Socially Optimal Altruism in a Game of Sequential Punishment

Richard Povey

Hertford College and St Hilda's College, University of Oxford

March 28th 2018

(With thanks to the Royal Economic Society, UK ESRC, Kevin Roberts, Chris Wallace, Peter Hammond, Godfrey Keller, Alan Beggs and Elizabeth Baldwin.)

Motivating Issues

- Rational non-paternalistic altruism of the form $u_i = v_i + \theta \sum_{k \neq i}^{n} [v_k]$ is recognized to have potentially have both positive and negative social welfare consequences, though negative results are more rare [Buchanan, 1975] [Stark, 1989] [Bernheim & Stark, 1988] [Hahn & Ritz, 2014].
- **Extrinsic versus Intrinsic Incentives** - Altruistic preferences provide an *intrinsic* motivation for individuals to behave well. Punishment systems provide an *extrinsic* motivation. Interact in interesting and sometimes perverse ways.
- **Theory of Sequential/Repeated Games** - Provides an analytic framework in which to explore the interaction between altruism and punishment systems.
- **Positive and Normative Limits to Altruism** - Vast literature exists on positive limits. Most normative analyses are tied to particular specific contexts. From an evolutionary persepective normative limits may provide part of the explanation for positive limits.

A Word on Idiom

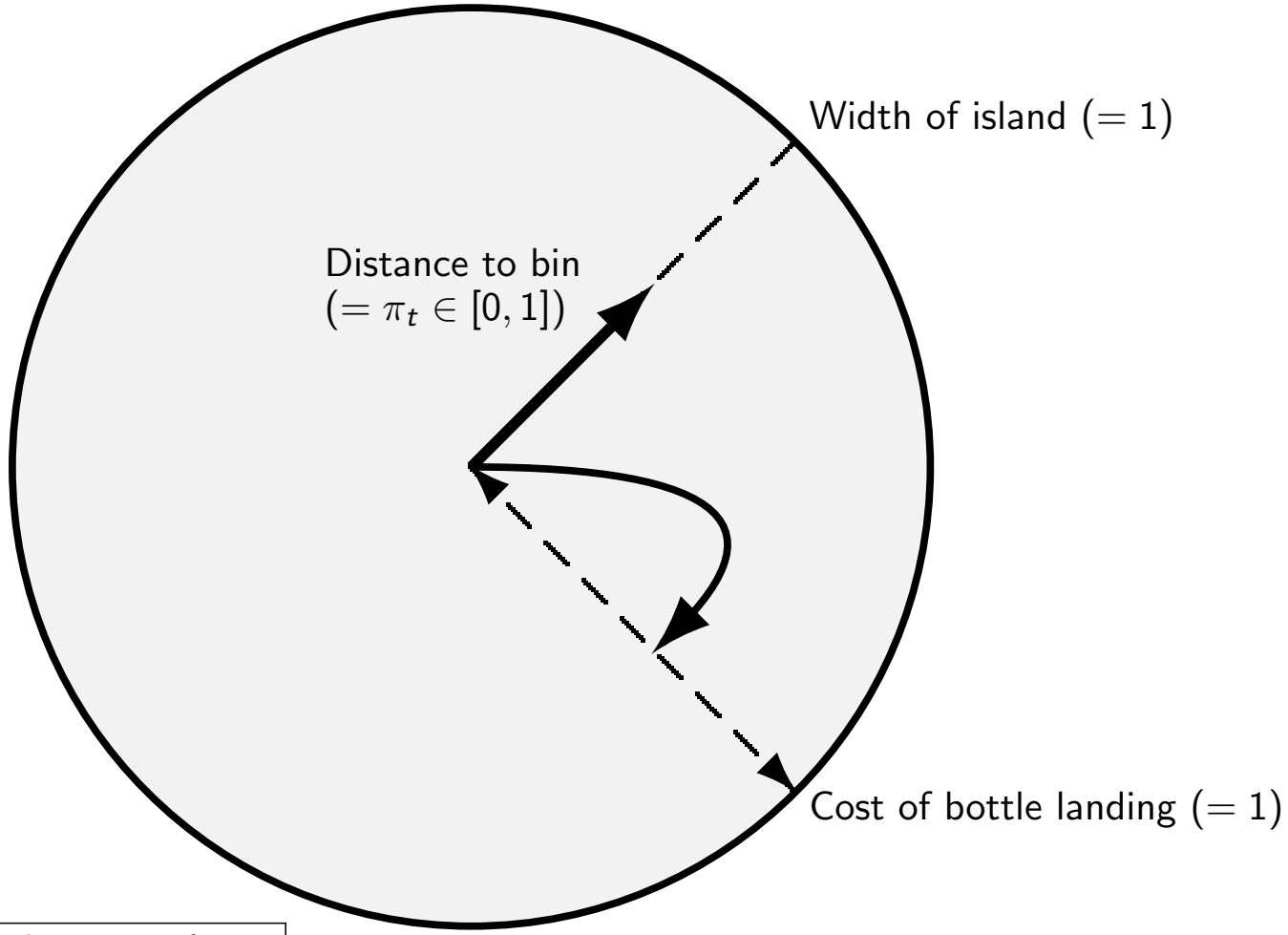
The sequential punishment model presented in this paper is intended as a highly abstract and stylized representation of social interaction, rather than as a realistic model of a specific situation. A simple "parable" can often help with the intuition. Models with a similar idiom include:

- Robinson Crusoe economy.
- Samuelson's "chocolate pension game" [Samuelson, 1958].
- Diamond's model of fiat money in a "coconut economy" [Diamond, 1984].

So, in that spirit, a desert island parable seems appropriate...

An Island Parable

Individuals (who have been on the island long enough to set up a "back garden") finish off a cold beer one at a time and must decide whether to walk to the bin or just throw their bottle into another individual's garden:



Coefficient of altruism - $\theta \leq 1$

Discount factor - $0 \leq \delta < 1$

Players' Preferences

- **Felicity** represents "private utility" from "economically fundamental" goods.
- In period t, player t moves so as to maximize his expected discounted (social) utility. The weighting on own felicity is 1 but the weighting on the felicity of others is θ.
- This is of course only one among a number of alternative ways to specify altruism. The advantage is that it enables us to simplify away from "multiplier effects" and focus on the normative analysis of rational non-paternalistic altruism.
- **Social welfare function is utilitarian in felicities**, or we can argue that Pareto efficiency (in either felicities or utilities) requires an equilibrium where no bottles are thrown.

The Model

- Infinite sequential game of perfect information. Players move once but live forever.
- In period t, player t receives a **harm/punishment opportunity** which if taken, yields a felicity benefit $\pi_t \in [0, 1]$ distributed according to continuous differentiable p.d.f., which is known in advance of the choice.
- If they take the opportunity, player t chooses another player to be the "target" who then suffers a felicity cost of 1.
- Model represents the idea that many aspects of social interaction could be conceived to take the form of sequential opportunities to impose exteralities.

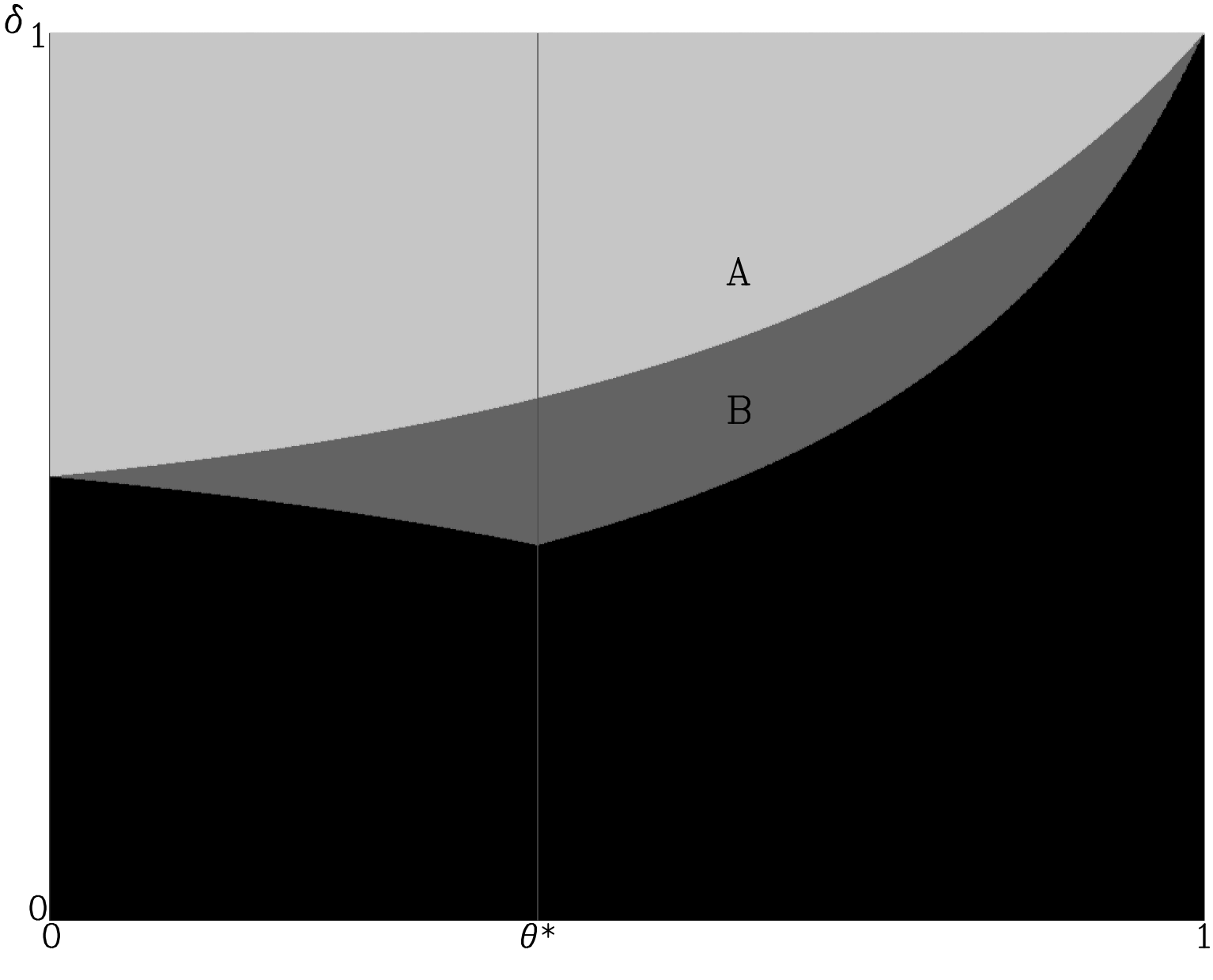
Three Effects

- **Temptation Effect** - Individuals with higher altruism are less *tempted* to inflict harm upon another individual for their own gain. (This is the main benefit from higher altruism.)
- **Willingness Effect** - Individuals with higher altruism are less *willing* to punish another individual for a previous misdemeanour by inflicting harm upon them. (This is a cost to higher altruism.)
- **Severity Effect** - Individuals with higher altruism also find some kinds of punishment less severe. In particular, if a fine was imposed, and some or all of the revenue is redistributed to another individual whose felicity has some weight in the utility function of the individual we are trying to punish, then any given size of fine is less severe for the punishee. (Another cost to higher altruism.)

Notation

- δ - **Discount factor**.
- θ - **Coefficient of altruism**.
- $\pi_t \in [0, 1]$ - **Benefit** from harming / punishing in period t (randomly distributed between 0 and 1).
- θ^* - **Socially optimal level of altruism** - Enables socially efficient equilibrium to be sustained for largest possible range of δ.
- δ^* - Lowest possible value of δ for which the socially efficient outcome can be sustained. (Corresponds to θ^*.)
- $\kappa(\theta)$ - Net loss in utility when individual deviates from socially efficient equilibrium when optimal punishment is applied. (So $\kappa(\theta) \geq 0$ is good.)

Overview - Socially Efficient Equilibria



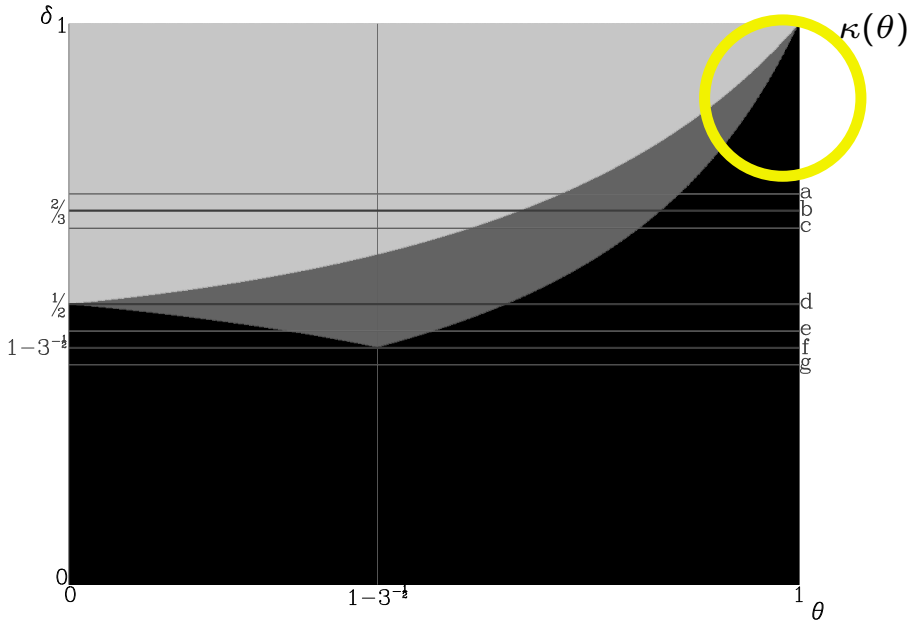
A - Supportable using Nash-reversion path

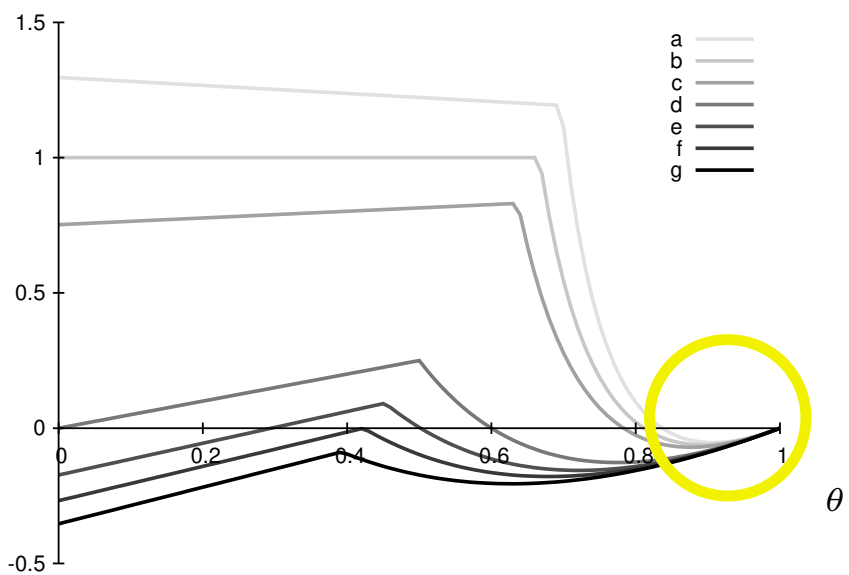
B - Supportable using optimal path

Solution Concepts

- **Folk Theorem** - [Aumann & Shapley, 1992] [Rubinstein, 1979] [Fudenberg & Maskin, 1986] For any given θ, as $\delta \longrightarrow 1$, the socially efficient outcome becomes supportable. We are interested here, however, in what happens as $\theta \longrightarrow 1$ for any given $\delta < 1$.
- **Optimal Penal Codes** - [Abreu, 1988] Abreu's framework of optimal penal codes in the form of punishment paths provides a natural framework that can be adapted to analyse socially efficient equilibria in the sequential punishment model.
- **Renegotiation Proofness** - [Farrell & Maskin, 1989] [Benoit & Krishna, 1993] We assume that society is able to avoid the temptation to let malefactors "off the hook". Thus we stick with subgame perfection rather than further refining the equilibrium criterion.

Results - Illustrated Using Uniform Distribution of Benefit





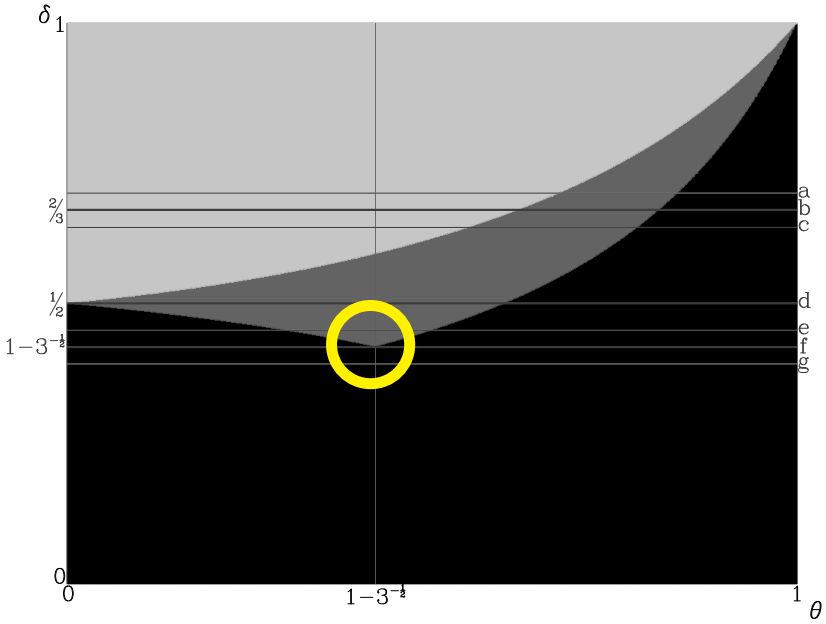
Theorem

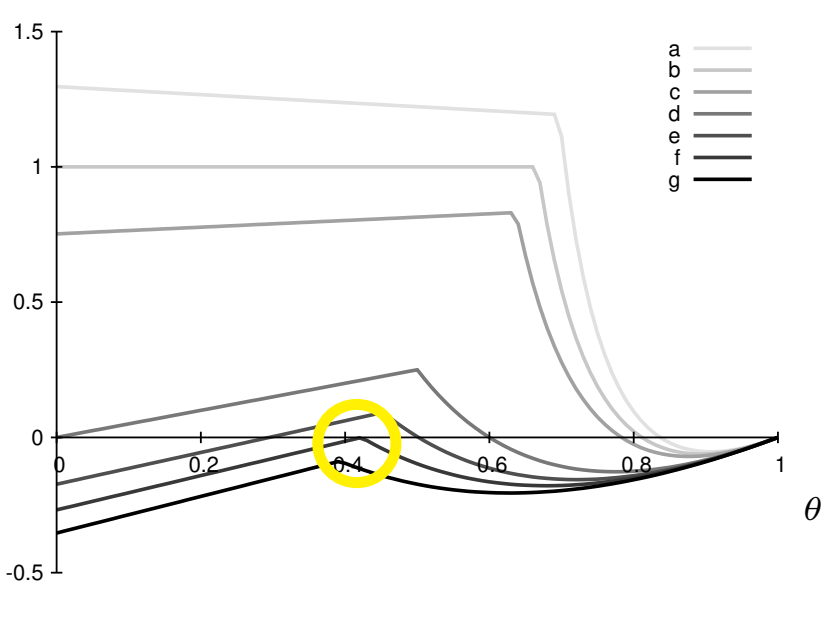
As $\theta \longrightarrow 1^-$, $\kappa(\theta) \longrightarrow 0^-$.

Proof.

Intuition: If $\theta = 1$ then $\kappa(\theta) = 0$. As $\theta \longrightarrow 1^-$, the willingness and severity effect become negligible, and the temptation effect ensures that $\frac{d\kappa}{d\theta} > 0$. □

Results - Illustrated Using Uniform Distribution of Benefit





Theorem

$\theta^* \in (0,1)$ $\left(\theta^* = 1 - \frac{1}{\sqrt{3}} \approx 42\%$ *for uniform benefit distribution* $\right)$

Proof.

Intuition: Let $\delta = \delta^* = \theta^*$. If $\theta = \theta^* + \epsilon$ then willingness and severity effect dominate temptation effect, so $\frac{d\kappa}{d\theta} < 0$. If $\theta < \theta^*$ then (because punishment is maximal) willingness effect is 0, temptation effect dominates severity effect, so $\frac{d\kappa}{d\theta} > 0$. □

The Socially Optimal Level of Altruism

Find a "knife-edge" where punishment is maximal and the socially efficient outcome is barely sustainable.

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

$$\theta = \delta \tag{2}$$

Solving (1) and (2) similtaneously yields:

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

Since $\tilde{\theta} < \theta^* < 1$, the socially optimal level of altruism must be greater than or equal to $\frac{3-\sqrt{5}}{2} \approx 38\%$.

Other Benefit Distributions

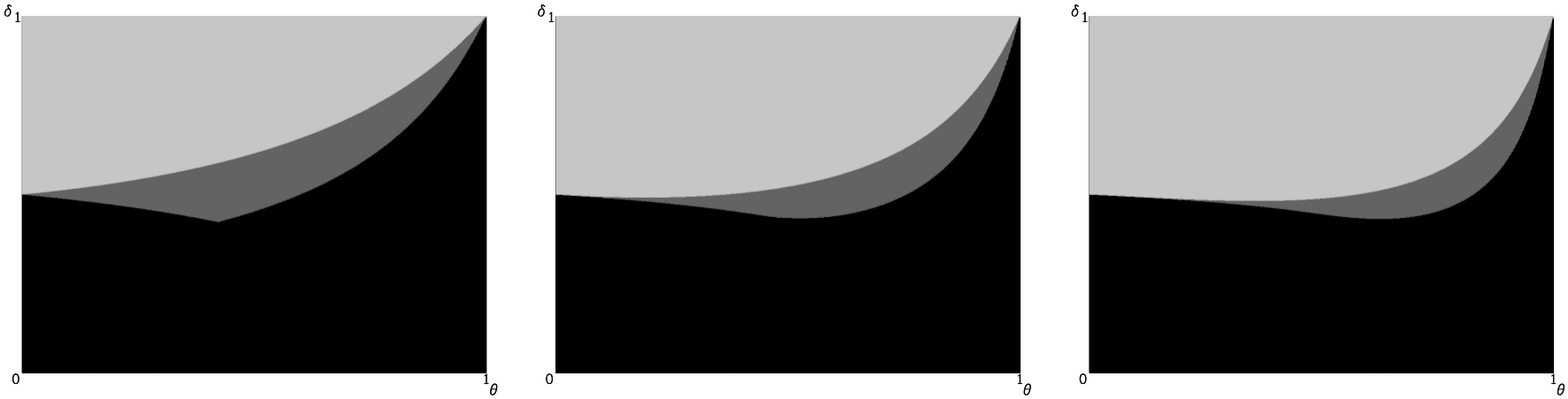


Figure: Socially efficient equilibria for $g(\pi) = 1$, $g(\pi) = 2\pi$ and $g(\pi) = 3\pi^2$.

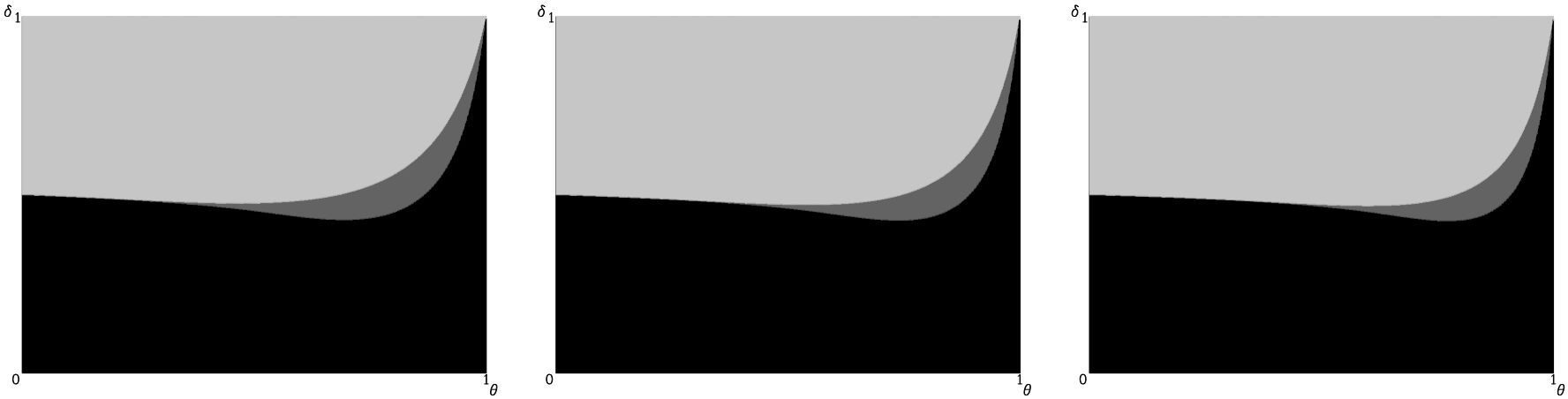


Figure: Socially efficient equilibria for $g(\pi) = 4\pi^3$, $g(\pi) = 5\pi^4$ and $g(\pi) = 6\pi^5$.

Conclusion - Implications / Applications

- Three effects framework might be useful in analysing altruism in other and more specific contexts.
- **Historical development of "extended order"** [Hayek, 1988] - as extrinsic incentive mechanisms become more sophisticated, less need for intrinsic moral controls.
- **International agreements** - Best chance of success (short of global federation) may be in world of imperfectly altruistic sovereign states.
- **Antisocial collusion** (rent seeking) may be easier to sustain in situations of partial/imperfect rather than (close to) perfect altruism.

References I

Abreu, Dilip (1988).

"On the Theory of Infinitely Repeated Games with Discounting".
Econometrica, 56(2), 383–396.

Aumann, Robert J. and Shapley, Lloyd S. (1992).

"Long Term Competition - A Game Theoretic Analysis".
UCLA Economics Working Papers 676, UCLA Department of Economics.

Benoit, Jean-Pierre and Krishna, Vijay (1993).

"Renegotiation in Finitely Repeated Games".
Econometrica, 61(2), 303–23.

Bernheim, Douglas B. and Stark, Oded (1988).

"Altruism within the Family Reconsidered: Do Nice Guys Finish Last?".
The American Economic Review, 78(5), 1034–1045.

Buchanan, James (1975).

The Samaritan's Dilemma.
In E. S. Phelps (Ed.), *Altruism, Morality and Economic Theory* (pp. 71–86). Russell Sage Foundation, New York.

Diamond, Peter (1984).

"Money in Search Equilibrium".
Econometrica, 52(1), 1–20.

Farrell, Joseph T. and Maskin, Eric S. (1989).

"Renegotiation in Repeated Games".
Games and Economic Behaviour, 1(1), 327–360.

References II

Fudenberg, Drew and Maskin, Eric (1986).
"The Folk Theorem in Repeated Games with Discounting or with Incomplete Information".
Econometrica, 54(3), 533–54.

Robert Hahn and Robert Ritz (2014).
"Optimal Altruism in Public Good Provision".
Cambridge Working Papers in Economics 1403, Faculty of Economics, University of Cambridge.

Hayek, Friedrich A. (1988).
The Fatal Conceit: The Errors of Socialism.
Routledge.

Rubinstein, Ariel (1979).
"Equilibrium in Supergames with the Overtaking Criterion".
Journal of Economic Theory, 21(1), 1–9.

Samuelson, Paul A. (1958).
"An Exact Consumption-Loan Model of Interest with or without the Social Contrivance of Money".
Journal of Political Economy, 66, 467.

Stark, Oded (1989).
"Altruism and the Quality of Life".
The American Economic Review, 79(2), 86–90.