Bounded Rationality and the Limits to Altruism

Richard Povey

Hertford College, University of Oxford

May 27th 2022

Presentation given at the "Models of Bounded Rationality" symposium at St Hilda's College, University of Oxford.

Overview

- **Limits to Altruism** - *Evidence for Imperfect Altruism*
  - Experimental / Behavioural Game Theory
  - Economics of Happiness
- **Explanations**
  - Information Processing
  - Empathy - With Other Humans and Future Self
- **Evolutionary Foundations**
  - Multilevel Selection Theory
  - Haystacks Model
  - Intrinsic and Extrinsic Incentives
- **Conclusion**

Strong and Weak Reciprocity

- **Weak reciprocity** is the form of altruism that can be seen as enlightened self-interest – individuals do each other a good turn because they rationally expect to be "paid back". However, this is insufficient to explain the complex functional integration of human societies [Fehr & Gächter, 2000].
- **Strongly reciprocal altruism** takes a **positive** and **negative** form, where individuals either help or harm others at material cost to themselves. This acts as "glue" holding institutions together, because the willingness of strong reciprocators to punish "cheats" forces selfish individuals to also behave well.
- **Altruistic punishment** - This is a key mechanism which acts as an "altruism amplification device", because it is usually less costly to punish another individual (e.g. by ostracising them) than it is to make a sacrifice for their benefit [Sober & Wilson, 1999].

Experimental / Behavioural Game Theory

- Some of the most important games that have been extensively tested in laboratory environments are:
  - **Finitely-repeated prisoners' dilemma**
  - **Public goods games**
  - **Dictator or ultimatum games**
  - **Centipede game**
- There is a vast literature which it would be foolish to attempt to summarise here, but a strong consensus that the predictions of classical game theory (based upon self interest and perfect common knowledge of rationality) are systematically violated. There is also a consensus that in order to explain observed behaviour it is necessary to introduce both:
  1. **Limitations upon perfect common knowledge of rationality**
  2. **Other-regarding preferences**

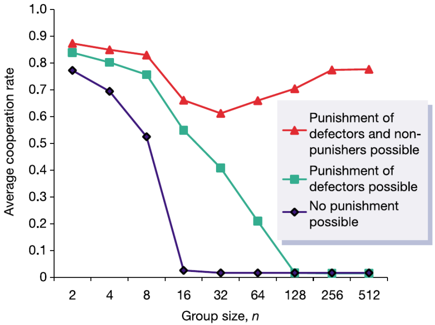
Prisoners' Dilemma

	C	**D**
C	$b-c$ $b-c$	$\underline{b}$ $-c$
D	$-c$ $\underline{b}$	$\underline{0}$ $\underline{0}$

- If agents play a co-operative strategy (C) then convey a benefit b upon the other player but themselves incur a cost c.
- Since D is a strictly dominant strategy, backwards induction can be used to show that any finitely-repeated prisoners' dilemma results in a unique subgame-perfect Nash equilibrium with D played by both players in every period.
- However, significant co-operation occurs in finitely-repeated experimental prisoners' dilemma games. Andreoni and Miller conclude on the basis of experimental evidence that rational reputation-building on the part of most agents plus true altruistic preference on the part of a minority offers the best explanation for this phenomenon [Andreoni & Miller, 1993].

Public Goods Games

- Public Goods games are similar to N-player prisoners' dilemma but where each player can choose how much to contribute, with each unit of contribution creating a benefit b which is shared over the group but at a cost $b > c > \frac{b}{N}$.
- Evidence [Dawes & Thaler, 1988] shows that for small groups average contributions are usually in the region of 40%-60% of the optimal level. When the game is repeated with the same individuals playing, the average level of contributions tends to drop over time. However, the ability to altruistically punish non-co-operators and non-punishers greatly increases the ability to sustain co-operation [Fehr & Gächter, 2000] [Fehr & Fischbacher, 2003].

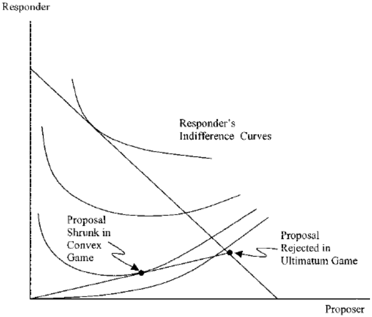


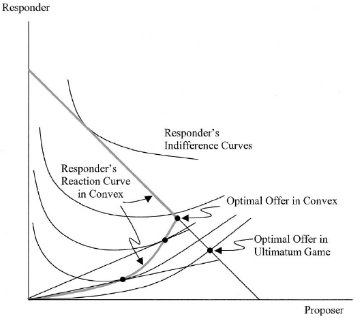
The Ultimatum Game

- The "ultimatum game" is played between two individuals. The first individual proposes the division of £x between the two individuals and the second individual can either accept the offer or refuse, in which case both get a payoff of 0.
- If both individuals are selfish and there is full common knowledge of rationality, classical game theory predicts that the first individual offers the smallest amount they can that is higher than 0 (i.e. 1p) and that the second individual accepts.
- However, when the game is played in experimental situations, the predicted outcome occurs extremely rarely, and there is significant variation between cultures regarding the amount that the first individual offers to the second. The empirical evidence has been summarised as showing that offers are usually between 30% and 40%, with the mode often being 50%. Very few offers are below 20%, and those which are this low are often rejected [Camerer & Thaler, 1995].

The Ultimatum Game

- Andreoni et al. have extended the ultimatum game to convexify the strategy space of the second individual by allowing them to continuously shrink the "pie" after the allocation is chosen by the proposer. Around 40% of subjects were found to have convex preferences for equity as illustrated by the diagrams below, whilst around 50% were found to have selfish preferences [Andreoni et al., 2003]:


Responder
Responder's Indifference Curves
Proposal Shrunk in Convex Game
Proposal Rejected in Ultimatum Game
Proposer


Responder
Responder's Indifference Curves
Responder's Reaction Curve in Convex
Optimal Offer in Convex
Optimal Offer in Ultimatum Game
Proposer

Centipede Game

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

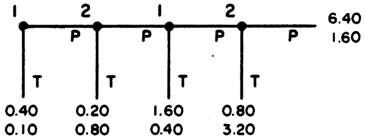


FIGURE 1.—The four move centipede game.

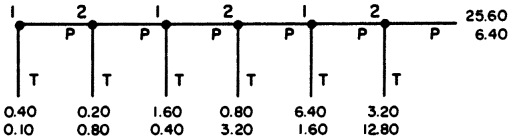
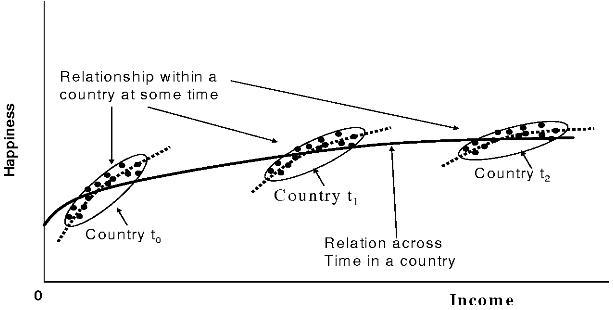


FIGURE 2.—The six move centipede game.

Economics of Happiness

A key result of this literature is that the positive relationship between happiness and income is greater within a society than it is over time as a society develops. This strongly suggests the presence of negative relative income effects ("keeping up with the Jones' ") which would seem to have a connection with negative strong reciprocity / preferences for fairness. [Easterlin, 1974] [Clark et al., 2008] [Layard, 2006].


Relationship within a country at some time
Happiness
Country t_0
Country t_1
Country t_2
Relation across Time in a country
0
Income

Information Processing

- The importance of heterogeneity of preferences and uncertainty in explaining empirical regularities in experimental game theory strongly suggests that imperfect information is critical to both limits to altruism and bounded rationality.

- Uncertainty about other agents' types plus information processing costs creates the need for "rules of thumb" (e.g. tit-for-tat "strategy" in repeated prisoners' dilemma).

- This also places direct limits on rational altruism under risk aversion since trying to benefit another agent is generally riskier due to uncertainty about their needs.

- Another important consideration is the issue of moral hazard. Not enough time to sat much about this here but there is a large area of work on the economics of altruism within the family [Becker, 1974] [Bergstrom, 1989] [Bruce & Waldman, 1990].

Empathy - With Other Humans and Future Self

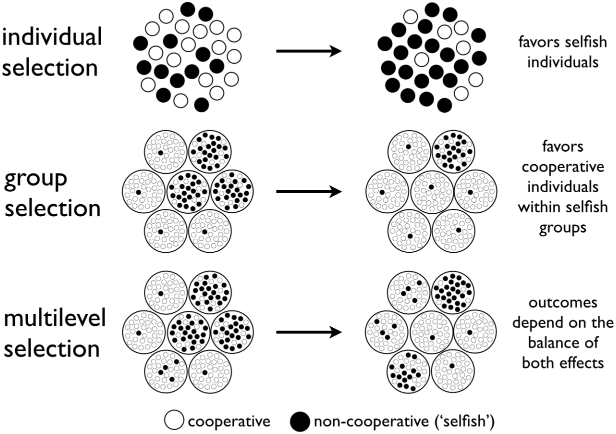
- **Empathy** is crucial to other-regarding preferences but empathising is mentally and emotionally resource-intensive.

- **Myopia** - There is a strong analogy between the limits to human ability to empathise with less-proximate others and with ones' future self, leading to myopia.

- The problem an individual faces when motivating their current self to act in the best interest of a future self is analogous to an organisation setting incentives for its members [Thaler & Shefrin, 1981].

- Making decisions in advance to take from one future self to give to another future self (e.g. saving for a pension) is easier than when sacrificing the immediate interest of the current self for a future self (e.g. temptation to spend on frivolities today).

Multilevel Selection Theory



individual selection → favors selfish individuals

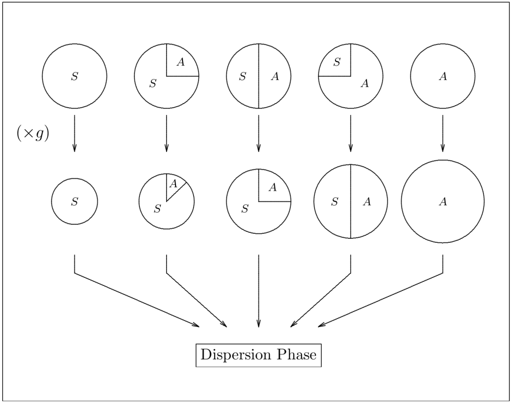
group selection → favors cooperative individuals within selfish groups

multilevel selection → outcomes depend on the balance of both effects

○ cooperative ● non-cooperative ('selfish')

Source: www.ecologyandsociety.org

Haystacks Model



Note: Although the Haystacks model provides a neat framework for analysing group selection, there are other mechanisms in play that have a similar effect and make group selection even more plausible for cultural evolution:

- **Ostracism**
- **Inter-group conflict**

Haystacks Model

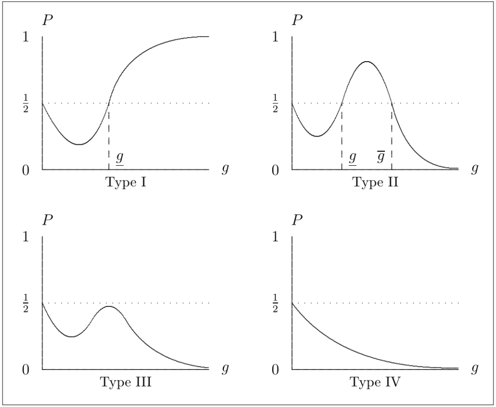


Figure: Source: [Cooper & Wallace, 2004].

A constant positive probability of mutations rules out Type I cycles, and we then need the number of periods of isolation to be in a "Goldilocks zone" to ensure Type II cycles and enable altruism to survive in the population.

Intrinsic and Extrinsic Incentives

- **Intrinsic Incentives** - Altruistic preferences provide an *intrinsic* motivation for individuals to exhibit altruistic behaviour.
- **Extrinsic Incentives** - Punishment systems provide an *extrinsic* motivation.
- Often it is empirically difficult to distinguish between the two (e.g. enlightened self-interest in the repeated prisoners' dilemma) [Hammond, 1975].
- These two forms of incentives represent alternative "social technologies" that can potentially be used to achieve socially beneficial outcomes, but which can interfere with one another in a perverse manner [Povey, 2014].
- The moral preferences and institutions which have evolved in human society represent a particular "policy mix" which may (or may not) be socially optimal.

Conclusion

- Ultimately bounded rationality and altruism (and its limits) are intimately connected evolutionary phenomena.
- A full picture involves classical and evolutionary game theory, the role of altruistic punishment in building coercive institutions (the state) and the ability of evolved resource allocation mechanisms (markets, the legal system) to aggregate more information than any individual human mind can comprehend, as emphasised by economists of the Austrian school such as Hayek [Hayek, 1960] [Hayek, 1988].
- In the sense that our individual rationality (and empathy) operates within limits, the normative task is to reform institutions to generate better outcomes. An evolutionary perspective does not imply that everything that has currently evolved is optimal. It does however predict that there will be hidden balances and unanticipated functionalities, leading to potential unintended consequences from reform. Hence a piecemeal approach is advisable.

References I

Andreoni, James and Castillo, Marco and Petrie, Ragan (2003).
“What Do Bargainers’ Preferences Look Like? Experiments with a Convex Ultimatum Game”.
The American Economic Review, 93(3), 672–685.

Andreoni, James and Miller, John H. (1993).
“Rational Cooperation in the Finitely Repeated Prisoner’s Dilemma: Experimental Evidence”.
The Economic Journal, 103(418), 570–585.

Becker, Gary S. (1974).
“A Theory of Social Interactions”.
Journal of Political Economy, 82, 1063–1093.

Bergstrom, Theodore C. (1989).
“A Fresh Look at the Rotten Kid Theorem–and Other Household Mysteries”.
The Journal of Political Economy, 97(5), 1138–1159.

Bruce, Neil and Waldman, Michael (1990).
“The Rotten-Kid Theorem Meets the Samaritan’s Dilemma”.
The Quarterly Journal of Economics, 105(1), 155–165.

Camerer, Colin and Thaler, Richard H. (1995).
“Anomalies: Ultimatums, Dictators and Manners”.
The Journal of Economic Perspectives, 9(2), 209–219.

Clark, Andrew E. and Frijters, Paul and Shields, Michael A. (2008).
“Relative Income, Happiness, and Utility: An Explanation for the Easterlin Paradox and Other Puzzles”.
Journal of Economic Literature, 46(1), 95–144.

References II

Cooper, Ben and Wallace, Chris (2004).
"Group Selection and the Evolution of Altruism".
Oxford Economic Papers, 56, 307–330.

Dawes, Robyn M. and Thaler, Richard H. (1988).
"Anomalies: Cooperation".
The Journal of Economic Perspectives, 2(3), 187–197.

Easterlin, Richard A. (1974).
"Does economic growth improve the human lot? : some empirical evidence".
Nations and households in economic growth : essays in honor of Moses Abramovitz, (pp. 89–125).

Fehr, Ernst and Fischbacher, Urs (2003).
"The Nature of Human Altruism".
Nature, 425, 785–791.

Fehr, Ernst and Gächter, Simon (2000).
"Cooperation and Punishment in Public Goods Experiments".
The American Economic Review, 90(4), 980–994.

Hammond, Peter (1975).
Charity: Altruism or Cooperative Egoism?
In E. S. Phelps (Ed.), *Altruism, Morality and Economic Theory* (pp. 115–131). Russell Sage Foundation, New York.

Hayek, Friedrich A. (1960).
The Constitution of Liberty.
Routledge and Kegan Paul Ltd, London.

References III

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

Layard, Richard (2006).
“Happiness and Public Policy: A Challenge to the Profession”.
The Economic Journal, 116(510), pp. C24–C33.

McKelvey, Richard D. and Palfrey, Thomas R. (1992).
“An Experimental Study of the Centipede Game”.
Econometrica, 60(4), 803–836.

Povey, Richard (2014).
“Punishment and the potency of group selection”.
Journal of Evolutionary Economics, 24(4), 799–816.

Sober, Elliot and Wilson, David S. (1999).
Unto Others.
Harvard University Press, Cambridge, Massachussets.

Thaler, Richard and Shefrin, Hersh (1981).
“An Economic Theory of Self-Control”.
Journal of Political Economy, 89, 392–406.