1$.

6 *The Infinite-Move Game - Equilibria Using Flat Punishment Paths*

In this section, we proceed to characterise the first-best efficient equilibria which can be supported in the infinite-move sequential punishment game (i.e. where $0 < \delta < 1$) using the optimal flat punishment scheme. Figures 2 through 7 show a variety of different possible shapes to the black area in (θ, δ) space (depending on the benefit distribution $g(\pi)$) where the first-best efficient outcome is not achievable. Propositions 2 through 4 derive the key features of this region in a general manner. The uniform benefit distribution will be further used as a specific illustrative example, but Propositions 2 through 4 apply for any continuously differentiable benefit distribution.

The first key result to be established is that as $\theta \longrightarrow 1^-$ (individuals become perfectly altruistic), the interaction of the three effects leads to a breakdown of the first-best 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 agents 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,

Fig. 2 Socially efficient equilibria for $g(\pi) = 1$

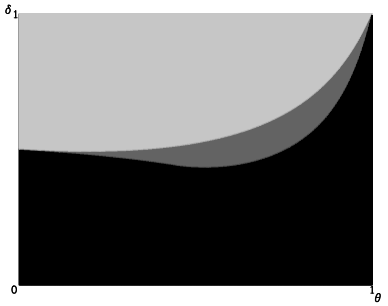

Fig. 3 Socially efficient equilibria for $g(\pi) = 2\pi$

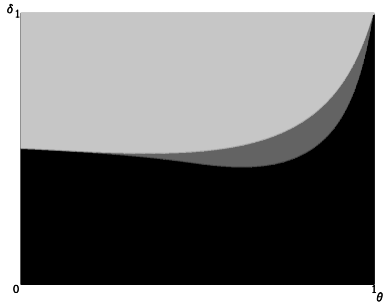

Fig. 4 Socially efficient equilibria for $g(\pi) = 3\pi^2$

Fig. 5 Socially efficient equilibria for $g(\pi) = 4\pi^3$

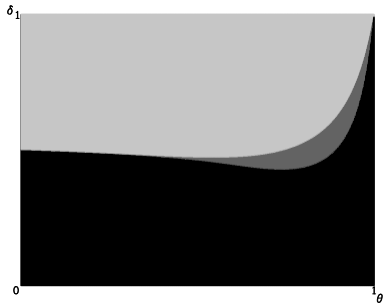

Fig. 6 Socially efficient equilibria for $g(\pi) = 5\pi^4$

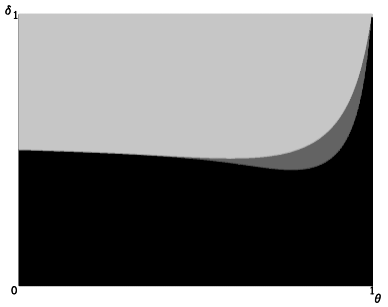

Fig. 7 Socially efficient equilibria for $g(\pi) = 6\pi^5$

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 will occur as $\theta \longrightarrow 0^+$, meaning that social efficiency can be supported with pure self interest. (This phenomenon is driven by the severity effect - see Proposition 3.)

There will in general exist a threshold level of altruism $\widehat{\theta}_{thr}$ (where $0 < \widehat{\theta}_{thr} < 1$), which has the property that it is low enough for the punishment path to involve inflicting harm for the widest possible range of benefit values (i.e. punishment trigger level $\bar{\psi} = 0$), but not any lower, in order to prevent the temptation effect outweighing the severity effect (and thus π^* falling) as θ is further reduced. There will be a corresponding $\widehat{\delta}_{thr}$ (where $0 < \widehat{\delta}_{thr} < 1$) which results in $\bar{\psi} = 0$ for $\delta \geq \widehat{\delta}_{thr}$ if $\theta = \widehat{\theta}_{thr}$. We will show that it must be the case either that the first-best over most possible worlds coefficient of altruism $\theta^* = \widehat{\theta}_{thr}$ (and corresponding $\delta^* = \widehat{\delta}_{thr}$), or that $\theta^* \in \left(\widehat{\theta}_{thr}, 1\right)$ and $\delta^* \in \left(0, \widehat{\delta}_{thr}\right)$. (See Proposition 4.) Hence $\widehat{\theta}_{thr}$ provides a lower bound for θ^*.

An expression for the punishment "trigger level" $\bar{\psi}$ can be found by constructing a condition for co-operation with the punishment path (co-operation meaning "going through with" the punishment) and making it bind when the benefit takes the least attractive value 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. Hence along a punishment path the "target" for punishment A_t is set to be the most recent deviator at all points along the punishment path. A deviation from the punishment path (refusal to inflict harm) at time t will therefore occur if the following inequality holds:

$$\theta - \pi_t \geq \frac{\delta}{1-\delta}\left(\int_{\bar{\psi}}^{1}(1-\theta)g(\pi)d\pi\right) \tag{9}$$

The LHS of (9) is the gain in utility by player t by refusing to take the harm opportunity, which is equal to the utility gain of θ (from avoiding harming another individual whose resultant net felicity gain of 1 is then weighted by θ) minus the utility (and felicity) loss to player t of π_t (due to player t not receiving the benefit from the harm opportunity). The RHS is the expected present discounted value of the infinite stream of utility lost by player t when all future punishment is "focussed" onto them (assuming all future players co-operate by inflicting harm for all realized benefit values where $\pi_t > \bar{\psi}$).

Making (9) bind when $\pi_t = \bar{\psi}$ and rearranging gives us the following:

$$\bar{\psi}(\theta) = \theta - (1-\theta)\frac{\delta}{1-\delta}\int_{\bar{\psi}(\theta)}^{1} g(\pi)d\pi \tag{10}$$

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 $\frac{d\bar{\psi}}{d\theta}$. This determines the direction of the willingness effect - the impact of a change in the coefficient of altruism upon the optimal punishment "trigger level". We should note at this point that as $\theta \longrightarrow 1^-$, this expression becomes unambiguously positive. Hence, as θ increases at an already high level of altruism, then the willingness effect reduces the effectiveness of punishment.

$$\frac{d\bar{\psi}}{d\theta} = \frac{(1-\delta) + \delta\int_{\bar{\psi}}^{1} g(\pi)d\pi}{(1-\delta) - \delta(1-\theta)g(\bar{\psi})} \tag{11}$$

We now apply the following notation (from Definition 3) for the gain in utility from co-operating with the first-best efficient initial path, $\kappa(\bar{\psi}(\theta);\theta)$, and with the punishment path, $\lambda(\bar{\psi}(\theta);\theta)$ (assuming in both cases that the realized benefit value is the one that makes co-operation least attractive, i.e. $\pi_t = 1$ on the initial path and $\pi_t = \bar{\psi}(\theta)$ on the punishment path). For an equilibrium, we require $\kappa(\bar{\psi}(\theta);\theta) \geq 0$ and $\lambda(\bar{\psi}(\theta);\theta) \geq 0$. $U(\bar{\psi}(\theta);\theta)$ represents the per period expected present discounted utility looking forwards along a punishment path which begins next period for an individual who is being punished. $V(\bar{\psi}(\theta);\theta)$ represents the per period expected present discounted utility looking forwards along a punishment path for a "neutral observer" who is not being punished.

$$\begin{aligned} U(\bar{\psi}(\theta);\theta) &= \int_{\bar{\psi}(\theta)}^{1} (\theta\pi - 1)\, g(\pi) d\pi \\ V(\bar{\psi}(\theta);\theta) &= \int_{\bar{\psi}(\theta)}^{1} (\theta\pi - \theta)\, g(\pi) d\pi \end{aligned} \tag{12}$$

$$\lambda(\bar{\psi}(\theta);\theta) = \left(\frac{\delta}{1-\delta}\right) V(\bar{\psi}(\theta);\theta) - \left(\frac{\delta}{1-\delta}\right) U(\bar{\psi}(\theta);\theta) + \bar{\psi}(\theta) - \theta \tag{13}$$

$$\kappa(\bar{\psi}(\theta);\theta) = -\left(\frac{\delta}{1-\delta}\right) U(\bar{\psi}(\theta);\theta) + \theta - 1 \tag{14}$$

If θ is less than δ, individuals will be willing to punish for all possible values of the benefit. This implies that $\lambda(\bar{\psi}(\theta);\theta) \geq 0$ when $\bar{\psi} = 0$ and $\delta \geq \theta$. The intuition for this result is that if individuals are sufficiently patient, the cost of having an infinite stream of future punishment "focussed" upon them is sufficiently great that they will never want to deviate from a punishment path for any possible value of the benefit π_t.

Lemma 1 *If $\theta \leq \delta$ then punishment will occur for all benefit values along the optimal flat punishment path. (If $\theta \leq \delta$ then $\bar{\psi}(\theta) = 0$, otherwise $\bar{\psi}(\theta) = \theta - (1-\theta)\frac{\delta}{1-\delta}\int_{\bar{\psi}}^{1} g(\pi)d\pi$.)*

Proof Substituting (12) into (13) and setting $\lambda(\bar{\psi}(\theta);\theta) \geq 0$ where $\bar{\psi}(\theta) = 0$ gives us $(1-\theta)\frac{\delta}{1-\delta}\int_0^1 g(\pi)d\pi - \theta \geq 0$. Since $\int_0^1 g(\pi)d\pi = 1$, rearranging yields the stated inequality. If $\theta > \delta$, on the other hand, then $\bar{\psi}$ must be where $\lambda(\bar{\psi}(\theta);\theta) = 0$ in an interior solution (i.e. $0 < \bar{\psi}(\theta) < 1$) as described by equation (10).

Having derived the optimal flat punishment path, we are now in a position to characterise the first-best efficient equilibria which can be supported using it. Substituting (12) into (14) gives us the following for $\kappa(\bar{\psi}(\theta);\theta)$, (along with its total derivative with respect to θ, with $\bar{\psi}(\theta)$ suppressed to $\bar{\psi}$ for clarity and decomposed into temptation, willingness and severity effects):

$$\kappa(\bar{\psi}(\theta);\theta) = -\frac{\delta}{1-\delta}\int_{\bar{\psi}(\theta)}^{1} (\theta\pi - 1) g(\pi) d\pi + \theta - 1 \tag{15}$$

$$\frac{d\kappa}{d\theta} = \underbrace{1}_{\substack{\textit{temptation}\\ \textit{effect}}} + \underbrace{\left(-\frac{\delta}{1-\delta}(1-\theta\bar{\psi})\,g(\bar{\psi})\frac{d\bar{\psi}}{d\theta}\right)}_{\substack{\textit{willingness}\\ \textit{effect}}} + \underbrace{\left(-\frac{\delta}{1-\delta}\left(\int_{\bar{\psi}}^{1}\pi g(\pi)d\pi\right)\right)}_{\substack{\textit{severity}\\ \textit{effect}}} \tag{16}$$

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 $\kappa(\bar{\psi}) < 0$).

Proposition 2 *As altruism becomes perfect, the optimal flat punishment path cannot support a first-best efficient equilibrium, for any value of the discount factor. (As* $\theta \longrightarrow 1^-$, $\bar{\psi}(\theta) \longrightarrow 1$, $\kappa(\bar{\psi}(\theta);\theta) \longrightarrow 0$ *and* $\frac{d\kappa}{d\theta} \longrightarrow 1$, *therefore as* $\theta \longrightarrow 1^-$, $\kappa(\bar{\psi}(\theta);\theta) \longrightarrow 0^-$.*)*

Proof As $\theta \longrightarrow 1$, it can be seen from expression (10) that $\bar{\psi} \longrightarrow 1$. The RHS of (15) thus goes to 0. Meanwhile, the RHS of (16) goes to 1. Since $\kappa(\bar{\psi}(\theta);\theta)$ is a continuously differentiable function, it must therefore be the case that $\kappa(\bar{\psi}(\theta);\theta)$ 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 $\bar{\psi} = 0$, so that punishment is occurring for all possible values of the benefit, the severity effect will dominate. This means that as $\theta \longrightarrow 0^+$ social efficiency becomes unambiguously supportable. The following proposition derives the required condition on δ.

Proposition 3 *If* $\delta > \frac{1}{1+\bar{\pi}}$ *then a first-best efficient equilibrium can be supported with pure self interest. (If* $\delta > \frac{1}{1+\bar{\pi}}$ *then* $\kappa(\bar{\psi}(\theta);\theta) > 0$ *as* $\theta \longrightarrow 0^+$.*)*

Proof By Proposition 1, when $\theta \leq \delta$ and so $\bar{\psi} = 0$, there is no further willingness effect and so $\frac{d\bar{\psi}}{d\theta} = 0$. As $\theta \longrightarrow 0^+$, this must occur. Therefore, as can be seen from (15), as $\theta \longrightarrow 0^+$, $\kappa(\bar{\psi}(\theta);\theta) > 0$ provided that $\frac{\delta}{1-\delta} > \frac{1}{\int_0^1 \pi g(\pi)d\pi}$. Letting $\bar{\pi} = \int_0^1 \pi g(\pi)d\pi$ and rearranging yields the stated result.

The intuition for Proposition 3 is 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 of values of δ where too high a level of altruism renders the first-best 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 may potentially cause a breakdown of efficiency. (We will shortly show that $\delta^* < \frac{1}{1+\bar{\pi}}$ - see Proposition 4.)

As discussed in section 4, there are a number of approaches which we can take in defining the socially optimal level of altruism in the sequential punishment game.

In a world in which we were unable to achieve the first-best solution, we could ask what impact a change in the coefficient of altruism has upon the efficiency of the second-best equilibrium. This we do in section 9. For the remainder of this section, however, we continue to concentrate on worlds where the first-best solution may be available, and hence ask what value of θ allows the first-best efficient outcome to be supportable for the widest range of δ. Therefore, although we do later on consider the best we can do in each possible world, we primarily focus here upon the broader and more "philosophical" issue of which coefficient of altruism we would choose as a social planner if we did not know which world (specified entirely by δ) that we would end up in.

The following proposition defines the first-best over most possible worlds level of altruism, θ^* and corresponding minimum δ, δ^*. The optimal coefficient of altruism has a number of key features. Firstly, it must be a "knife-edge" first-best efficient equilibrium so that $\kappa(\bar{\psi}(\theta);\theta) = 0$. 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 (i.e. given Proposition 1 we set $\theta = \delta$). Lemma 2 below will be necessary:

Lemma 2 *If there exist first-best optimal values of θ when $\delta = \delta'$, and all those which exist satisfy the inequality $\theta \geq \theta'$, then there must exist a first-best over most possible worlds level of altruism θ^* such that $\theta' < \theta^* < 1$ and which is knife-edge optimal for a δ^* such that $0 < \delta^* < \delta'$.*

Proof If $\Theta^*(\delta')$ (see Definition 4) is a non-empty and non-singleton set (so that $\delta' > \delta^*$) then if δ is infinitesimally reduced to δ'' then $\{\theta^*\} \subset \Theta^*(\delta'') \subset \Theta^*(\delta')$. Thus, as δ limits to δ^* from above, $\Theta^*(\delta)$ will limit to a singleton value, which we can define as $\theta^* = \lim_{\delta \longrightarrow \delta^{*+}} \{\Theta^*(\delta)\}$. Finally, Proposition 2 implies that $\theta^* < 1$ and Proposition 1 implies that $\delta^* > 0$.

Proposition 4 *The first-best over most possible worlds level of altruism θ^* is always strictly positive and strictly less than 1. ($\theta^* \in [\widehat{\theta}_{thr}, 1)$ and $\delta^* \in \left(0, \widehat{\theta}_{thr}\right]$, where $\widehat{\theta}_{thr} = \frac{3-\sqrt{5-4\bar{\pi}}}{2(1+\bar{\pi})}$.)*

Proof First observe that $\widehat{\theta}_{thr}$ is where both $\kappa(\bar{\psi}(\theta);\theta) = 0$ and $\theta = \delta$. The following two equations must therefore hold simultaneously: (Equation (18) is derived from Proposition 3. Equation (17) is derived from setting equation (15) equal to 0 and plugging in $\bar{\psi} = 0$ and $\bar{\pi} = \int_0^1 \pi g(\pi) d\pi$.)

$$\frac{\delta}{1-\delta} = \frac{1-\theta}{1-\theta\bar{\pi}} \tag{17}$$

$$\theta = \delta \tag{18}$$

Equations (17) and (18) together form a quadratic equation system, yielding the following solution: (Note that the second solution to the quadratic can be discounted since we require that $\widehat{\theta}_{thr} < 1$ in order for (18) to be satisfied with $\widehat{\delta}_{thr} < 1$.)

$$\widehat{\theta}_{thr} = \widehat{\delta}_{thr} = \frac{3-\sqrt{5-4\bar{\pi}}}{2(1+\bar{\pi})} \tag{19}$$

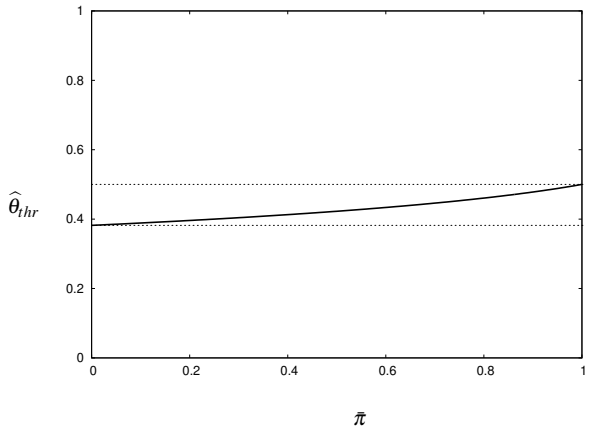

Fig. 8 $\widehat{\theta}_{thr}$ as a function of $\bar{\pi}$

There are now two logical possibilities to consider. The first is that $\Theta^*\left(\widehat{\delta}_{thr}\right)$ is an empty set, in which case $\left(\widehat{\theta}_{thr}, \widehat{\delta}_{thr}\right)$ is knife-edge optimal and $\lim_{\delta \to \left(\widehat{\theta}_{thr}\right)^+} \{\Theta^*(\delta)\} = \tilde{\theta}$, hence $\theta^* = \delta^* = \widehat{\theta}_{thr} = \widehat{\delta}_{thr}$. The second possibility is that $\Theta^*\left(\widehat{\delta}_{thr}\right)$ is non-empty. In that case, applying Lemma 2 with $\theta' = \delta' = \widehat{\theta}_{thr}$, we have $\theta^* \in \left[\widehat{\theta}_{thr}, 1\right)$ and $\delta^* \in \left(0, \widehat{\theta}_{thr}\right]$.

Since $3 - \sqrt{5 - 4\bar{\pi}} < 2$, Propositions 3 and 4 have also together established that $\delta^* < \frac{1}{1+\bar{\pi}}$ 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 when δ is sufficiently close to (but remains above) δ^* (so $\delta^* < \delta < \frac{1}{1+\bar{\pi}}$).

Figure 8 shows the value of $\widehat{\theta}_{thr}$ as a function of $\bar{\pi}$. It can immediately be seen that $\widehat{\theta}_{thr} \in \left[\frac{3-\sqrt{5}}{2}, \frac{1}{2}\right]$. So, the first-best over most possible worlds level of altruism θ^* must be greater than or equal to $\frac{3-\sqrt{5}}{2} \approx 38\%$.

In order to make the welfare economic justification for singling out the first-best over most possible worlds coefficient of altruism θ^* as clear as possible, we now present a final Proposition which elaborates its desirable properties:

Proposition 5 *If a first-best efficient outcome is supportable for any values of θ, then it will be supportable for the first-best over most possible worlds level of altruism θ^*.(If $\Theta(\delta) \neq \emptyset$ then $\theta^* \in \Theta^*(\delta)$.)*

Proof Applying the result from Proposition 4, (θ^*, δ^*) is knife-edge optimal so that $\kappa(\bar{\psi}(\theta); \theta) = 0$ as described by equation (15). If $\Theta(\delta) \neq \emptyset$ then $\delta > \delta^*$. Since the RHS of (15) is increasing in δ then $\theta = \theta^*$ implies $\kappa(\bar{\psi}(\theta^*); \theta^*) > 0$. Thus a first-best outcome is supportable for (θ^*, δ) and so $\theta^* \in \Theta^*(\delta)$.

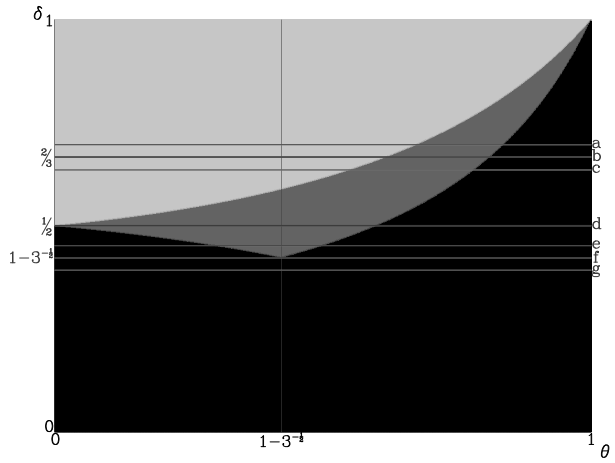

Fig. 9 Socially efficient equilibria where $g(\pi) = 1$ for $0 \leq \pi \leq 1$

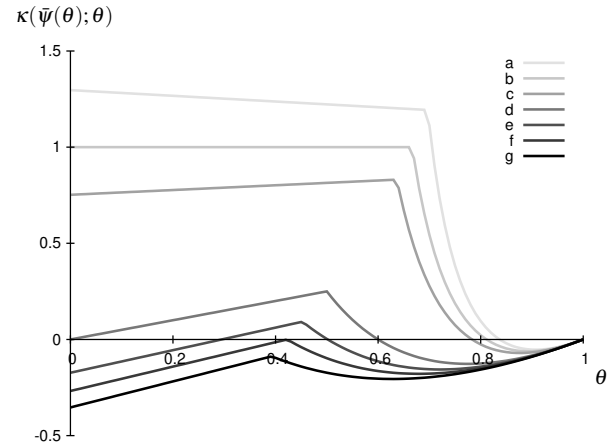

Fig. 10 Values of $\kappa(\bar{\psi}(\theta);\theta)$ in cross-sections *a*-*g*

6.1 Illustration: continuous uniform distribution

Figure 9 illustrates the application of Propositions 2, 3 and 4 to the case of a uniform distribution with support between 0 and 1 and therefore where $\forall_{0\leq\pi\leq1}$: $g(\pi) = 1$ and $\bar{\pi} = 1/2$. The key features are the first-best over most possible worlds level of altruism θ^* and corresponding δ^* (by substituting $\bar{\pi} = 1/2$ into expression (19), we find that this is at $\theta^* = \delta^* = 1 - \frac{1}{\sqrt{3}} \approx 42\%$) and the value of $\delta = \frac{1}{1+\bar{\pi}} = \frac{2}{3}$ above which the severity effect dominates as $\theta \longrightarrow 0^+$. Figure 10 illustrates how the

value of $\kappa(\bar{\psi}(\theta);\theta)$ changes (on the y-axis) for a series of "cross sections" taken through figure 9 (labelled *a*-*g*) where δ is fixed and θ is allowed to vary along the x-axis. When $\delta = 1 - \frac{1}{\sqrt{3}}$,[26] it can be seen that $\kappa(\bar{\psi}(\theta);\theta)$ lies below 0 for any value of θ apart from $\theta^* = 1 - \frac{1}{\sqrt{3}}$ and $\theta = 1$.[27] Another key "cross-section" is at the value of $\delta = 1/2$.[28] The significance of this point is that it is where the co-operation constraint for the optimal "Nash-reversion" punishment path 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). A third important value is $\delta = \frac{1}{1+\bar{\pi}} = \frac{2}{3}$.[29] The unique property of this particular value of δ is that the severity and temptation effects exactly cancel, as shown by the fact that the value of $\kappa(\bar{\psi}(\theta);\theta)$ is constant for any $\theta < \delta$. It is instructive to compare this to values of δ slightly above and below $2/3$ respectively.[30] Here we notice that the value of $\kappa(\bar{\psi}(\theta);\theta)$ 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 $\delta < \frac{2}{3}$. Cross-sections are also shown for δ values slightly above and below $1 - \frac{1}{\sqrt{3}}$.[31] The main feature to note here is that when δ is below $1 - \frac{1}{\sqrt{3}}$, the only value of θ for which $\kappa(\bar{\psi}(\theta);\theta)$ is not negative is 1.

7 *Quasi-Flat Paths*

The next two sections will primarily be concerned with extending the result from Proposition 2 to the general case where the optimal punishment path is not restricted to be flat. We begin by characterizing the equilibria supportable using the optimal *quasi-flat* punishment path. We then proceed, in section 8, to establish that, as $\theta \longrightarrow 1^-$, the socially efficient equilibrium becomes unsupportable even using the optimal *generic* punishment path.

Definition 6 *A quasi-flat path is one which is flat from point 2 onwards, and is denoted by $\tilde{\psi}$. The point 1 trigger level is $\tilde{\psi}_1$. The point 2 and after trigger level is $\tilde{\psi}_2$.*[32]

An important concept that will be used repeatedly in the lemmas and propositions 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 punishee and for a "neutral observer".

[26] This corresponds to line f in figure 9.

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

[28] This corresponds to line d in figure 9.

[29] This corresponds to line b in figure 9.

[30] These correspond to lines a and c in figure 9.

[31] These correspond to lines e and g in figure 9.

[32] This is the simplest punishment path 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.

Definition 7 *Let the U-average and the V-average be respectively denoted as* $U^{-1}(U_k(\psi))$ *and* $V^{-1}(V_k(\psi))$. *These two averages are defined 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* $U^{-1}(U_k(\psi))$.

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

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

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

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

$$\frac{\frac{d}{d\psi_{i+k}}V^{-1}(V_k(\psi))}{\frac{d}{d\psi_{i+k}}U^{-1}(U_k(\psi))} = \left(\frac{1-\psi_{i+k}}{1-\theta\,\psi_{i+k}}\right)\left(\frac{g\left(U^{-1}(U_k(\psi))\right)}{g(V^{-1}(V_k(\psi)))}\right)\left(\frac{1-\theta\,U^{-1}(U_k(\psi))}{1-V^{-1}(V_k(\psi))}\right) \tag{20}$$

The following two lemmas will prove very useful in this and subsequent sections. Importantly, Lemma 3 applies to all optimal punishment paths, not just quasi-flat ones. It states that U_k must we weakly minimized at point 0 along an optimal path. Lemma 4 states that constraint (5) must bind at point 1 along an optimal quasi-flat path.[33]

Lemma 3 *The U-average must be weakly minimized at the beginning of an optimal punishment path. ((a) If a punishment path* ψ^* *is optimal then* $\forall_k : U_k(\psi^*) \geq U_0(\psi^*)$. *(b) If a punishment path* ψ^* *is optimal then* $\psi_1^* \leq U^{-1}(U_0(\psi^*)) \leq U^{-1}(U_1(\psi^*))$.*)*

[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.

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:

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

$$\lambda_i(\psi') = \mu_i(\psi') = \left(\frac{\delta}{1-\delta}\right) V_i(\psi') - \left(\frac{\delta}{1-\delta}\right) U_0(\psi') - \psi'_i + \theta$$

These can be rewritten as:

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

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

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

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

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

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:

$$\begin{aligned} &\frac{\delta}{1-\delta}\left(\int_{U^{-1}(U_0(\psi))}^{1} (\theta\pi - 1) g(\pi) d\pi\right) = \\ &\delta \int_{\psi_1}^{1} (\theta\pi - 1) g(\pi) d\pi + \frac{\delta^2}{1-\delta} \int_{U^{-1}(U_1(\psi))}^{1} (\theta\pi - 1) g(\pi) d\pi \end{aligned} \tag{21}$$

This can be rewritten as:

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

Since we know from the argument made above that $U_0(\psi^*) \leq U_1(\psi^*)$, 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 (21) can also be rewritten as:

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

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 3 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 $\forall_{\tilde{\psi}^*} : \tilde{\psi}_1^* \leq \tilde{\psi}_2^*$. 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 $\tilde{\psi}$, $\forall_{k\geq 1} : \tilde{\psi}_2 = V^{-1}(V_k(\tilde{\psi})) = U^{-1}(U_k(\tilde{\psi}))$.

Lemma 4 *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 $\lambda_1(\tilde{\psi}^*) = 0$.*

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 path. $\frac{d\phi}{d\tilde{\psi}_1} = -\delta(1-\theta\tilde{\psi}_1)g(\tilde{\psi}_1)$ is also negative, so 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 0 before the period 1 constraint binds. Since $\tilde{\psi}_2 = U^{-1}(U_1(\tilde{\psi}))$, Lemma 3 implies that $\tilde{\psi}_2^* \geq \tilde{\psi}_1^*$. This 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 $\tilde{\psi}_1^* = 0$, 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.

7.1 A taxonomy of optimal quasi-flat paths

The general 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 game. Although they all 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 analyzing the equilibria supportable by optimal generic punishment paths in the sequential punishment game. This is because, as will be shown in section 8, it is always possible to construct a quasi-flat path whose severity (ϕ) value forms an upper bound for all sustainable generic paths, and to derive sufficient structure from this to extend Proposition 2 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 game, 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.

The positioning of trigger levels in a quasi-flat path involves a key trade-off, allowing greater punishment to be "bought" at point 1, but at the cost of less severe punishment at later points in the path. The "sacrifice ratio" will be given by expression (20), which measures the increase in "carrot" (measured as a higher V-average) for a

given reduction in severity of 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. (The benefits and costs are discounted, so the exchange ratio, given a particular trigger level, looks the same for all future periods.) It is therefore optimal to "spread out" the punishment evenly over the entire tail of the path.

The gain in severity of the punishment path from a reduction in the point 1 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 2 and after trigger level. The *cost* of incentivizing co-operation with a lower trigger level at point 1, however, does *not* depend upon the probability density at the point 1 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 then this cost will always outweigh the benefit of making the quasi-flat punishment path non-flat.

There are a number of possibilities for the precise form that the optimal quasi-flat path might take. By Lemma 3, it is impossible for the trigger level at point 1 to be higher than at point 2 onwards. Also, the optimal path cannot possibly be "fully-minimal" (i.e. exhibit $\forall_k : \psi_k = 1$ because it can be seen from Lemma 1 and condition (10) that there will always exist a sustainable, and more severe, flat path.[34] This then leaves six logical possibilities, types A-F, illustrated in figures 11 through 16. Type A is the *maximal* path characterized in Lemma 1. 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. The next type, D, 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. This happens because we reach the top of the support of the distribution for π (i.e. $\tilde{\psi}_2$ reaches 1). We shall call this a *quasi-minimal* path. With type E paths, on the other hand, the amount of punishment at point 1 runs up against constraint (6) 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 sixth possibility, F, is that there is an optimal marginal trade-off between punishment at point 1 and at point 2 and after. We shall call this a *carrot-maximized* path.

We shall also find it essential in the lemmas and propositions to follow to distinguish between two different types of optimal path. *Fully-constrained paths* must satisfy all co-operation conditions defined by (5) and (6). *Semi-constrained paths* only need satisfy the conditions defined in (5). 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).

[34] Lemma 1 implies that there will always exist an optimal flat path with $\bar{\psi}^* < 1$, which is strictly more severe than the "fully-minimal" path (where $\bar{\psi}^* = 1$). (This is unless $\theta = 1$, which is ruled out by assumption.)

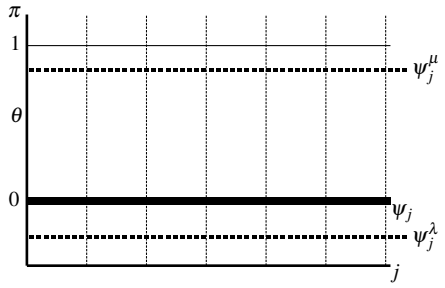

Fig. 11 A - A maximal path

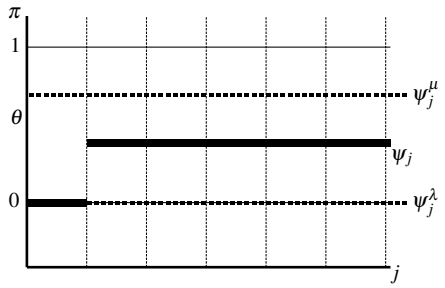

Fig. 12 B - A quasi-maximal path

Definition 8 *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 $\bar{\Psi} \subset \tilde{\Psi}$ and 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. Other optimal paths can be defined using these sets. So, for example, $\tilde{\psi}^*_{sc} = \sup \tilde{\Psi}_{sc} = \max_{\psi \in \tilde{\Psi}_{sc}} \{\phi(\psi)\}$ is 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 (6) does not apply. Also, observe that the optimal semi-constrained path will be at least as severe 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, $\phi(\sup \tilde{\Psi}_{sc}) \geq \phi(\sup \tilde{\Psi}_{fc})$. This observation is key in enabling Proposition 2 to be generalized to the case where the optimal generic path is used to punish deviations from the socially efficient initial path.

Definition 9 *Define ψ_k^{λ} and ψ_k^{μ} as the values of ψ_k that satisfy* (5) *and* (6) *with equality. ψ_k^{λ} and ψ_k^{μ} 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. Note that, for a quasi-flat path, they will be the same for any k because $\forall_k$: $V_k(\tilde{\psi}) = V_1(\tilde{\psi})$.*

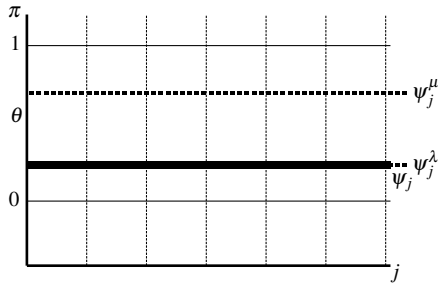

Fig. 13 C - A flat path

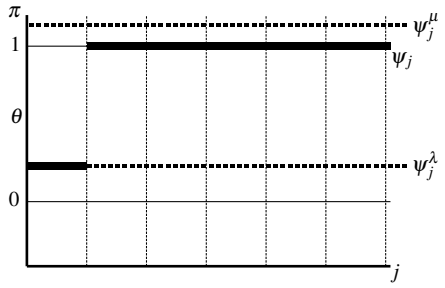

Fig. 14 D - A quasi-minimal path

Propositions 1, 3 and 4 continue to hold for the equilibria supportable by generic optimal paths (and, therefore, quasi-flat paths) without alteration. As we shall argue shortly, in general the quasi-maximal case only occurs when θ is close to the boundary established in Lemma 1 below which the optimal path is maximal. We shall also see that the quasi-minimal case cannot possibly be optimal. The flat, carrot-constrained and carrot-maximized paths illustrated in figures 13, 15 and 16 therefore respectively represent the three possibilities for an interior solution.

7.2 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 $\bar{U} = (\frac{\delta}{1-\delta})U_0\left(\underline{\psi}\right)$ be the expected utility for the punishee along the fixed path $\underline{\psi}$. Let $\bar{\lambda}_k(\tilde{\psi}) = (\frac{\delta}{1-\delta})V_k(\tilde{\psi}) - \bar{U} + \tilde{\psi}_k - \theta$ be the co-operation constraint at point k for quasi-flat path $\tilde{\psi}$ given the use of fixed path $\bar{\psi}$ to punish a deviation. Since the trigger level at point 1 should be set so that $\bar{\lambda}_1(\tilde{\psi}) = 0$, we know that the following condition must hold:

$$\tilde{\psi}_1 = \theta - \left(\frac{\delta}{1-\delta}\right)V_1(\tilde{\psi}) + \bar{U} = \theta + \left(\frac{\delta}{1-\delta}\right)\int_{\tilde{\psi}_2}^{1}(\theta - \theta\pi)g(\pi)d\pi + \bar{U} \qquad (22)$$

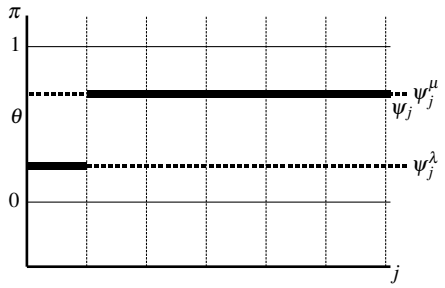

Fig. 15 E - A carrot-constrained path

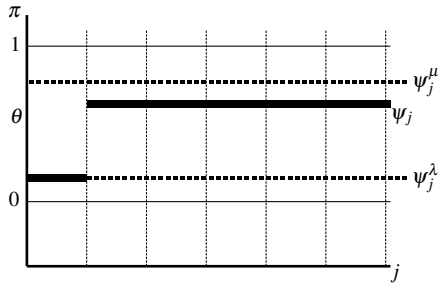

Fig. 16 F - A carrot-maximized path

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:

$$\phi = -\left(\frac{\delta}{1-\delta}\right) U_0(\tilde{\psi}) = \delta \int_{\tilde{\psi}_1}^{1} (1-\theta\pi)\, g(\pi) d\pi + \left(\frac{\delta^2}{1-\delta}\right) \int_{\tilde{\psi}_2}^{1} (1-\theta\pi)\, g(\pi) d\pi \tag{23}$$

Totally differentiating (22) gives us:

$$\frac{d\tilde{\psi}_1}{d\tilde{\psi}_2} = -\left(\frac{\delta}{1-\delta}\right)(\theta - \theta\tilde{\psi}_2)\, g(\tilde{\psi}_2) \tag{24}$$

Totally differentiating (23) with respect to $\tilde{\psi}_2$ gives us:

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

Substituting in (24) and simplifying yields:

$$\frac{d\phi}{d\tilde{\psi}_2} = \left(\frac{\delta^2}{1-\delta}\right) g(\tilde{\psi}_2)\left((1-\theta\tilde{\psi}_1)(\theta - \theta\tilde{\psi}_2)\, g(\tilde{\psi}_1) - (1-\theta\tilde{\psi}_2)\right) \tag{25}$$

This is unambiguously negative if and only if the following condition holds:

$$g(\tilde{\psi}_1) < \frac{1-\theta\tilde{\psi}_2}{(1-\theta\tilde{\psi}_1)\,\theta\,(1-\tilde{\psi}_2)} \tag{26}$$

The RHS of the above expression is increasing in $\tilde{\psi}_1$ and $\tilde{\psi}_2$. This means that the most stringent condition will be where $\tilde{\psi}_1 = \tilde{\psi}_2 = 0$. 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:

$$\forall_\pi : g(\pi) < \frac{1}{\theta} \tag{27}$$

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 3. Condition (27) is therefore sufficient for the optimal quasi-flat path to be flat. (This result is extended to a generic punishment path in Proposition 7 in section 8.)

A *necessary* condition for the optimal quasi-flat path to be flat (or rather for a particular flat path to be optimal in the set of quasi-flat paths) can be found by substituting $\tilde{\psi}_1 = \tilde{\psi}_2 = \bar{\psi}^*$ into condition (26) to give the following:

$$g\left(\bar{\psi}^*\right) < \frac{1}{\theta\left(1 - \bar{\psi}^*\right)} \tag{28}$$

7.3 Conditions for non-flat paths

We now proceed to lay out the conditions which must hold for the various possible configurations of a non-flat fully-constrained optimal quasi-flat path. We already know the necessary and sufficient condition for a maximal (type A) path from Proposition 1. Next, taking the case of a quasi-maximal (type B) path, we know that $\tilde{\psi}_1 = 0$. 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.

For a carrot-constrained (type E) path, constraint (5) binds at point one and constraint (6) 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.[35] The optimal carrot-constrained path is characterized 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 constraints (5) and (6)) gives us:

$$\tilde{\psi}_2 = 2\theta - \tilde{\psi}_1 \tag{29}$$

The fact that a carrot-maximized (type F) path is also a possibility can be seen by observation of condition (26). As $\tilde{\psi}_2 \longrightarrow 1$, the RHS of (26) 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 thus have the property that condition (26) is satisfied with equality. Rearranging this yields:

[35] 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.

$$\tilde{\psi}_2 = \frac{g(\tilde{\psi}_1)(1-\theta\tilde{\psi}_1)\theta - 1}{g(\tilde{\psi}_1)(1-\theta\tilde{\psi}_1)\theta - \theta} \tag{30}$$

The above argument from condition (26) also shows why a quasi-minimal (type D) path is impossible, since as $\tilde{\psi}_2 \longrightarrow 1^-$ the RHS of the inequality goes to ∞. Thus, 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.

Lemma 5 *(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 $\theta \longrightarrow 1^-$, $\tilde{\psi}^*_{sc} \in \bar{\Psi}$. (b) If $\forall_\pi : g(\pi) < \frac{1}{\theta}$ then $\tilde{\psi}^*_{sc} \in \bar{\Psi}$.)*

Proof As already argued in subsection 7.3, the possible types of quasi-flat path are exhaustively categorized by types *A*-*F* from the taxonomy in subsection 7.1. Also, we have already seen from (25) that *quasi-minimal* type *D* paths are not possible. Since we are only considering *semi-constrained* quasi-flat paths, type *E* (carrot-constrained) paths are also not possible.

In order to prove claim (a), we need to show that, as $\theta \longrightarrow 1^-$, types *A*, *B* and *F* are also impossible, leaving type *C* (flat) paths as the only possibility. Consider firstly type *A* (maximal) paths. From Lemma 1, we know that the optimal path will be maximal if and only if $\theta \leq \delta$. This cannot possibly occur as $\theta \longrightarrow 1^-$. Moving to type B (quasi-maximal) paths, by Lemma 4 these can only be optimal if $\lambda_1(\tilde{\psi}^*) = 0$ holds where $\tilde{\psi}^*_1 = 0$. As $\theta \to 1^-$ this equation limits to yield $-\delta \int_{\tilde{\psi}_2}^1 (1-\pi) g(\pi) d\pi = 1 - \delta(1-\bar{\pi})$, which is impossible since the RHS is unambiguously positive whilst the LHS is unambiguously negative. Finally, type F (carrot-maximized) paths are also not possible as $\theta \to 1^-$ because condition (30) limits to yield $\tilde{\psi}_2 = 1$ and so the $\lambda_1(\tilde{\psi}^*) = 0$ condition simplifies to give $1 - \tilde{\psi}_1 = \delta \int_{\tilde{\psi}_1}^1 (1-\pi) g(\pi) d\pi$, for which the only possible solution is $\tilde{\psi}_1 = 1$, yielding a flat path (which is also "fully-minimal" in the limit).

Additionally, for claim (b), note that Lemmas 3 and 4 together imply that condition (27) is sufficient for types *B* and *F* to be impossible, given the derivative of the severity, $\phi(\tilde{\psi})$ function as derived in equation (25). Unless $\theta \leq \delta$, therefore, the optimal path must be flat (type *C*).

8 *The Optimal Generic Path*

We are now ready to generalize Proposition 2 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 5), and we already know (from Proposition

2) 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 generic path, then it must also follow that, as $\theta \longrightarrow 1^{-}$, the socially efficient equilibrium cannot be supported by *any* sustainable path.

Proposition 6 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[36] $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$. It will therefore continue to be in the set of semi-constrained paths, and will be as severe as the path it was generated from.

Definition 10 *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 $\phi(\gamma(\psi)) = \phi(\psi)$.*[37]

Applying Lemma 5 from the previous section[38], 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 the optimal path is flat in a particular case, it is thus established that intermediate values of the coefficient of altruism are best able to allow a socially efficient equilibrium to be supported using globally optimal punishment paths. Proposition 6 will, as a result, be crucial to establishing the general applicability of the key results from section 6, which form the core contribution of this paper.

Lemma 6 below is an essential building block necessary for the proof of Proposition 6. It establishes that for any generic punishment path looking forward from any point k, the U-average, $U^{-1}(U_k(\psi))$, *must* lie equal to or above the V-average, $V^{-1}(V_k(\psi))$. It should be noted that the result from Lemma 6 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 variability of benefit values under which punishment is inflicted. The sensitivity of a "neutral observer" to the "wastefulness" of 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 altruists 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. This result should therefore have wider implications in other models involving altruism and punishment.

Lemma 6 *For any punishment path, at any point, the U-average is weakly greater than the V-average. ($\forall_{\psi}\forall_k : U^{-1}(U_k(\psi)) \geq V^{-1}(V_k(\psi))$.)*

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

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

[38] In Lemma 5, we impose only that the optimal path be semi-constrained, partly to simplify the proof but, more importantly, because we are able to generate semi-constrained quasi-flat paths by "flattening-out" generic paths.

Proof This result follows from an application of the principle of stochastic dominance, a key principle in the economics of risk (Rothschild and Stiglitz, 1970). First, note from definition 3 that the per period average utility functions for the person being punished and a neutral observer can be rewritten as $U_k(\psi) = \sum_{i=1}^{\infty}[p_i U(\psi_{i+k})]$ and $V_k(\psi) = \sum_{i=1}^{\infty}[p_i V(\psi_{i+k})]$ 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 and Stiglitz, 1974). For the two types of agent under consideration (with utility functions $U(\psi)$ and $V(\psi)$ respectively), the CARA works out as the following.

$$R_a^U = \frac{\theta}{1-\theta\psi} - \frac{g'(\psi)}{g(\psi)} \tag{31}$$

$$R_a^V = \frac{\theta}{\theta-\theta\psi} - \frac{g'(\psi)}{g(\psi)} \tag{32}$$

Since (32) is always unambiguously greater than (31), 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 3, is precisely analogous to $V^{-1}(V_k(\psi))$ (as opposed to $U^{-1}(U_k(\psi))$ for the punishee).

The intuition for the result in Lemma 6 is also closely related to the "sacrifice ratio" derived in expression (20). 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. Given a particular value of $U^{-1}(U_k(\psi))$, $V^{-1}(V_k(\psi))$ is maximized when this is generated by a flat path looking forwards.

Lemma 7 *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 6, $U_0(\gamma(\psi^*)) = U_0(\psi^*)$ and $\forall_k : V_k(\gamma(\psi^*)) \geq V_1(\psi^*)$. To see this, note the following:

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

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

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

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

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 6, $U^{-1}(U_1(\psi)) \geq V^{-1}(V_1(\psi))$.

Now we can proceed to note that:

$$\lambda_1(\gamma(\psi)) = \left(\frac{\delta}{1-\delta}\right)V_1(\gamma(\psi)) - \left(\frac{\delta}{1-\delta}\right)U_0(\gamma(\psi)) + \gamma_1(\psi) - \theta \tag{33}$$

$$\forall_{k\geq 2} : \lambda_k(\gamma(\psi)) = \left(\frac{\delta}{1-\delta}\right)V_k(\gamma(\psi)) - \left(\frac{\delta}{1-\delta}\right)U_0(\gamma(\psi)) + \gamma_2(\psi) - \theta \tag{34}$$

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

$$\lambda_1(\psi^*) = \left(\frac{\delta}{1-\delta}\right)V_1(\psi^*) - \left(\frac{\delta}{1-\delta}\right)U_0(\psi^*) + \psi_1^* - \theta \geq 0 \tag{35}$$

Given the result from Lemma 3 that $\psi_1^* \leq U^{-1}(U_1(\psi^*))$, along with all the observations noted so far, condition (35) is sufficient for (33) and (34) to be weakly positive for all relevant k.

Lemma 8 *The optimal semi-constrained quasi-flat punishment path is at least as severe as the optimal fully-constrained generic punishment path. ($\phi(\tilde{\psi}_{sc}^*) \geq \phi(\psi^*)$.)*

Proof Consider the optimal generic path $\psi^* \in \Psi_{fc}$. By Lemma 7, $\gamma(\psi^*) \in \tilde{\Psi}_{sc}$. Also, by definition, $\phi(\gamma(\psi^*)) = \phi(\psi^*)$. 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 $\phi(\tilde{\psi}_{sc}^*) \geq \phi(\psi^*)$.

Proposition 6 *As altruism becomes perfect, the optimal generic punishment path cannot support the socially efficient equilibrium, for any value of the discount factor. (As $\theta \longrightarrow 1^-$, $\kappa(\psi^*) < 0$.)*

Proof First, note, from expressions (7) and (8) that:

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

By Lemma 8, this implies that:

$$\kappa(\psi^*) \leq \phi(\tilde{\psi}_{sc}^*) + \theta - 1 \tag{36}$$

Proposition 2, combined with Lemma 5, has already established that the RHS of expression (36) goes to 0^- as $\theta \longrightarrow 1^-$. Therefore the LHS of (36) must be strictly negative as $\theta \longrightarrow 1^-$.

Proposition 7 *If the benefit distribution is sufficiently flat, then the optimal punishment path is flat. (If $\forall_\pi : g(\pi) < \frac{1}{\theta}$ then $\phi(\bar{\psi}^*) \geq \phi(\psi^*)$.)*

Proof From Lemma 5, we know that, if $\forall_\pi : g(\pi) < \frac{1}{\theta}$, then $\tilde{\psi}_{sc}^* \in \bar{\Psi}_{fc}$. Therefore, $\phi(\bar{\psi}^*) \geq \phi(\tilde{\psi}_{sc}^*)$. Applying Lemma 8, the result follows straightforwardly.

9 *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 first-best efficient initial path will not be supportable. There will still, however, exist a second-best optimal equilibrium, supported by the optimal flat punishment path, in the sense that the associated initial path maximizes efficiency by minimizing the range of benefit values for which harm opportunities are taken.

Along the second-best optimal initial path, the punishment path will be initiated when an individual punishes for a value of the benefit below benefit level π^*. (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 using equation (10). So, π^* and its total derivative with respect to θ will be described by the following equations:

$$\begin{aligned} \pi^*(\theta) &= \theta + \frac{\delta}{1-\delta}\left(\int_{\bar{\psi}(\theta)}^{1}(1-\theta\pi)g(\pi)d\pi\right) \\ \frac{d\pi^*}{d\theta} &= \underbrace{1}_{\substack{\textit{temptation}\\ \textit{effect}}} + \underbrace{\left(-\frac{\delta}{1-\delta}(1-\theta\bar{\psi})\,g(\bar{\psi})\frac{d\bar{\psi}}{d\theta}\right)}_{\substack{\textit{willingness}\\ \textit{effect}}} \\ &\quad + \underbrace{\left(-\frac{\delta}{1-\delta}\int_{\bar{\psi}}^{1}\pi g(\pi)d\pi\right)}_{\substack{\textit{severity}\\ \textit{effect}}} \end{aligned} \tag{37}$$

Figure 17 shows the highest π^* which is supportable along the initial path given different values of θ (along the x-axis) and δ. It can be seen (and is proven in the final Proposition 6 below, which characterises the properties of second-best equilibria) that the curve always has a slope of 1 as $\theta \longrightarrow 1^-$ and that the gradient is always positive for all θ when δ is low but is sometimes negative when δ is high. Importantly, there is always a level of θ high enough but lower than 1 where π^* falls below one and then back up to one as $\theta \longrightarrow 1$. This corresponds to the black region in figure 9 and derived analytically in Propositions 2 through 4. Finally, for high enough δ with a low enough θ, π^* goes above one (i.e. the graph gets "cut off"). This is corresponds to the region where the first-best efficient equilibrium is supportable.

Lemma 9 *In a second-best equilibrium, if* $\frac{d\pi^*}{d\theta} > 0$ *then* $\frac{dW}{d\theta} > 0$*, if* $\frac{d\pi^*}{d\theta} < 0$ *then* $\frac{dW}{d\theta} < 0$*.*

Proof Social welfare (simple sum of felicities) in any given time period on the initial path is given by $W = \int_{\pi^*(\theta)}^{1}(\pi - 1)g(\pi)d\pi$. Differentiating with respect to θ yields $\frac{dW}{d\theta} = (1-\pi^*)\frac{d\pi^*}{d\theta}$. Since $\pi^* < 1$ in a second-best equilibrium, the result follows straightforwardly.

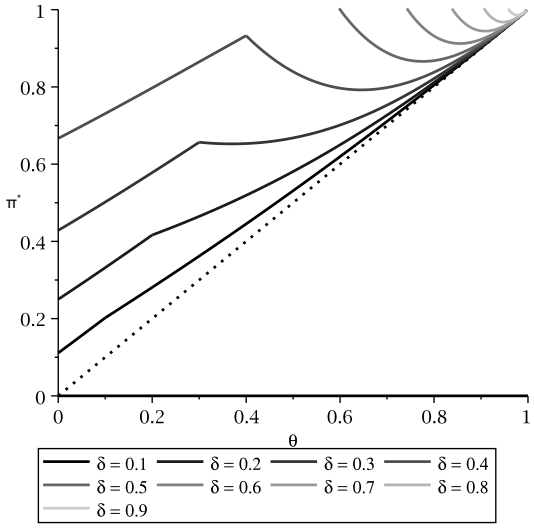

Fig. 17 Second-best equilibria where $g(\pi) = 1$ for $0 \leq \pi \leq 1$

Proposition 8 *In a second-best equilibrium, social welfare is always rising in θ for values of θ close enough to* 1. *In a second-best equilibrium, provided that $\delta < \frac{1}{1+\bar{\pi}}$, social welfare is rising in θ for $\theta < \delta$. In a second-best equilibrium, provided that $\delta < \frac{1}{1+\bar{\pi}}$ and $g(0)$ is sufficiently high, social welfare is falling in θ for intermediate values of θ.*

(As $\theta \longrightarrow 1$, $\frac{dW}{d\theta} > 0$. If $\delta < \frac{1}{1+\bar{\pi}}$ and $\theta < \delta$ then $\frac{dW}{d\theta} > 0$. If $\theta = \delta$ and $g(0) > \left(\frac{1-\delta}{\delta}\right)\left(\frac{1-\delta(1+\bar{\pi})}{1+(1-\delta)(1-\delta(1+\bar{\pi}))}\right)$ then $\frac{dW}{d\theta} < 0$.)

Proof For the first claim: From equation (37), as $\theta \longrightarrow 1^{-}$, $\bar{\psi} \longrightarrow 1$ (as can be seen from (10)) and so $\frac{d\pi^*}{d\theta} \longrightarrow 1$. By Lemma 9, the sign of the derivative of social welfare is the same as $\frac{d\pi^*}{d\theta}$.

For the second claim: By Lemma 1, if $\theta < \delta$ then there is no willingness effect in equation (37). Since in that case $\bar{\psi} = 0$, if $\delta < \frac{1}{1+\bar{\pi}}$ (with $\bar{\pi} = \int_0^1 \pi g(\pi) d\pi$) then the temptation effect dominates the severity effect in equation (37) and so $\frac{d\pi^*}{d\theta} > 0$. By Lemma 9, the sign of the derivative of social welfare is the same as $\frac{d\pi^*}{d\theta}$.

For the third claim: If $\theta = \delta$ then $\bar{\psi} = 0$. Therefore, from equation (11), $\frac{d\bar{\psi}}{d\theta} = \frac{1}{(1-\delta)(1-\delta g(0))}$. Substituting this into equation (37) yields: $\frac{d\pi^*}{d\theta} = 1 - \frac{\delta}{1-\delta}\left(\frac{g(0)}{(1-\delta)(1-\delta g(0))} + \bar{\pi}\right)$. Assuming that $\delta < \frac{1}{1+\bar{\pi}}$, in order to have $\frac{d\pi^*}{d\theta} < 0$, we need: $g(0) > \left(\frac{1-\delta}{\delta}\right)\left(\frac{1-\delta(1+\bar{\pi})}{1+(1-\delta)(1-\delta(1+\bar{\pi}))}\right)$. By Lemma 9, the sign of the derivative of social welfare is the same as $\frac{d\pi^*}{d\theta}$.

An important conclusion to draw from figure 17 is that the efficiency loss from too high a level of altruism can be non-negligible. Although as $\theta \longrightarrow 1$, $\pi^* \longrightarrow 1$, 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 markedly lower. Altruism is in many realistic cases a "double-edged sword" in the sequential punishment game, and too high a level of altruism will in general be significantly socially detrimental.

Proposition 8 shows that there will exist in some (though not all) second-best situations a local maximum of social welfare for a value of $\theta < 1$ (i.e. where $\theta = \delta$ if $g(0) > \left(\frac{1-\delta}{\delta}\right)\left(\frac{1-\delta(1+\bar{\pi})}{1+(1-\delta)(1-\delta(1+\bar{\pi}))}\right)$ though it is possible in other situations to have $\theta > \delta$ and $\frac{d\pi^*}{d\theta} = 0$ at a second-best optimum[39]). However, the global second-best optimum will always be where $\theta \longrightarrow 1^-$, showing that *if* a first-best outcome were impossible, it would be second-best optimal to make the coefficient of altruism as high as possible.

10 *Conclusion*

This paper has taken two areas of economic theory, the modelling of altruistic preferences and the structure of optimal punishment in dynamic games, and shown how they can interact to produce interesting results in a new model, the sequential punishment game - a simple infinite-move sequential game with perfect information and discounting, where players move by choosing whether or not to take opportunities to benefit themselves by inflicting harm upon others. The model is intended as an abstract representation of a wide variety of different possible human interactions occurring in many types of society with varying 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.

The results of the theoretical analysis in this paper could be applied in a number of contexts. Firstly, the temptation, willingness and severity effects generating these results in the sequential punishment game should occur in other more specific policy models, such as in applications to optimal taxation theory (the three effects might be a useful way to understand how the income and substitution effects of different types of taxes are determined under different forms of altruism) and the economics of crime and punishment (in particular, it would be interesting to explore the interaction between altruism, the severity effect and the optimality or otherwise of "wasteful" punishments such as vengeance).

Secondly, some light is shed on the apparent historical connection between the evolution of the state into increasingly sophisticated forms and the development of

[39] In particular, this can occur when there is a non-uniform benefit distribution, as illustrated in figures 2 through 7.

the moral norms of the market, in that extrinsic inventive mechanisms have been shown to work better when agents are at least partly intrinsically self-interested.[40]

Thirdly, if one recognises that altruism is at least partly socially determined and therefore malleable, the results of the sequential punishment game reinforce a normative warning - echoing those in the existing literature (Stark, 1989) - against a naive attempt to try to make society better by making individual people "nicer". Totalitarian and utopian ideologies that seek to remake human nature might well cause grave harm by upsetting the delicate institutional balance between the intrinsic and extrinsic incentives that sustain the existing social order.

Fourthly, broadening the interpretation of the players in the sequential punishment game to include multi-individual organisations, the possibility is suggested that an efficient international order (for instance in issues of conflict avoidance or in international agreements such as those to limit climate change) might in fact (and counter-intuitively) be more easily achievable if individual states are partly motivated by self-interest rather than entirely by the global interest. On the other hand, co-operation between firms in order to collude at the expense of the consumer is suggested to be potentially made easier rather than harder by the fact that the firms are (at least to some degree) motivated by individual profit rather than purely by joint profits. This might have implications for competition policy, in that moderate and imperfect levels of inter-firm "altruism" (e.g. family or kinship connections) might be even more conducive to collusion than more extreme cases of inter-firm "altruism" such as horizontal mergers (Malueg, 1992).[41]

Finally, we should briefly consider the generality of the modelling assumptions that have been made and the game theoretic solution concepts employed. The non-flat punishment paths analyzed in sections 7 and 8 of this paper are arguably fairly rich in that they capture the idea that those who are required to punish and are sufficiently altruistic will be "squeamish" about doing so and so must be incentivized to do so, whilst also requiring optimal inter-temporal distribution of punishment in order to achieve this. However, it would be desirable to further generalize the key results to allow for malevolent individuals ($\theta < 0$) and martyrs ($\theta > 1$), apply different criteria for credibility of punishment (e.g. renegotiation-proofness) as well as to allow for heterogeneity of coefficients of altruism between individuals. All of these issues would provide profitable avenues for further research.

Acknowledgements I would like to express my gratitude to the United Kingdom Economic and Social Research Council and to the Royal Economic Society for funding this research. I would also like to thank Kevin Roberts, Peter Hammond, Chris Wallace, Godfrey Keller, Alan Beggs, James Forder, Scot Peterson, Michael Griebe and a number of anonymous reviewers for their very helpful advice on various drafts. The usual disclaimers apply.

[40] For an extended analysis using evolutionary game theory of how such social costs to altruism can help to explain why societies with particular structures have evolved with particular complementary forms of imperfect altruism, see Povey (2014).

[41] Whether the sequential punishment game provides useful additional insights beyond the standard infinitely-repeated Bertrand and Cournot collusion models would depend on whether competition could be thought of as having additional dimensions beyond price, quantity or product design that could be meaningfully modelled as harm opportunities of the form analysed in this paper. Arguably the sequential punishment game would be most useful in modelling economies where firms are fused together with other political or social units so that competition is not purely market-based, such as in developing, feudal or command economies.

References

Abreu D (1986) Extremal equilibria of oligopolistic supergames. Journal of Economic Theory 39:191–225

Abreu D (1988) On the theory of infinitely repeated games with discounting. Econometrica 56(2):383–396

Asheim G, Nesje F (2016) Destructive intergenerational altruism. Journal of the Association of Environmental and Resource Economists 3(4):957 – 984

Aumann RJ, Shapley LS (1992) Long term competition - a game theoretic analysis. Ucla economics working papers, UCLA Department of Economics

Bénabou R, Tirole J (2003) Intrinsic and extrinsic motivation. The Review of Economic Studies 70(3):489–520

Bergstrom TC (2006) Benefit-cost in a benevolent society. American Economic Review 96(1):339–351

Bernheim DB, Stark O (1988) Altruism within the family reconsidered: do nice guys finish last? The American Economic Review 78(5):1034–1045

Cremer J (1986) Cooperation in ongoing organizations. The Quarterly Journal of Economics

Diamond P (1984) Money in search equilibrium. Econometrica 52(1):1–20

Diamond PA, Stiglitz JE (1974) Increases in risk and in risk aversion. Journal of Economic Theory 8(3):337–360

Fudenberg D, Maskin E (1986) The folk theorem in repeated games with discounting or with incomplete information. Econometrica 54(3):533–54

Hammond P (1975) Charity: altruism or cooperative egoism? In: Phelps ES (ed) Altruism, Morality and Economic Theory, Russell Sage Foundation, New York, pp 115–131

Harsanyi JC (1986) Utilitarian morality in a world of very half-hearted altruists. In: Heller RM Walter P Starr, Starrett DA (eds) Social Choice and Public Decision Making, Cambridge University Press, pp 57–73

Kandori M (1992) Repeated games played by overlapping generations of players. The Review of Economic Studies 59(1):81–92

Kotlikoff LJ, Persson T, Svensson LEO (1988) Social contracts as assets: a possible solution to the time-consistency problem. The American Economic Review 78(4):662–677

Lambson VE (1987) Optimal penal codes in price-setting supergames with capacity constraints. Review of Economic Studies 54(3):385–97

Lindbeck A, Weibull J (1988) Altruism and time consistency: The economics of fait accompli. Journal of Political Economy 96(6):1165–82

Malueg D (1992) Collusive behavior and partial ownership of rivals. International Journal of Industrial Organization 10(1):27–34

Meade JE (1973) Theory of Economic Externalities: The Control of Environmental Pollution and Similar Social Costs. Sijthoff, Leiden.

Messner M, Polborn MK (2003) Cooperation in stochastic OLG games. Journal of Economic Theory 108(1):152–168

Povey R (2014) Punishment and the potency of group selection. Journal of Evolutionary Economics 24(4)

Rawls J (1999) A Theory of Justice. Oxford University Press, Oxford.

Roberts KWS (1980) Interpersonal comparability and social choice theory. Review of Economic Studies 47(2):421–39

Rothschild M, Stiglitz JE (1970) Increasing risk: I. a definition. Journal of Economic Theory 2(3):225 – 243

Rubinstein A (1979) Equilibrium in supergames with the overtaking criterion. Journal of Economic Theory 21(1):1–9

Ruffin RJ (1972) Pollution in a crusoe economy. The Canadian Journal of Economics 5(1):110–118

Samuelson PA (1958) An exact consumption-loan model of interest with or without the social contrivance of money. Journal of Political Economy 66:467

Sobel J (2005) Interdependent preferences and reciprocity. Journal of Economic Literature 43(2):392–436

Stark O (1989) Altruism and the quality of life. The American Economic Review 79(2):86–90

Wen Q (2002) A folk theorem for repeated sequential games. The Review of Economic Studies 69(2):493–512