# Religion as a Seed Crystal for Altruistic Cooperation

Inaugural-Dissertation

zur Erlangung des Grades

Doctor oeconomiae publicae (Dr. oec. publ.)

an der Ludwig-Maximilians-Universität München

2006

vorgelegt von

Wolfgang Pfeuffer

| | |
|---|---|
| Referent: | Prof. Dr. Ekkehart Schlicht |
| Korreferent: | Prof. Dr. Gerhard Illing |
| Datum der mündlichen Prüfung: | 24. Juli 2006 |
| Promotionsabschlussberatung: | 26. Juli 2006 |

# Acknowledgements

This doctoral dissertation about the evolution of altruism is part of the results from my affiliation as a teaching assistant with the Institutional Economics Group at University of Munich. In its preparation I have been lucky to benefit from the real-life altruism extended by many people, who enabled me to begin and complete this piece of work and have contributed to making the past four and a half years a highly rewarding experience.

Professor Ekkehart Schlicht, who leads the Institutional Economics Group, served as my thesis supervisor and introduced me to the exciting field of institutional analysis. I would like to thank him for his constant encouragement and invaluable advice, and for making his group a highly inspiring place of enormous scientific diversity where real-world problems are addressed by combining the best of economics and its neighboring disciplines. The excellent atmosphere at our chair is also due to my fellow TAs Oliver Nikutowski and Florian Schwimmer, both of whom provided numerous insightful comments and suggestions concerning my work, as did seminar and conference participants in Munich, Barcelona and Tutzing. The same holds true for my former colleagues Thorsten Gliniars and Stefan Schubert and for one of "my" former students, Martin Leroch. Our chair secretary Maria Morgenroth kept my back clear of many administrative issues which would otherwise have impeded the progress of this work. Current and former student helpers, including Nicole Fröhlich and Roberto Cruccolini, provided support by getting me all sorts of digital and analog documents and, as far as the latter are concerned, dealing with library deadlines. Professors Gerhard Illing and Sven Rady kindly agreed to join the dissertation committee.

My thanks go out to all of the above, and to my Munich-based friends Maximilian Grasl, Matthias Hell and Tilmann Rave, for all your help and for making those years one of the best times of my life.

Writing about altruism and at the same time enjoying its blessings finally leads me to dedicate this piece of work to my mother, the most altruistic person I have ever met.

August 2006 Wolfgang Pfeuffer

# Contents

# List of Figures

# Chapter 1

# Introduction

"*The necessity for collective rationality is no guarantee that it will obtain.*"[1] In fact, social and especially economic life would be much easier if problems of collective action did not abound. Underprovision of public goods, overuse of common pool resources, and agency costs are well-known examples of economists' discomfort when it comes to the conflict of individual and collective rationality. Unlike in realms where Adam Smith's *Invisible Hand* can operate smoothly, the temptation for each individual to opportunistically pursue one's own advantage at the expense of others is insufficiently kept in check by market forces in these instances. In response, an important fraction of output in modern industrialized economies is expended year after year to enforce taxes, secure property rights and monitor behavior at the workplace. The associated cost represents a considerable deadweight loss: Resources could be saved and channelled toward genuinely productive uses if agents could effectively refrain from opportunism and credibly commit themselves to cooperative behavior. Everyone involved would then be better off. Notwithstanding the significance of this type of 'transaction costs' (as they might be termed) in advanced economies, people in dysfunctional societies where corruption and

[1] Macy (1998), § 1.1, his italics

nepotism reign marvel at the 'social capital' that alleviates problems of collective action and facilitates superior outcomes at least to a certain extent in successful economies. In other words, institutions (understood in an informal sense which includes e. g. good citizenship, the respect for property and the honoring of contractual agreements) are natural candidates when it comes to the determinants of economic performance and growth.

Empirical economists have of course taken up the task and confirmed the positive correlation between economic performance of society and institutions of social capital. Among the different dimensions of social capital, many studies highlight the role of personality traits which enable individuals to trust one another in business dealings.[2] The significance of trust in the economic sphere resides in agents' willingness to make themselves vulnerable to the opportunism of their partners: Trusting individuals in a world that dispenses with or cannot afford resource-consuming safeguards run the danger of being cheated and exploited by untrustworthy partners — at least near the end of business relationships. The precarious nature of trust as well as its efficiency implications have always been an important topic in economics: In the words of Kenneth Arrow (1974, p. 23), "trust is an important lubricant of a social system. It is extremely efficient; it saves a lot of trouble to have a fair degree of reliance on other people's word. Unfortunately this is not a commodity which can be bought very easily. If you have to buy it, you already have some doubts about what you've bought." Systematic analyses of trust and trustworthiness in economic contexts are thus called for, and game theory provides a range of formalizations for this endeavor. The Trust Game is offered as a model in which to appreciate the core meaning of trust in an economic context, i. e. the jeopardy and the potential involved in putting one's material fate in the hands of somebody else. Yet rationality on the part of both players in the basic Trust Game forces analysts into the prediction that trust will not be honored and hence cannot be shown in the first place. Rational players' incapacity to trust one another may be deplored by many standards, yet what is important

[2] See e.g. Tabellini (2005, Table 12) who finds trust to be the single most important facet of "social capital" in a cross-country growth regression.

to economists are its sobering implications for economic outcomes. These are driven home by the symmetric Prisoners' Dilemma, which portrays the conflict of individual and collective rationality in its most pronounced form. Individuals concerned with their own welfare turn out as both untrustworthy and mistrustful as they try to exploit their partners and at the same time protect themselves against being exploited. The result is a lock-in at an inferior state of mutual defection despite the fact that all parties involved would prefer an outcome of mutual cooperation where transactions are carried out smoothly.

Given the crucial importance of trust and other dimensions of social capital for the functioning of social systems and economies, the predictions derived in the framework of game theory are alarming. Fortunately, subjects in experimental game theorists' laboratories and in real-life transactions alike give proof of considerable endowments of social capital each day. The piece of work at hand is about the age-old question of how theory can accommodate this finding, the aim being twofold: In a first step, selected parts of the literature on how the conflict of individual and collective rationality is studied in economics and game theory will be reviewed along with the solutions proposed. Yet the overwhelming majority of these solutions, which rely on repeated-game effects, do not to apply when one wishes to adopt a scenario of — in the words of Douglass North (1984) — "impersonal exchange". The literature review hence sets the stage and is intended to motivate the contribution of the thesis: An argument will be proposed that religious involvement may be taken as an example in a 'signaling' approach to the evolution of cooperation, where the aim is to understand the persistence of collectively beneficial behavior in competitive settings which require — but do not favor — what Schlicht (2002) calls "anonymous trust".

The exposition proceeds as follows: In chapter 2 it will be urged that the one-shot (as opposed to the Iterated) Prisoners' Dilemma lends itself to formalize problems of collective action in societies where interactions are sporadic, ruling out future punishments for defective moves which might sustain cooperation as one of many Nash equilibria. Referring to biologists' understanding of the term, cooperation will be portrayed as al-

truistic in this scenario. Multilevel (or 'group') selection as a framework for studying the evolution of altruism will be discussed and adopted in chapter 3. Selected pieces of the literature on group selection enabled by assortative encounters (i. e. altruists tending to associate and interact preferentially with each other) will be reviewed in chapter 4. Adding to this literature, an ecological model of 'religious signaling' will be proposed in chapter 5. Religious involvement will be shown to possibly retain its signaling value and sustain the evolution of altruistic cooperation in long-run equilibrium. Chapter 6 concludes.

# Chapter 2

# Game Theory and the Problem of Collective Action

The theory of noncooperative games provides a natural framework for a theoretical analysis of collective action problems. Players in a game decide on how to behave in a situation of strategic interdependence, where the outcome is contingent on the strategies pursued by themselves as well as by their opponents. Players' behavior is modeled as rational: When choosing among the different strategies at hand, players opt for strategies which best serve their self-interest as given by a preference relation they have over the different outcomes. Preferences and the rational nature of all decision-making in a game are typically assumed to be common knowledge, meaning that each player knows the von Neumann-Morgenstern utility payoffs her opponents seek to maximize, and that opponents know to be known to do so, and that a player knows about her opponents' knowledge to be known to do so, *ad infinitum*. Common knowledge enables players to inform themselves about how their opponents might behave, which they often need to do because of the strategic nature of the interaction. Game theorists then try to determine the outcome to be obtained when all players' decisions are in equilibrium, meaning that no one can by switching to a different strategy bring about an outcome which ranks higher

according to their preference.

This chapter begins by reviewing the Trust Game, which is designed as a paradigmatic representation of the precarious nature of trust in economic settings. Its implications for collective action problems will be driven home by invoking the Prisoners' Dilemma, which abstracts from the sequentiality of moves and treats players symmetrically. The canonical way of allowing for cooperative outcomes in the Prisoners' Dilemma — repeated play in idiosyncratic relationships — will be discussed and rejected not only because of theoretical shortcomings but also as a solution for important collective action problems arising in modern large-scale societies with high degrees of geographic and social mobility: When transactions take place in increasingly anonymous settings, the 'shadow of the future' cannot be invoked to sustain cooperative equilibria. This shifts attention back to the one-shot case, where cooperating is a strictly dominated strategy never to be pursued by rational players. The extensive evidence that sporadic interactions do not prevent people from cooperating — neither in the laboratories of experimental game theorists nor in the field — remains enigmatic in this regard and calls for an alternative approach.

## 2.1 Trust Game

In the introduction it was argued that trust is an important component of the 'social glue' which fosters societies' economic performance and success by mitigating collective action problems associated with, e. g., asymmetric information or public goods.[3] On an abstract level, the economic issues inherent in trust are formalized in the Trust Game which was adapted from the 'Centipede' game (see e. g. Kreps 1990) and put to an experimental test by — *inter alia* — Berg *et al.* (1995).

The basic Trust Game is played by two individuals. Player 1 has received a resource

[3] See James (2002) for a more exhaustive list of examples and references.

endowment of $E > 0$, while player 2 has access to an investment project and is thus in a position to augment the first player's money. The two players move in two consecutive stages. In the first stage, player 1 may decide to pass her endowment on to player 2 or alternatively keep the endowment for herself. If player 1 keeps the money, the game ends with earnings[4] of $E$ and 0 for player 1 and player 2, respectively. Stage 2 of the game arises only if player 1 has handed over her endowment. Player 2 in this case has an augmented amount of $(1+r)E > E$ at her hands and full discretion as to its use: In particular, she is free to keep the augmented amount for herself, leaving player 1 with zero earnings. Alternatively she can hand the original amount of $E$ back to player 1 and split the increment of $rE$ in some way, e. g. keeping $rE/2$ for herself and paying the same amount to player 1. Players are taken to be concerned only with earnings of their own, preferring more income to less, when they evaluate the different outcomes of the game and make their choices.

Figure 2.1 shows the basic Trust Game described above in an extensive game tree representation. This game is easily solved by backward induction: Were the second stage to be reached, player 2 would choose to exploit player 1's trust and keep the endowment and the interest for herself, $(1+r)E > rE/2$. Common knowledge enables player 1 to anticipate player 2's reasoning when making her choice in the first stage of the game. Therefore she decides not to trust player 1 and contents herself with her endowment. Otherwise she would lose her endowment and end up with resources of $0 < E$. The game ends after the first stage with earnings of $E$ for player 1 and 0 for player 2, despite the fact that — by investing the endowment and splitting the returns, e. g. equally, among the two individuals — player 1 could end up with $(1+\frac{r}{2}E) > E$ and player 2 with $\frac{r}{2}E > 0$. Both players would prefer this outcome to the unique subgame perfect equilibrium of 'no trust'. The efficient outcome is however out of reach as players' individual rationality advises player 1 not to make herself vulnerable to an opportunistic move which player 2

[4]The term earnings, although not very common, is used here to emphasize that the matrix shows players' *material* (resource) incomes as opposed to utilities. This distinction will be important in the context of the indirect evolutionary approach in section 5.1.

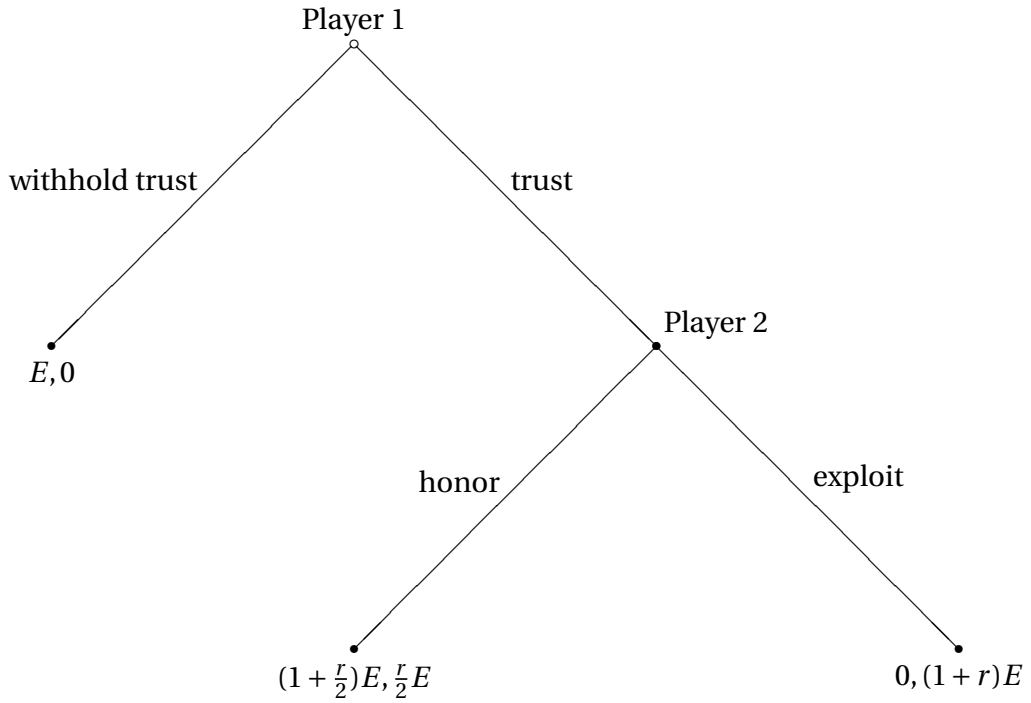


Figure 2.1: Basic Trust Game

would find expedient.

## 2.2 Prisoners' Dilemma

The Trust Game is of course not the only framework in which to study collective action problems and their consequences. In the words of Cook *et al.* (2005), its special appeal is that it captures the "*core* of an act of trust or trusting behavior - risk taking and thus making oneself vulnerable." (Their italics) The aspect of risk is however easily dealt with by the first-moving player who can effortlessly predict her opponent's untrustworthiness. Players in the Trust Game are hence in a situation of Prisoners' Dilemma, which succinctly drives home perhaps not the core meaning but in any case the welfare implications of the

| | cooperate | defect |
|---|---|---|
| cooperate | $R$ | $S$ |
| defect | $T$ | $P$ |

Figure 2.2: Prisoners' Dilemma

problem of trust.[5]

Players in the Prisoners' Dilemma move simultaneously. Figure 2.2 gives resource earnings for the row player, conditional on his choice to either cooperate or defect and the strategy pursued by her opponent's . Take earnings to satisfy $T > R > P > S$ and $2R > T + S$, where the latter inequality makes sure that the Prisoners' Dilemma is "pure" in the sense that mutual cooperation maximizes aggregate earnings and from this point of view entails efficiency. Agents' preferences in a Prisoners' Dilemma leave them with effectively no choice but to defect, as it portrays an interaction where free rides on others' efforts unambiguously result in a net resource gain. As a result they prefer outcomes in which they defect, regardless of whether their opponent cooperates or not: It is individually as rational to exploit a cooperator as to protect oneself against getting exploited by a defector. In terms of the previous discussion, players in a Prisoners' Dilemma turn out as untrustworthy and mistrustful at the same time. The dominant strategy equilibrium in which both players defect immediately follows as the unique solution to this game, despite the fact that both players would prefer the efficient outcome of mutual cooperation yielding its fruit over mutual defection.

Defection being a dominant strategy, players in a Prisoners' Dilemma can spare themselves any reasoning about their opponent's behavior. If this distinguishes the Prisoners'

[5] See Williamson (1993) for an outright rejection of trust as a concept with a meaning of its own in economics.

Dilemma from the trust game, where agents decide on whether to trust or to withhold trust after making up their mind as for their opponent's trustworthiness, the implications for collective action problems are the same: Much like in the trust game, where the inferior outcome is the unique subgame-perfect equilibrium, players in a Prisoners' Dilemma are faced with zero room for efficiency. This is essentially because they cannot credibly commit themselves to cooperative play in a world where binding contracts are not provided for. For such a commitment to be meaningful, it must be in the individuals' own interest to act on their promise. Both players of course prefer the outcome of mutual cooperation to a state of mutual defection. Mutual cooperation is however ruled out by the dominant-strategy nature of the PD, where the safe option to unilaterally defect and exploit one's opponent excludes the efficient outcome of mutual cooperation from the feasible set. Accordingly, efforts to rationalize cooperation in the one-shot PD in the following vein are hopeless[6]: Symmetrically rational players in a symmetric game should expect themselves to make identical choices and should therefor cooperate as they prefer mutual cooperation over mutual defection. The fact that both players would in fact prefer the unattainable outcome of mutual cooperation over what they end up with raises the problem rather than contributing to its solution.

## 2.3 Iteration

The standard approach in game theory to allow for cooperation in PD-like problems — taken in the 'folk theorems' and rigorously formalized by Aumann (1959) — is to argue that the immediate gains from exploiting a cooperative move come at a cost, namely punishment in subsequent encounters. In the Iterated Prisoners' Dilemma (IPD), the stage game is periodically played by the same individuals over and over again, constituting a 'supergame' in which players can condition their actions on their opponent's behavior in previous interactions. A pure strategy in a supergame specifies which action to take

[6] See Binmore (1994, ch. 3.3 and 3.4), who likens such attempts to efforts of "squaring the circle".

in each of the periods for all of the histories of play that may have materialized in the periods before, and players choose among the different strategies in order to maximize e. g. the present value of the flow of (usually discounted) von-Neumann-Morgenstern utilities they attach to the stage game outcomes.

### 2.3.1 Cooperative Equilibrium

The number of strategies conceivable in the IPD is of course immense. In order to see how iteration can foster cooperative outcomes it is convenient to highlight a specific trigger strategy which means to basically cooperate but equally to switch to continual defection against an opponent who has exploited one's cooperativeness. Depending on von-Neumann-Morgenstern utility differences and the possible discounting of future as opposed to present utilities, a player may find it in her interest to refrain from defecting even if utilities are strictly monotonic in the stage-game earnings: Defecting against a cooperator-retaliator would confer a one-time advantage as in the one-shot case, but this must be traded against a lost stream of future gains from successful cooperation over mutual defection when no definite end of the sequence of play is in sight. If future utilities are not discounted too much over immediate ones, the cost of defecting may well outweigh the benefit.

Two players who face each other in an IPD and expect each other to do this kind of reasoning will both opt for the trigger strategy considered and attain a subgame-perfect Nash equilibrium of self-serving mutual cooperation. This equilibrium is sustained by vigilance, i. e. the credible threat on the part of either side to protect themselves and retaliate against an act of exploitation by taking the measure which best serves their interest in such an event, namely to defect from then on. Two *caveats*, however, are in order: Cooperation can only obtain in the purely theoretical case of an infinite time horizon: If the sequence of play were known to end after some number of interactions with certainty, individuals would prefer to defect in the last stage as they need not fear

future punishment from then on. If defection cannot be averted in the last stage, there is however no convincing point in investing in the relationship by cooperating in the second-to-last interaction. In the case of strictly rational players, cooperation unravels across the board as such reasoning applies to all encounters by way of backward induction, even if the end of the play sequence is moved further and further into the future.[7] If an infinite time horizon is thus necessary for 'rational cooperation' to obtain in the IPD, it is however not sufficient: The best response to an opponent expected to go through with a strategy of continual defection is again to invariably defect just like her and avoid the painful (and costly) lesson of being exploited on the first encounter. A problem of coordination arises.

### 2.3.2 Coordination Problems

Introducing the trigger strategy makes the inefficient state of mutual defection less compelling compared to the one-shot PD case. It does not, however, solve the commitment problem: Cooperation can be reconciled with individual rationality when agents play the Prisoners' Dilemma repeatedly in the context of never-ending idiosyncratic relations, but only if both expects their opponent to embark on the trigger strategy as well. What expectations to form is however a highly non-trivial question: Continual mutual defection is just as consistent with Nash equilibrium as the successful cooperation. To achieve mutual cooperation requires agents to *coordinate* themselves on a specific equilibrium.

In fact, as pointed out by Brian Skyrms (2004, p. 5), the present value matrix of the utility flows for two agents choosing among the cooperative-retaliatory and the defective strategy can take the form of a paradigmatic coordination game, the Stag Hunt. The Stag Hunt formalizes a story originally told by Jean-Jacques Rousseau about the material rewards and risks of investing oneself in a cooperative venture: Mutual investments entail

[7] See Radner (1980) for the case of "$\varepsilon$-rational" players who do not insist on their most preferred outcome but content themselves with a result that comes (arbitrarily) close to their optimum. These players can achieve mutual cooperation with finitely repeated play.

the highest possible earnings for each of the two individuals, but unilateral investments are in vain, as costs are incurred without the corresponding benefits.

Figure 2.3 shows the present values of the utility[8] flows for two players in an IPD who discount the future with rate $\delta > 0$. The discount applied to future utilities may reflect players' rate of time preference or the fact that the sequence of play may end after each period with constant probability $\delta$. In symmetric interactions by two individuals who both employ the trigger (defective) strategy, mutual cooperation (defection) will ensue over the entire sequence of play and the present values are simply $\frac{R}{\delta}$ ($\frac{P}{\delta}$). An outcome where both players employ the trigger strategy and thus achieve mutual cooperation is unanimously preferred over the defective outcome just like in the one-shot case. The good news here is that mutual cooperation qualifies as an equilibrium outcome: Approaching a trigger strategy by acting the same can be preferable to trying to exploit the trigger strategy by defecting from the beginning on, which would result in a present value of $T - P + \frac{P}{\delta} = T + \frac{1-\delta}{\delta}P$. This is the case if and only if $0 < \delta < \frac{R-P}{T-P}$, i. e. if players do not discount the future to an extent which would make the initial exploitation gain $T-R$ weigh more than the punishment to be suffered from the second interaction on, $\frac{1-\delta}{\delta}(R-P)$. Not surprisingly but equally important, the trigger strategy played against a defector results in $S + \frac{1-\delta}{\delta}P$, which represents a loss of $P - S > 0$ compared to the case of matching a defector's behavior from the start and averting an act of exploitation in the first stage.

| | trigger | continual defection |
|---|---|---|
| trigger | $\frac{R}{\delta}$ | $S + \frac{1-\delta}{\delta}P$ |
| continual defection | $T + \frac{1-\delta}{\delta}P$ | $\frac{P}{\delta}$ |

Figure 2.3: Present Value Matrix of Payoff Flows in an Iterated Prisoners' Dilemma

[8] The symbols used are the same as in section 2.2 in order to facilitate comparison with the one-shot case, although they denote utilities instead of earnings. A linear utility function with zero intercept and unity slope lends itself in this context.

Provided that $\delta < \frac{R-P}{T-P}$, this IPD takes the structure of a Stag Hunt: It pays to "invest" in the relationship if the opponent acts the same while unilateral investments are in vain. Each player's utility-maximizing decision will thus depend on her expectations about her opponent's behavior, which will be confirmed in either of the two pure equilibria: Both will either embark on the trigger strategy and consequently cooperate in all stages or pursue a strategy of defection from the beginning on. Considering that these equilibria (along with a mixed one where players embark on the trigger and the defective strategy with strictly positive probability) are not the only possible outcomes in more general versions of the IPD, the coordination problem is even compounded: Deferring an analysis of the many other possible history-dependent strategies but generalizing the trigger strategy to implicit agreements by both players to randomize jointly over cooperating and defecting, the set of feasible (per-period) earnings vectors expands to the convex hull of the stage game earnings. As is known from the Folk Theorem, Nash equilibrium for sufficiently low discount rates is consistent with all feasible earnings vectors which are not inferior to the stage game equilibrium of mutual defection, i. e. the whole of the shaded area in figure 2.4.

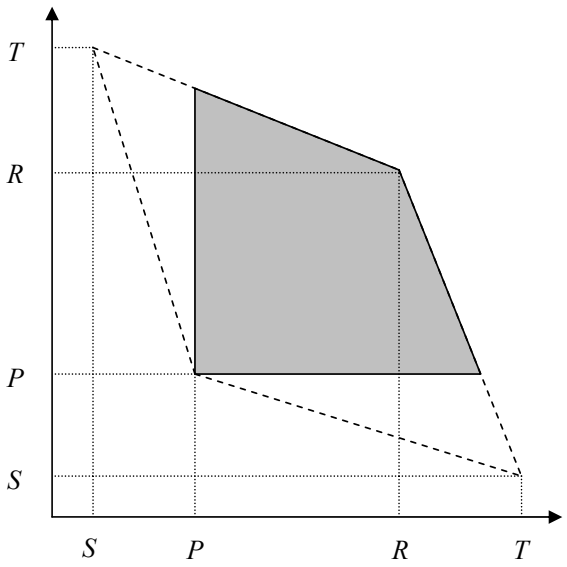


Figure 2.4: Set of Feasible and Individually Rational Per-Period Earnings in the IPD

In other words, virtually any 'amount' of cooperation over time (including zero) is possible

in the IPD, raising again the question of which outcome (if any) can be predicted to obtain.

Summing up, a problem of equilibrium selection is faced, the crucial point being players' expectations about their opponents' behavior. Classical game theory has little to say about how players form expectations when faced with multiple equilibria which survive the uncontroversial refinement of subgame perfection — as is the case not only in figure 2.3 but also in the general IPD[9] — and thus remains largely silent in this context.

### 2.3.3 The Axelrod Tournaments and Evolutionary Game Theory

The IPD tournaments organized by Robert Axelrod (1984) speak to the equilibrium selection problem in the IPD and to the wide variety of possible strategies, which goes well beyond the trigger strategies considered this far. Axelrod invited scientists to submit IPD strategies, formalized as finite automata coded in computer programs, and had them play one another a number of symmetric PD interactions in round-robin tournaments.[10] Strategies were at the end of a tournament ranked on the basis of total cumulative payoffs in all encounters. In the first tournament (repeated five times), each strategy played all others for 200 interactions. Although only continual defection was defensible on grounds of commonly known rationality, 14 different strategies were submitted. Top-scoring among these was *Tit-For-Tat* (TFT), submitted by Anatol Rapaport, which means to cooperate initially and, in the word of Robert Trivers (1971), "reciprocate" whatever an opponent did on the previous move from then on. Axelrod modified the design for the second tournament to rule out possible endgame effects and to align the experiment with the theoretically interesting case of infinite or uncertain length. In the five rounds of the second tournament, play duration was drawn from a random distribution which implied a median length of again 200 interactions. TFT emerged victorious from the second tour-

[9] See Friedman (1971).

[10] Each strategy was also confronted with a copy of itself and a randomizing strategy.

nament just as it had won the first. To check whether TFT's success depended in fact on the specific mix of strategies submitted to the tournaments, Axelrod subsequently took an "ecological" approach in a third experiment. In the ecological simulation, the composition of the strategy pool was updated after each round to reflect strategies in proportion to their relative success: Strategies with above-average payoffs would be represented to a higher degree then before, while below-average scoring strategies would get crowded out of the pool. As it turned out, TFT caught on in the ecological tournament as well — at least in the sense that it was represented in the long-run strategy pool to a higher proportion than any other strategy.[11]

Axelrod attributed TFT's success to it being "nice" (never the first to defect), retaliatory, forgiving and easy to understand at the same time and highlighted the importance of these characteristics for the resolution of collective action problems. Despite its intuitive appeal and impressive performance in the computer tournaments, TFT as a strategy in the infinite-horizon IPD appears in a different light when deductively analyzed in terms of the Evolutionary Stable Strategy (ESS) concept. The notion of an ESS was originally introduced by evolutionary biologists John Maynard Smith and George R. Price[12], who were the first to apply the theory of games to the study of animal behavior and conflict. Maynard Smith and Price observed that the resource value conferred upon an animal by a specific behavior, much like the earnings for a strategy in a game, depends on the behaviors pursued by the rest of the population. Evolutionary game theory was born when they proposed to formalize the competition for resources governing survival and reproduction as an uncooperative game. An evolutionary game is played in a large population with random pairings over and over again. Selection according to relative success as well as random influences — mutation — determine how the mix of strategies pursued in a population changes from generation to generation.[13] Maynard Smith and

[11] See Binmore (1994, ch. 3.2.5) for an adjustment of the widespread belief that TFT 'won' the ecological tournament.

[12] See Maynard Smith & Price (1973).

[13] Axelrod used the term "ecological" instead of "evolutionary" for his simulations in the third tournament. In fact, mutations were not accounted for in his analysis.

Price defined an ESS as a behavior which — when pursued by the vast majority in a population — will resist the competition from an alternative strategy (or a combination of strategies) injected into the system in small doses. In other words, a population settled in an ESS equilibrium fends off an attempted invasion by 'mutants'.

While the ESS concept allows to select among multiple Nash equilibria in many games, in the IPD it cannot lend support to an all-TFT equilibrium which Axelrod's ecological tournament might inspire: Follow Bendor & Swistak (1997) in considering e.g. the event of a slightly less provokable strategy of "Tit-For-Two-Tats" (TF2T) being injected into an all-TFT pool. TF2T, which tolerates a single defection and retaliates only when defected against twice, will easily find and defend its place in the population: In confrontations with TFT, provocability is uncalled for and hence no asset, and TF2T therefore does just as well as the incumbents. TFT is therefore evolutionary stable at best in a 'weak' sense: It may be argued that such an invasion is neutral because neither will TFT's frequency further decrease nor has observed play changed. Yet TFT and hence cooperation in general are prone to disaster in the updated strategy pool via a cascade of further invasions by strategies with decreasing 'niceness', at least when earnings monotonicity is the only restriction placed on the evolutionary process.[14] This allows Bendor and Swistak to conclude that TFT is in general not even "weakly" stable. Apart from the fact that TFT and hence reciprocity were from this point of view oversold in the wake of Axelrod's tournaments, thinking in ESS terms turns out as a source of frustration with regard to the problem of equilibrium selection in the IPD in general: As is formally shown by Farrell & Ware (1989) and Marinoff (1990), *no* pure strategy (in fact, not even a finite mixture of pure strategies) can be considered an ESS in the IPD[15]. In this sense, TFT's success was due

[14] Bendor and Swistak's example is the arrival of a strategy called "Suspicious Tit-For-Tat" (meaning to defect on the first encounter and return to TFT only if one was not defected against) which poses no problem for TF2T but will lock-in with TFT in an indefinite sequence of mutual defection. TF2T consequently emerges at the top of the earnings hierarchy. Under e. g. an "imitate-the-winner" dynamic where only the highest earning strategies develop, the population will then tend to an all-TF2T pool, which in turn invites a takeover by (first) a strategy which alternates between defection and cooperation and (then) continual defection.

[15] See however Lorberbaum *et al.* (2002) who show that three ESS in one-move-memory strategies exist in an IPD where players are sufficiently "error-prone" in strategy implementation, i. e. respond to their

to Rapaport's accurately "forecasting" the high number of cooperative strategies among the submissions by other participants (see Field 2001, ch. 3). Essentially not much is done about the problem of equilibrium selection by invoking the evolutionary approach, except that the initial distribution of strategies in the population takes the place of players' beliefs about their opponents' behavior.

### 2.3.4 Iterated Prisoners' Dilemma in Evolving Artificial 'Societies'

If the evolutionary approach to the IPD has not proved as helpful in increasing predictive power as one might have hoped, Axelrod's work laid the foundations for a strand of research which adds the dimension of interaction structure to the analysis. In the tournaments, each strategy was paired with all opponents in a round-robin fashion and the resulting relationships were 'closed' in that all of them persisted for a common (expected) number of interactions. The only way to punish a defective move was hence to treat like with like. Recent 'agent-based' studies, in contrast, have boundedly rational agents interact within in artificial 'societies', the structure of which is taken to evolve conjointly with agents' behavior in the IPD[16]. Strategies in these simulations are hence not limited to rules of behavior to apply in any given IPD relationship. 'Agents' rather decide for themselves with whom to play or at least have the option to end relationships as they please, which gives cooperative behavior an additional margin of defense against being cheated.

The exit option of ending the sequence of play is precisely what favors cooperation in Schüßler (1989). All agents in Schüßler's model — unconditional cooperators as well as a variety of 'trigger' strategies which defect after having displayed a specific number

---

opponent's previous action by making the 'wrong' move with a certain probability. Ironically, these do *not* include the "$\epsilon$" analogue of TFT but certainly that of continual defection.

[16] See Tesfatsion (forthcoming) for a survey of the expanding field of "agent-based computational economics".

of cooperative moves — transit an anonymous "search pool" to find partners for the IPD. While agents in the search pool have no information about potential opponents whatsoever, the IPD is played in an "iteration pool" within ex-ante stable relationships to be exogenously broken only with a small probability. In contrast to the 'canonical' IPD, however, play sequences in Schüßler's model end for an endogenous reason as well: Agents can dismiss their opponent if they no longer wish to continue the match. Such endogenous exit lends itself to unconditional cooperators as an act of retaliation against being cheated and is equally taken by defectors when cheated upon during their initial phase of cooperativeness. Following a relationship break-up, the parties involved (both the aggressor and the victim in the endogenous case) are sent back to the search pool. While it is true that anonymity in the search pool enables defective strategies to exploitative acts of the 'hit-and-run' type, ex-ante stability of matches in the interaction pool clearly favors cooperation: Unlike defectors, cooperators are given the chance to form relationships which remain intact and yield the fruit from mutual cooperation for the involved until exogenously cut. Uncooperative strategies therefore have to look for new partners and transit the search pool more often than cooperators and tend to be confined to the earnings from mutual defection as they "have to face new partners who might be as mean or meaner than they are themselves" (p. 737). The upshot of Schüßler's simulations is clear: Endogenous exit is sufficient for otherwise unconditional cooperators to collect higher earnings than defectors and catch on in the population when the exogenous break-up probability is small and evolution proceeds slowly, i. e. when the long-term benefits accruing to cooperators in stable relationships overcompensate the 'hit-and-run' exploitation gains from defection.

While agents in an anonymous search pool have no choice concerning with whom to enter into a play sequence, cooperators in Schüßler's IPD can selectively stay with their like. The resulting population structure gives cooperators an important edge over defectors. A natural way to explore the role of population structure in the IPD further is to have agents decide not only on how long to stay with a given partner but also on whom to play in the first place. This is achieved in a series of models with "choice and refusal" (CR) à la

Gale & Shapley (1962), where agents can identify and remember potential opponents and make use of agent-related information not only in existing relationships but equally in their initiation. The logic shared by these models is that play sequences are started, continued or resumed only when an offer is made by an agent to play a specific partner and accepted by the addressee. In other words, each IPD stage including the first requires both parties to establish or renew mutual consent to the transaction.

In Ashlock *et al.* (1996) agents base their choice and refusal decisions on expected payoffs which are the symmetric at the beginning of a simulation but are updated after each iteration stage: Whenever two agents agree to play one another, the earnings received from the interaction enter into an agents' books which keep a weighted average of past and present earnings achieved by playing that particular opponent. Offers are made and accepted only if the expected earnings against the potential opponent exceed a "minimum tolerance level". More precisely, each agent ranks potential opponents on the basis of expected earnings and extends one offer towards the most promising agent if the associated expected earnings are tolerable. Conversely, all offers received from tolerable agents are accepted. Agents who find all potential opponents intolerable and neither extend nor accept any offers earn a "wallflower" amount, while a "refusal" amount is earned by agents whose offer is rejected by an uninterested addressee. The dynamics of the society in this model is governed by average cumulated earnings — normalized by the total number of earnings achieved either by engaging in IPD stages and/or receiving wallflower or refusal earnings — and formalized as a genetic algorithm inspired by Holland (1975): As in Axelrod's tournaments, IPD behaviors are represented as finite automata and formalized as bit strings which code for an initial move as well as the moves to make in response to an opponent's previous behavior. In synchronized "genetic steps" at the end of a generation, couples are formed among a specified number of agents selected from the top of the earnings hierarchy, resulting in offspring with genetic material recombined and mutated from their "parents"' bit strings. Behaviors with above-average success thus tend to spread in the population as the genetically evolved offspring replace agents ranking at the bottom of the earnings scale. Ashlock et al. report their simulation results in terms of

the overall average earnings observed and find that cooperative behavior is favored in comparison with the 'unstructured' case of round-robin or random matching. Besides allowing for the emergence of interesting social networks, preferential partner selection due to a suitably chosen minimum tolerance level (given the other parameters) enables cooperation to spread in the population rather quickly in a many simulation runs. On the other hand — in particular for high refusal or wallflower earnings — populations may degenerate into "wallflower ecologies" of autistic agents where inactivity becomes a general phenomenon.

A closely related model by Hauk (2001) further increases the viability of cooperation and addresses the problem of wallflower ecologies by endogenizing the minimum tolerance level. In most of their simulations, Ashlock et al. treated the tolerance level as a commonly shared parameter. Some runs, however, had it evolve as part of agents' genetic makeup and surprisingly resulted in lower average earnings compared to parameterized cases where the tolerance level was chosen to equal the wallflower earnings. Hauk takes a different approach and has agents continually adjust their tolerance levels as part of an individual learning process which also encompasses periodical revision of IPD behaviors and thus takes the place of population learning in Ashlock et al. Also in contrast to Ashlock et al., only five IPD automata of reduced behavioral complexity are considered: Unconditional cooperation, TFT and TF2T (which are all 'nice') as well as 'Rip' and continual defection who are not. (Rip was originally conceived as a strategy to take advantage of TF2T's forbearance by defecting every other time in the 'paradigmatic' IPD.) Agents in Hauk's model keep track of a moving average of earnings achieved against all previously played opponents and use this average as their minimum tolerance level. The endogeneity of the tolerance level allows cooperators to free themselves out of "parasitic" relationships with a defecting agent, which are observed in Ashlock et al. for tolerance levels chosen too low given the remaining parameters. Instead their minimum tolerance level rises over time to reflect the potential experienced in encounters with other (nicer) opponents, and the parasite is eventually unmasked and rejected as a partner. Although not playing an iteration results in zero earnings and hence unambiguously in an opportunity cost

as all earnings in the IPD stages are nonnegative, the cost associated with refusing an intolerable offer is affordable for cooperative agents: After all, they quickly come to form a subpopulation of their own and interact exclusively with each other, achieving the benefits from mutual cooperation over mutual defection. Defecting strategies in turn "learn" to be content with interacting among themselves rather than becoming wallflowers and are eventually induced to switch to 'nicer' behavior at the next strategy revision step, which is when they are informed about the high earnings harvested by the cooperative agents. The result indicated by Hauk's simulations is stable cooperative play among the nice strategies with occasional and unsuccessful invasions by Rip mutants.

In the choice-and-refusal models agents endorse a specific (possibly complex in Ashlock et al.) IPD strategy and in each iteration decide with whom (if at all) to play based on previous experiences of their own with the available opponents. The resulting population structure clearly favors agents with cooperative IPD strategies including (but not limited to) TFT and thus supports the workings of reciprocity. Reciprocity as modeled by TFT is direct in nature and perhaps best summarized in the saying, "I won't scratch your back if you don't scratch mine." The model by Nowak & Sigmund (1998) differs in this regard as agents can adopt an attitude of "I won't scratch your back if you don't scratch theirs" and hence act as indirect reciprocators (see Alexander 1987). Interactions in this model are evocative of the Trust Game described in section 2.1 not only because they take an asymmetric form: In each encounter, a potential donor is given the option to extend a helping hand to her opponent, where the cost of helping is smaller than the benefit conferred upon the recipient. (If no help is given, both parties' earnings are zero.) Providing helping is thus collectively desirable. Yet given that recipients are like a vegetable in this game, without any capacity to reward helpful donors, rational players in the donor position have no incentive to help much like first-movers in the trust game choose not to trust as they anticipate rational second-movers' untrustworthiness. In Nowak and Sigmund's model, the donor-recipient game is played in a population with random pairing, and the roles of potential donor and recipient within an encounter are assigned randomly. Agents do not condition their behavior on individual play histories

with specific partners (in fact, the chance to meet one and the same agent more than once is negligibly small in the model). All agents are however assigned an individual and fraud-resistant "image score" which allows to address the phenomenon of indirect reciprocity. At the beginning of a generation the image score is zero for all agents. In the course of a generation (where any given agent plays the donor-recipient game once, several times or not at all; agents who are not selected to play receive zero earnings), an agent's image score summarizes how she performs as a donor within the general population in the following way: Whenever an agent is chosen as a potential donor and provides (refuses to) help, her image score increases (decreases) by one. Image scores of potential or actual recipients are unchanged. Potential donors can condition their action — if not on their own previous experience with a particular opponent — on their opponent's image in society by looking at her score. An agent's strategy hence amounts to a number which specifies the minimum score a potential recipient has to have for the agent to provide help when chosen as a potential donor. At the end of each generation, all agents have offspring according to their cumulated earnings. In the absence of mutations, offspring inherit the strategy of their parents, yet their image score starts off at zero. In their simulations of the basic case where each agent's image score is visible for all agents and mutations are excluded, Nowak and Sigmund find that indirect reciprocity is indeed established in the population as the following strategy eventually takes over: Help others if and only if they have an image score of zero or more. In the presence of mutations, endless cycles of cooperation and defection are observed: Unconditional cooperators who appear due to random influences can invade a cooperative population of indirect reciprocators, and the society gets vulnerable to defectors who never provide help. With a large proportion of defectors in the population, indirect reciprocators' vigilance against getting exploited comes back into play again and facilitates a resurgence of cooperative behavior.

## 2.4 Empirical Salience and Theoretical Alternatives

The models reviewed in the previous section make an important point by indicating that structured interactions favor the evolution of cooperation in the IPD. Predictive power is increased in comparison to the 'canonical' evolutionary IPD with round-robin or random matching in that cooperative equilibria gain independence from the initial composition of the strategy pool. In these models, they rather obtain — given the right circumstances — within "evolving systems of autonomous interacting agents"[17]. If taking interaction structure into account can help us understand the evolution of cooperation in repeated games, the assumptions involved in the IPD approach as such remain a source of dissatisfaction. Is the 'shadow of the future' an appropriate representation of modern large-scale societies? In today's market economies, business transactions take place in an increasingly anonymous setting. Business gets increasingly depersonalized as transportation costs and other barriers to trade become less and less important. With an increasing pace of social and technological progress, business partners change frequently and relationships get more and more ephemeral. For the repeated-game argument to go through, however, players must be in longer-term idiosyncratic relationships which allow them to keep a log of their opponents' observed behavior in (at least a few of) past encounters.

Authors of agent-based IPD models are of course aware of these informational requirements: Adding to the basic case where each agent's image score is visible for all others, Nowak & Sigmund (1998) conducted simulation studies with incomplete information about image scores. Agents in these runs can update their image of a potential donor in an encounter between two individuals only if they happened to observe the transaction, which is the case with probability smaller than one. Nowak and Sigmund find that incomplete information concerning image scores does not impede the evolution of indirect reciprocity but makes the model sensitive to the number of interactions taking

[17] Tesfatsion (forthcoming)

place within a generation and to population size: Indirect reciprocators depend on a large number of interactions taking place in a small population to counteract the reduced accuracy of image scores. In a related vein, Hruschka & Henrich (2006) study a model of preferential partner choice among unconditional defectors along with "cliquers", whose memory covers a small subgroup of preferred partners rather than the entire population. Cliquers in their model try to pair and cooperate — unless they make a mistake and defect — with members of their clique, whose attractiveness as a partner they summarize in an individual-related index similar to the expected earnings in Ashlock et al. When none of their preferred partners is available, cliquers give as yet unknown opponents a chance by cooperating with a probability reflecting their degree of niceness towards strangers. Depending on their performance as IPD partners, members of a clique can get deleted from the cliquer's list and replaced by a new partner. Simulations of the model suggest that personalized interactions with a limited set of partners are sufficient to sustain the evolution of cooperation in cliques, i. e. niceness and forbearance towards preferred partners and skepticism towards strangers.

These are only two examples of how the informational requirements in evolutionary models of the IPD have been relaxed. Nevertheless, a repeated-game argument will by definition always remain at the core of the IPD approach to the evolution of cooperation. Yet it is precisely the non-iterated case in which the importance of trust is greatest and at the same time the problem of trust most pronounced: Modern market economies — dispensing to an important extent with idiosyncratic relations — provide ample room for easily anticipated opportunism which one would expect to require costly ex-ante safeguards or block mutually beneficial transactions altogether. Yet casual observation of everyday transactions (which are so self-evident that they go largely unnoticed) speaks a clear language in favor of people's willingness and capacity to cooperate: Travelers often leave significant tips *after* meals in motel restaurants which they know to never visit again, participants in experiments give a fraction of the cake to responders in *Dictator* Games even when stakes are high, and *anonymous* donations to charity are surprisingly common. No repeated-game effects can be at work in these cases as future interactions

between the parties involved can safely be ruled out. Yet (many) people allow business and social life to operate smoothly by refraining from the safe and profitable option of refusing to tip, share or donate. These people are prepared to give part of their own wealth or, equivalently, neglect opportunities in sporadic interactions where a free ride at others' efforts and expense would go unpunished with certainty. This amounts to cooperating in a one-shot PD situation, where rationality in the standard game-theoretic sense would leave players with no other choice than to defect.

## 2.5 Summary and Discussion

Game theory portrays the chances for cooperation in Prisoners' Dilemma situations in a sobering light. This comes of course as no surprise, as the PD is constructed precisely with the impossibility of cooperation in mind. A dilemma is posed for society more than for the players, who have a dominant strategy at hand. Things are different yet still problematic in the iterated case. An infinite play sequence between any given two players and successful coordination is required for sustained mutual cooperation to obtain as one of many Nash equilibria. The explanatory power of game theory in this context can be enhanced by the evolutionary approach, especially when the level of abstraction of standard game-theoretic models is reduced like in agent-based models with e. g. choice and refusal of partners. If the evolutionary approach can also remedy the backward induction problem — as long as no strategies are admitted which know the end of a given relationship in advance and defect in the final stage, provoking strategies that defect on the second-to-last stage, and so on — its applicability still depends crucially on repeated-game effects. How can the evolutionary approach be applied to the one-shot case, which is so important in today's mass societies?

The model by Orbell & Dawes (1991) provides an illustration of how evolutionary thinking accommodates structured interactions in an analysis of sporadic interactions in anonymous settings. Players in this model cannot recognize former acquaintances, let alone

condition their actions on opponents' previous behavior. Whenever they engage in a (one-shot) PD interaction, their pure strategy is hence either unconditional cooperation or unconditional defection. Interaction structure arises in the model from players' decision whether to play at all: Orbell and Dawes consider a specific earnings matrix, where the earnings from playing a defector are strictly negative and the earnings from playing a cooperator strictly positive. For both cooperative and defective strategies it is hence worthwhile to voluntarily engage in an interaction if and only if they expect to be faced with a cooperative move. Orbell and Dawes model their agents as "cognitive misers": When forming expectations about any potential opponent, they tend to use a heuristic of projecting their own intentions onto others. Cooperators' subjectively perceived probability of meeting a cooperator is thus higher than defectors'. Cooperators as a consequence offer and accept to play a PD more often than defectors, which provides them with a competitive advantage over defectors as they tend to harvest the fruit from mutual cooperation, while defectors tend to content themselves with the (zero-valued) outside option of not playing.

Repeated-game effects are hence not essential to the evolution of cooperation, at least when other 'suitable' circumstances are met — in Orbell and Dawes, agents' heuristic when forming expectations and an outside option which yields higher earnings than playing a defective move. The crucial factor is interaction structure, which can obtain with the most sporadic of interactions in *a priori* anonymous settings — including the animal kingdom. The next chapter is therefore concerned with evolutionary biology.

# Chapter 3

# Evolutionary Biology, Group Selection, and Altruism

Following the strictly dominated strategy of cooperating in the one-shot PD means not only to content oneself with lower earnings than would result from defecting, but equally to advance the resources of one's opponent. Such behavior is referred to as a genuinely altruistic act by biologists, who have been studying altruism for a much longer time than evolutionary game theorists. Building on Charles Darwin's paradigm of the "survival of the fittest", evolutionary biology is concerned with the differential survival and reproduction of different gene sets which encode possibly different behaviors. Resulting differences in fitness for their bearers determine the dynamics of genotypes when natural selection operates on the population as the game is played over and over again, in the sense that individuals with below-average fitness give way to the more fit over time.

In evolutionary analyses it comes as no surprise that results depend crucially on the selection criterion applied. In the paradigmatic case, where selection pressures are taken to aim at individual fitness in unstructured populations, the hopelessness in explaining cooperation by rational players in the one-shot PD carries over directly to an evolutionary

account of altruism. When advantages in individual fitness foster survival and reproduction and altruists incur fitness penalties for the benefit of some selfish competition, they are bound to continually lose ground in the population and asymptotically become extinct. At the end of the day, purely individual selection just like individual rationality rules out altruistic behavior by its very definition: Individuals supposed to evolve when selected along the lines of individual fitness simply cannot afford altruistic resource transfers to competitors which leave them with below-average fitness. On the contrary, competition will eventually ensure that all 'survivors' in evolution turn out to behave in a way which best serves their turn in terms of individual resources. In fact, such reasoning is a methodological cornerstone in some parts of economics:[18] The differential expansion of firms on the basis of profits is precisely the grounds on which standard economics views — if not the process, then the results of — entrepreneurs' decision-making "as if" profit maximization was their goal. A related idea in game theory is that behaviors which result in high earnings get preferentially learned and adopted by a large number of players who try out different strategies and evaluate them according to their performance. John Nash in his doctoral thesis invoked the "mass action" interpretation to lend additional support to his solution concept: Equilibrium outcomes appear and can be conceived of as if each player went for maximum individual utility — typically monotone in individual earnings — given the actions of others. Yet if the workings of selection pressures offer themselves to support economists' maximizing hypothesis, an evolutionary framework offers alternatives to purely individual selection.

Concerning different targets of selection, Charles Darwin had higher-level selection mechanisms in mind right from the start. In *The Descent of Man*, he expands on the competition among "tribes" conquering each other according to relative success. Altruists are penalized in terms of individual fitness relative to nonaltruists within each tribe. One level above, however, altruists when largely among themselves, i. e. in tribes with a disproportionately high number of altruists, benefit from the fitness margin conferred

[18] See Friedman (1953, p. 21) but also Alchian (1950).

upon these tribes *vis-à-vis* low-altruism tribes.[19] After all, tribes with a relatively high number of altruists can be expected to operate at higher levels of productivity and grow at the expense of largely 'selfish' tribes.

## 3.1 Simpson's Paradox, and the Evolution of Cooperation and Within-Group Altruism

Darwin's appreciation of the group-adaptive qualities of altruism makes an important point with regard to the analysis of collective action problems in modern large-scale societies: Altruistic cooperation as exemplified in a game of (repeated) one-shot Prisoners' Dilemma is not necessarily ruled out by the criterion of selection among individuals who compete on the basis of individual wealth. At the heart of this claim is a phenomenon known in statistics as Simpson's Paradox (Simpson 1951). An oft-cited example of this phenomenon was encountered by the University of California in Berkeley in their admission process[20]: 'Although' women applicants were enjoying higher success rates than male applicants in almost every department, fewer women than men turned out to be accepted at the university as a whole. What explained this result was by no means the few departments where men were slightly more successful, but rather the fact that women tended to apply for programs which are harder to get into in general (irrespective of sex) than the programs envisioned by male applicants. This section provides an illustration of this idea in the context of cooperation in the Prisoners' Dilemma.

Consider a large number of individuals who engage in sporadic PD interactions and are genetically or socially predisposed to either cooperate or defect. By construction of the PD, a cooperator will receive lower earnings compared to a defector at her place in any

[19] See Darwin (1874, p. 134f).

[20] See Bickel *et al.* (1975).

given situation, regardless of the opponent's type. This is precisely why cooperating in a sporadic PD game amounts to irrationally pursuing a dominated strategy and remains enigmatic through the lens of classical game theory. Yet the PD has another property with potentially important implications in an evolutionary setting: The expected *levels* of earnings increase in either type's chances of playing a cooperator. This is true for defectors, as their hopes of taking a free ride on others' efforts increase, and is equally true for cooperators, as their risk of getting exploited is lowered:

$$E_C = p_{CC}R + \left(1 - p_{CC}\right)S = p_{CC}(R - S) + S \quad \text{and} \tag{3.1}$$

$$E_D = p_{CD}T + \left(1 - p_{CD}\right)P = p_{CD}(T - P) + P, \tag{3.2}$$

where $p_{CC} \in [0,1]$ and $p_{CD} \in [0,1]$ denote cooperators' and defectors' odds of playing a cooperator, respectively. In the standard evolutionary setting of a large unstructured population with random matching among $N_C$ cooperators and $N_D$ defectors, the odds of playing a cooperator correspond to the proportion of cooperators in the population and are the same for both types, $p_{CC} = p_{CD} = p_C := \frac{N_C}{N_C + N_D} \in [0,1]$. While inspection of (3.1) and (3.2) confirms that $\frac{dE_C}{dp_C}, \frac{dE_D}{dp_C} > 0$, the level effect cannot be invoked in favor of cooperators: $E_C < E_D \forall p_C$. An oversimplified yet instructive way to see how the PD nevertheless allows for precisely the type of group effect invoked by Darwin is by assuming a population partitioned into two groups. Let $N_{C1}$ and $N_{D1}$ denote the numbers of cooperators and defectors in the first group and $N_{C2}$ and $N_{D2}$ the corresponding numbers in the second group. Assume that the proportion of cooperators in group one, $p_{C1} := \frac{N_{C1}}{N_{C1} + N_{D1}} \in (0,1)$ is higher than in the second: $p_{C2} := \frac{N_{C2}}{N_{C2} + N_{D2}} < p_{C1}$ and restrict pairings to occur within each of two groups. Earnings for both cooperators and defectors will accordingly be higher in the first group compared to earnings for the same type in in the second: $E_{C1} = p_{C1}(R - S) + S > p_{C2}(R - S) + S = E_{C2}$ and $E_{D1} = p_{C1}(T - P) + P > p_{C2}(T - P) + P = E_{D2}$ as $p_{C1} > p_{C2}$. A structure like this allows for cooperators in the first group to have higher expected earnings than defectors in the second, reflecting the dominance of mutual cooperation over mutual defection which is just as essential to the

PD as the temptation to unilaterally defect:

$$\begin{aligned} & E_{C1} &&= p_{C1}(R-S)+S > p_{C2}(T-P)+P = E_{D2} \\ \Leftrightarrow \quad & p_{C1} &&> \left(\frac{T-P}{R-S}\right) p_{C2} + \frac{P-S}{R-S} \end{aligned}$$

can clearly occur for $p_{C1}, p_{C2} \in (0,1)$ as $\frac{P-S}{R-S} < 1$ by construction of the PD. While cooperators are still penalized in terms of resources and act altruistically[21] *vis-à-vis* defectors within each of the groups:

$$\begin{aligned} E_{C1} &= p_{C1}(R-S)+S < p_{C1}(T-P)+P = E_{D1} \quad \text{and} \\ E_{C2} &= p_{C2}(R-S)+S < p_{C2}(T-P)+P = E_{D2}, \end{aligned}$$

cooperators can in fact end up with a higher earnings than defectors when computing the receipts $\hat{E}_C$ and $\hat{E}_D$ of (fictitious) representative agent types across groups:

$$\hat{E}_C = \phi E_{C1} + \left(1-\phi\right) E_{C2} > \psi E_{D1} + \left(1-\psi\right) E_{D2} = \hat{E}_D$$

where $\phi := \frac{N_{C1}}{N_{C1}+N_{C2}} \in (0,1)$ and $\psi := \frac{N_{D1}}{N_{D1}+N_{D2}} \in (0,1)$ are the proportions of cooperators and defectors belonging to the first group instead of the second, respectively. To see this more clearly, assume a special case[22] where groups are of equal size, $N_{C1}+N_{D1} = N_{C2}+N_{D2}$, and where the population is equally divided into cooperators and defectors, $N_{C1}+N_{C2} = N_{D1}+N_{D2}$. Then $\phi = 1-\varphi = p_{C1}$ and $1-\phi = \varphi = p_{C2}$ along with $p_{C1}+p_{C2} = 1$. "Representative type" earnings read $\hat{E}_C = p_{C1}E_{C1} + p_{C2}E_{C2}$ and $\hat{E}_D = p_{C2}E_{D1} + p_{C1}E_{D2}$, so that $\hat{E}_C > \hat{E}_D$ becomes

$$\left(p_{C1}^2 + p_{C2}^2\right)(R-S)+S > 2p_{C1}p_{C2}(T-P)+P.$$

[21] If groups were continua of sufficiently small size, the within-group altruistic nature of cooperating might get compromised as the benefit conferred by an *individual* cooperator on all potential opponents (including herself) would be non-negligible and could offset the cost of cooperating. Cooperating then would raise one's own payoff over what would be earned by defecting, as if an $n$-person linear public goods game (see Cohen & Eshel 1976) was being played by the entire group.

[22] as in Fletcher & Zwick (2000)

Furthermore assuming that an additive PD is being played, this reduces to

$$\left(2p_{C1}-1\right)^2 > \frac{c}{b},$$

where $R-S=T-P=:b$ is the benefit conferred by cooperators upon their opponent (of either type) and $P-S(=T-R)=:c$ is the associated cost. Again, this can clearly occur for $p_{C1}<1$ as $\frac{c}{b}<1$. Assuming that high-earning strategies grow at the expense of underperformers by either differential survival and reproduction or imitation, the population ratio of cooperators will have risen in the next period, 'although' it has fallen in both of the groups. This dynamics, paradoxical as it may appear at first, is perfectly in line with the simultaneous effects of selection both within and between groups as alluded to by Darwin: Cooperators get exploited and crowded out by defectors within each group as expected, yet the biasedness of groups regarding the earnings level of both types favors cooperators in the first group over defectors in the second. The evolution of (within-group altruistic) cooperation depends on the relative importance of the two effects. In the example, the intergroup effect overcompensates the intragroup effect as long as the difference in the conditional probabilities of playing a cooperator (termed 'index of assortativity' by Bergstrom 2003),

$$p_{CC}-p_{CD}=\phi p_{C1}+\left(1-\phi\right)p_{C2}-\left[\psi p_{C1}+\left(1-\psi\right)p_{C2}\right]=\left(2p_{C1}-1\right)^2, \tag{3.3}$$

exceeds the cost-benefit ratio of cooperating.[23]

[23] The example, of course, does not speak to the *long-run* survival of altruism. Within-group competition would eventually have defectors crowd out cooperators in a system like the one portrayed here. See Fletcher & Zwick (2000) for computer simulations of the longer-run properties of populations which initially give rise to Simpson's Paradox.

## 3.2 Selfish Genes, Inclusive Fitness, and the Group Selection Debate

Darwin's account of the different levels at which selection works at the same time against and in favor of altruists did not prevent the idea of "group selection" from becoming the subject of an ongoing intellectual controversy among biologists and beyond. Resistance had formed in the 1960's against group selection models as in Wynne-Edwards (1962), where groups with high proportions of altruists were taken as adaptive units in their own right, turning a blind eye to competitive pressures within groups. George C. Williams's response to these 'naive' group selection models was to insist on the gene as the fundamental selection target[24], driving home the point that groups are too unstable and ephemeral to be considered as an object of evolution in themselves. The criticism met by group selection theory grew even stronger when biologists developed apparent alternatives where no reference to the adaptive qualities of groups was deemed necessary.[25]

---

[24] Williams (1966)

[25] One of the most widely cited examples in this context is Robert Trivers's analysis of reciprocity (Trivers 1971) which is however not particularly relevant for the present study: The explanatory value of the Trivers theory relates to how an equilibrium of mutual cooperation can be sustained when both agents play Tit-for-Tat in an IPD, which has been rejected as a modeling framework in the context of modern large-scale societies where idiosyncratic business relations cannot be taken for granted. From this perspective, Trivers does not necessarily address the phenomenon of genuine altruism (which is why he could equally well have used the term 'reciprocal selfishness' as opposed to 'reciprocal altruism', as had been suggested by a reviewer, see Field (2001, p. 126).)

In fact, Alexander Field points out that whether reciprocity in repeated games qualifies as selfish or altruistic depends on its current frequency in the population. This ambiguity about the nature of reciprocal behavior relates nicely to the phenomenon of multiple equilibria in an IPD where players choose between the trigger strategy and the strategy of continual defection contingent on expectations about their opponent's move, as already explained (see section 2.3.2). According to Field, the Trivers model — while it explains how reciprocity serves to *sustain* (nonaltruistic) cooperation — does not explain how reciprocity *originates* from a situation where (altruistically) operating in low frequency. In other words, the Trivers model just like classical game theory leaves the problem of why and how IPD participants coordinate on the cooperative outcome of the resulting Stag Hunt unaddressed. Brian Skyrms (2004), arguing that repeated-game effects *are* relevant to collective action problems, explores how population structure affects equilibrium selection in the Stag Hunt by facilitating the transition from the uncooperative to the cooperative equilibrium.

At the heart of William D. Hamilton's theory of *kin selection* (Hamilton 1963) is an idea which had surfaced before in Williams & Williams (1957), where attention was called to the fact that most early-in-life interactions necessarily occur among parent animals and their offspring. Hamilton's theory builds on the premise that altruistic acts by humans and other animals are preferentially conferred among close relatives, which are likely to share significant parts of their genetic composition. A subtle shift in perspective here makes self-sacrificing altruists appear to serve an interest of their (genetic) self: Individuals would willingly compromise themselves in part, expecting enhanced genetic prospects to enjoy. On a semantic level, Hamilton's original term *inclusive fitness* succinctly points out the essence of kin selection theory, defining individuals' fitness to extend to close relatives, backed up by genetic relatedness.

Richard Dawkins's theory of the *selfish gene* (Dawkins 1989) goes even further and focuses the analysis on alleles which combine to different genotypes in different individuals. Individuals here reduce to mere "vehicles of selection", admitted to the picture only as common destinies of different genes which happen to share the same boat in the fitness race. After the differential survival and reproduction of individuals, genes which had enjoyed above-average fitness across their different manifestations in different individuals will be overrepresented in the updated population. This observation leaves no room for altruistic individuals in a scenario where fitness is discerned purely at the level of genes.

Inclusive fitness and selfish gene theories were commonly perceived to play down the interest in groups (and in individuals, in the latter case) as adaptive units and discredit 'apparently' altruistic behavior as selfishness in disguise, which led many to believe that the group selection argument was not only implausible but also uncalled for.

## 3.3 Multi-Level Selection

Sober & Wilson (1998) argue persuasively that inclusive fitness and selfish gene theories

do not lend themselves to a rejection of the group selection idea. They propose to rather understand both paradigms as specific perspectives adopted within a more comprehensive approach to evolution, namely multi-level selection, which they view as a "unified evolutionary theory of social behavior". In fact, the group selection debate demonstrates that altruism as a behavioral phenomenon to acknowledge (or not) is to quite some extent at analysts' discretion. Arguing against group selection and in favor of inclusive fitness or selfish genes amounts to adopting a specific perspective and consolidating fitness accounts at the preferred level. On whichever level opted for, it is no surprise to conclude that anything which evolves simply cannot afford to incur fitness penalties. The alternative they propose is to make broader and consistent use of Dawkins's concept of vehicles of selection and apply it not only to individuals but equally on higher levels. According to this view, natural selection occurs and should be taken into account at *all* levels of association with fitness differences among the constituents. The hierarchy of possible selection targets ranges from chromosomes in a gene, genes in individuals, individuals in (kin or non-kin) groups, to groups in a population. Regarding kin selection theory, altruistic propensities are penalized within kin groups, yet families containing many altruists are favored over kin groups which do not. As for selfish gene theory, selfish genes are favored over altruistic genes within individuals, yet individuals with high proportions of altruistic genes outcompete individuals with high proportions of selfish genes which feed on and pose risks for their host organism.[26]

Charles Darwin in his example invokes two levels at which selection pressures simultaneously operate: the level of the individual and the level of the group (or "tribe"). Thinking along the lines of multi-level selection enables analysts to appreciate the same behavior as both altruistic at the lower level of the individual and adaptive at the level of the group. Of course, altruism will only evolve if the beneficial effect accruing to altruists

[26] As for (genuine) reciprocal altruism — in an IPD of fixed and known duration — Sober and Wilson's point is that a reciprocator is penalized within mixed 'groups' of a reciprocator facing an unconditional defector, yet two reciprocators when playing each other ('groups' of reciprocators) do better than defectors among themselves. Unlike in the infinitely repeated IPD, reciprocity here unambiguously qualifies as altruistic as it implies to refrain from safely exploiting an opponent on the last encounter from which on no future punishment is possible.

in high-performing groups as opposed to selfish types in low-performing groups offsets or overcompensates altruists' fitness penalties within both of the groups. In Sober and Wilson's terms, the virtue of the multi-level approach is to circumvent the "averaging fallacy", which would amount to consider only the net effect and (tautologically) conclude that any behavior which evolves cannot be genuinely altruistic.

The appropriate dose of averaging and accordingly the demarcation between altruism and selfishness is of course no trivial issue. Wilson and Sober's urging not to commit the averaging fallacy has recently been criticized in Tullberg (2003) as an invitation to the "average negligence". Among Tullberg's examples is firms' giving away of complimentary sample merchandise without any immediate returns, which he proposes to view as an "integral part of a chosen strategy" carried out (by no means altruistically) in order to introduce new products or increase sales numbers. In fact, it is precisely on these grounds that an outcome of mutual cooperation achieved by reciprocators in the infinitely repeated IPD qualifies as nonaltruistic: Two reciprocators in an IPD constantly refrain from exploiting their opponent, and one would be led to diagnose them with altruism when considering each interaction separately. Contemplating the entire sequence of play, however, the punishment costs of defection would outweigh the one-time gains from exploitation, so that defection at any time would lower rather than increase individual (cumulative) earnings. It seems reasonable to combine the returns of multiple transactions into an earnings aggregate for the strategy of playing TFT as opposed to continual defection, as they accrue to one and the same strategist. Yet it seems uncalled for to consolidate earnings accruing to distinct recipients — of the same type but in different groups as in Simpson's Paradox — into a hypothetical average. In fact, cooperating in the one-shot PD (as well as not defecting in the last stage of a fixed and known duration IPD) can be argued to be unambiguously altruistic *vis-a-vis* any opponent actually played, as it entails lower earnings for oneself and higher earnings for the opponent than would result from defecting.[27]

[27] See Field (2001, p. 123).

Multi-level selection thinking lends itself to an appreciation of cooperation in the one-shot PD. Besides rendering Simpson's Paradox less puzzling than it might appear at first glance, it advances the understanding of altruism by biologists and economists alike: The ground is cleared to acknowledge that cooperation in the example given in section 3.1 amounts to an act of (within-group) altruism although it was shown to have evolved. Likewise, the theory of multi-level selection points out favorable payoff margins for cooperators which escape the study of a given encounter but result from the interaction structure at the population level. It thus offers a possible explanation why behavior such as contributing anonymously to charity or providing unilateral help to others — although inconsistent with the maximization of individual wealth — may persist in a competitive environment, at least for some time. The next section is concerned with the longer run.

## 3.4 Modeling Group Selection: Islands and Haystacks

At the heart of Simpson's Paradox is a positive correlatedness of strategies at the population level: Unlike in the standard evolutionary setting, cooperative moves are more likely to be faced with cooperation than are defective moves. In the example of 'hard-wired' player types predisposed to either cooperate or defect in each encounter, such correlatedness arises from the assumed biased partitioning of the population with regard to types and the restriction on play to occur only within each of the groups. An evolutionary account of sustainable within-group altruism in this framework amounts to explaining how groups with sufficiently high variation regarding their cooperator-defector ratio are established and how this variation evolves over time. While parents and offspring naturally share important parts of their genetic composition as emphasized by the kin selection approach, multilevel selection models of altruistic cooperation in sporadic interactions (among unrelated members) must address the issue of how cooperative behavior tends to concentrate in isolated sub-populations. To complicate the picture, isolated cooperative milieus — while necessary for altruism to evolve in a group selection model

— are unstable in the long run: Due to the altruistic nature of cooperation, defectors grow at the expense of cooperators within all groups. If groups remain forever isolated, within-group dynamics will have defectors take over all groups as evolution proceeds, eventually eliminating the asymmetry of groups which is essential to the group effect.

John Maynard Smith (1964) proposed the widely influential haystack model as a framework in order to formally study the workings of group selection. In the haystack model, groups take the form of mice living in the different haystacks on a meadow which are set up in spring and torn down after the summer. After the haystacks have been cleared, the mice blend back into the meadow population for the winter, and the process repeats itself year after year. Part of Maynard Smith's motivation to build and study the haystack model were his doubts about the evolution of altruism by group as opposed to kin selection. In fact, he believed it to depend on unlikely events (like genetic drift pushing altruists to fixation in some of the groups as in the earlier 'islands' model by Wright 1945) while sibling groups come with considerable variance regarding their proportion of altruists without further ado. Maynard Smith assumes that haystacks are founded by females who have been fertilized as a result of random mating in the meadow population. In contrast to kin selection models — where siblings disperse and mate to start new sibgroups after each interaction — the group structure in a haystack model remains intact for several interaction rounds ("breeding generations"), and the sibgroups evolve into local subpopulations before dispersal. To analyze the model, Maynard Smith subjected altruists to drastically unfavorable within-group dynamics: Altruists not only reduce in frequency but are eliminated in the course of a breeding phase in all initially mixed groups. Such powerful within-group selection against altruists led Maynard Smith to conclude that group selection (viewed as an alternative to kin selection) cannot convincingly explain the evolution of altruism.[28]

[28]Ted Bergstrom (2002) argues that *if* cooperation is to survive in this scenario (which can only occur in initially homogeneous groups of two cooperators), it has to be of the "seemingly altruistic" as opposed to genuinely altruistic kind. Bergstrom looks at the cumulative numbers of offspring born to either type over the different breeding generations during a season and understands them as the earnings in a game played among the haystack founders. Such a game between founders at the beginning of a breeding phase can be shown to turn out as a Stag Hunt, where two cooperators play best responses to each other. As with all

Wilson (1987) considers a haystack model where altruists reduce in frequency instead of disappearing outright in all mixed groups. This specification allows him to address the role played by the number of generations spent in isolated groups: While the within-group dynamics can run its course and tends to eliminate altruists as this number increases, Wilson points out that altruists go by no means extinct during the first few generations when Maynard Smith's drastic assumption concerning mixed groups is relaxed. The resulting model is at the same time more realistic and more balanced than Maynard Smith's, in that it allows altruists to benefit from the fact that longer group duration reinforces not only the intragroup but also the intergroup effect: Wilson shows that for a number of interactions in isolated groups, the growing productivity differential between groups (much like accumulating sampling error as long as groups have not yet grown too large) causes sibgroups' initial variation regarding the altruistic allele to increase from generation to generation. When haystacks are cleared and recolonized at suitable intervals, the beneficial effect from spending multiple generations within groups can offset the detrimental one. Group selection as exemplified by the haystack model turns out as possibly more favorable for the evolution of altruism than "pure" kin selection.

Cooper & Wallace (2004) study a closely related model where groups of repeated one-shot PD players are randomly formed by drawing without replacement from a non-large population. An analytical result establishes that multigenerational groups are in fact *necessary* for within-group altruism to evolve in this case: Unlike in the haystack type models, no sibgroup effect operates in the first generation right after group formation. "[I]mmediate and random rematching" thus renders the group structure ineffective. The scenario reduces to the standard evolutionary setting where altruism cannot evolve. For lack of analytical solutions, Cooper and Wallace investigate the role of multigenerational

---

analyses of multi-level selection, the (non)altruistic qualities of cooperation depend on the level at which earnings are consolidated. In a similar vein but contemplating the interactions during a breeding phase separately as one-shot PDs, Skyrms (2004, p. 8) points out that the within-group dynamics of Maynard Smith's model induces "perfect correlation of types" after the first breeding. The evolution of cooperation then comes as no surprise, as "it is a defining characteristic of the prisoner's dilemma that cooperators do better against themselves than defectors do against defectors". Yet cooperating in the one-shot PD can safely be said to retain its altruistic flavor also in this case, as it means to forgo gains from exploiting a defenseless opponent.

groups by means of agent-based simulations. In line with Wilson's conclusions, they find that the global proportion of altruists after its initial decline can recover and rise above its initial level (after $\underline{g}$ interactions) as the time spent in intact groups increases. Eventually it must drop (falling short of its initial level at $\bar{g}$) and collapse when mutations are included to rule out persistently homogeneous groups full of altruists. The upshot is that within-group altruism among non-kin can survive even in this case rather than depending on the unlikely event of an initially homogeneous group full of altruists. Altruism within randomly formed groups can coexist in cycles with selfish behavior in the long run when the rhythm of group break-up and reform implies a number of interactions spent within groups in the range between $\underline{g}$ and $\bar{g}$ (which depend, of course, on the parameter values of the model).

## 3.5 Summary and Discussion

Group selection still remains a source of some controversy in evolutionary biology. This might come as a surprise considering that group selection is about nothing more (and nothing less) than a different (if possibly unorthodox) way of approaching uncontroversial phenomena. According to Reeve (2000), Sober and Wilson themselves are partly to blame as parts of their book suggest that group selection is materially distinct and in some cases superior to other frameworks of fitness maximization such as selfish gene and inclusive fitness theories.

In any case, there is no need to pit group selection against the 'orthodox' theories in order to demonstrate its usefulness in economics. As pointed out by Zywicki (2000), the language of multi-level selection lends itself perfectly to phrase the core idea of a theory of the firm à la Alchian & Demsetz (1972): Firms are collections of individuals, and more successful firms set themselves apart from less successful ones in that their employees achieve higher gains from teamwork or, for that matter, cooperation. The monitor comes into play because free riders see firms with high levels of teamwork as an invitation to

shirk and live on their co-workers' efforts, thereby undermining successful firms from within.

At the end of chapter 2 it was noted that structure plays a crucial role for the evolution of cooperation in the Iterated Prisoners' Dilemma. Multi-level selection theory can be seen as an open invitation to import this insight into an analysis of altruism in the one-shot case. It might however be argued that humans do not live in haystacks. The next chapter is about how to deal with this observation.

# Chapter 4

# Assortation, Signals, and Religion

The point of departure for this chapter is Cooper and Wallace's observation that when groups persist for several generations, matching of players becomes and for some time remains positively-assortative at the population level despite the fact that groups are initiated at random: Random sampling does not prevent some groups from containing more cooperators than others. As the high-altruism groups outgrow low-altruism ones, a cooperator's odds of meeting a cooperator will come to exceed a defector's for a certain period of time. Strategies thus get positively correlated and Simpson's Paradox obtains just like in the oversimplified two-period example (see section 3.1) where nonrandom matches were presupposed by assuming cooperators to mysteriously associate in the first group and of defectors in the second. Yet the redispersal of groups at suitable intervals is essential for such assortation to constantly re-develop in the long run. Otherwise it would get eliminated as defectors take over all groups from within.

If assortative play in the sense of nonrandom encounters can obtain despite random sampling as long as groups disband and reform at the 'right' pace, this does not however answer the question of why different groups exist in the first place, especially with regard to increasingly integrating large-scale economies. Signals are an evident mechanism in

this regard and can naturally explain why an otherwise anonymous society gets structured at all and how such structure leads to positively assortative interactions. Again, a simple example cannot harm in making issues clear. To investigate the role of signaling in the evolution of within-group altruism, consider a large population of $N_C$ players 'hard-wired' to cooperate and $N_D$ players conditioned to defect as before and introduce a commonly observable signal which players can send or not. Assume that the fraction $p_{SC} \in (0,1)$ of cooperators who send the signal is different from the fraction $p_{SD} \in (0,1)$ of defectors who do so. Interactions require the mutual consent of both partners involved and a large number of potential opponents are immediately available. Assume further that signal-sending cooperators just like signaling defectors for some reason (earnings-related or not) insist on playing only partners who send the signal like they do. In the simplest case, players have an in-built preference to interact only with the likes of them as far as the (non)display of the signal is concerned. Alternatively, agents could be understood to care about their material success: Cooperators could be seen as prudently escaping the role of an exploitation victim and defectors as consistently chasing their prey. When the indicative value of the signal, i. e. the connection between the display of the signal and individuals' types, is known to agents (consciously or implicitly), signaling cooperators like signaling defectors will refuse to play nonsignalers. In either case, play is limited to occur among signalers on the one hand and among nonsignalers on the other.

Interactions in this example are much as if the population were partitioned into two groups: the 'group' of signalers and the 'group' of nonsignalers. Comparing the proportions of cooperators within the two groups, it is straightforward to see that

$$\begin{aligned} & \frac{p_{SC}N_C}{p_{SC}N_C + p_{SD}N_D} < \frac{\left(1-p_{SC}\right)N_C}{\left(1-p_{SC}\right)N_C + \left(1-p_{SD}\right)N_D} \\ \Leftrightarrow \quad & p_{SC} < p_{SD} \end{aligned} \tag{4.1}$$

Quite intuitively, the signaling 'group' has a lower ratio of cooperators compared to the nonsignaling 'group' if the proportion of cooperators who display the signal is smaller than the proportion of defectors who do. The inequality in (4.1) can in this case be

rewritten as

$$(p_{SC} - p_{SD})\left(\frac{p_{SC}N_C}{p_{SC}N_C + p_{SD}N_D}\right) > (p_{SC} - p_{SD})\left(\frac{(1-p_{SC})N_C}{(1-p_{SC})N_C + (1-p_{SD})N_D}\right)$$

or equivalently

$$\begin{aligned} & p_{SC}\left(\frac{p_{SC}N_C}{p_{SC}N_C + p_{SD}N_D}\right) + (1-p_{SC})\left(\frac{(1-p_{SC})N_C}{(1-p_{SC})N_C + (1-p_{SD})N_D}\right) \\ > \quad & p_{SD}\left(\frac{p_{SC}N_C}{p_{SC}N_C + p_{SD}N_D}\right) + (1-p_{SD})\left(\frac{(1-p_{SC})N_C}{(1-p_{SC})N_C + (1-p_{SD})N_D}\right), \end{aligned}$$

which reduces to

$$p_{CC} > p_{CD},$$

where $p_{CC}$ ($p_{CD}$) are a cooperator's (defector's) population-level odds of playing a cooperator. When cooperators are less inclined to display the signal than defectors, cooperators are hence more likely to play a cooperator than are defectors. Exactly the same bias in cooperators' and defectors' chances to play a cooperator arises when $p_{SC} > p_{SD}$, i. e. in the more 'natural' case where cooperators are overrepresented in the 'group' where the signal is displayed. If interactions are sufficiently assortative (see equation (3.3) for an interpretation of the difference $p_{CC} - p_{CD}$ in terms of assortativity), within-group altruistic cooperation can evolve by Simpson's Paradox.

This example highlights the crucial issues to address in order to take signaling as a convincing explanation for the long-run survival of altruism: What is the nature of the signal, and why is it that cooperators and defectors differ persistently in their ability or inclination to send it? As a thought experiment, assume in the example that the (non)display of the signal is a genetic (or social) trait 'wired' into an individual just like her propensity to cooperate or to defect. Altruism then cannot survive in the long run even if a sufficiently high difference in the two types' signaling propensities — which narrows down as evolution proceeds — is presupposed for the first generation on whichever grounds.

Nonaltruistic signalers as well as nonsignalers eventually crowd out their altruistic 'in-group' opponents as the descendants or imitators of defectors and cooperators inherit the display of the signal with the same ease. The signal ceases to indicate agents' types and interactions become random as 'fake' signalers undermine the system. The population operates as if mixed groups in Simpson's Paradox were permanently isolated and must eventually converge to an all-selfish state. A natural avenue for the long-run survival of altruism hence is a differential ability or inclination to display the signal: The theory of costly signaling to be discussed in the following builds on the premise that it is more difficult or less advantageous for defectors to send a specific signal than for cooperators.[29] The next sections review and discusses different forms that such signals could take, religion include .

## 4.1 Moral Sentiments

Robert Frank (1988) in his book *Passions Within Reason* persuasively argues that an individual's emotional and motivational state becomes manifest in observable symptoms which are (at last in part) beyond conscious manipulation and can therefore serve as a reliable signal of one's commitment to cooperate even when individual self-interest would counsel defection, as in the one-shot PD. Frank's analysis can be seen as elaborating on the idea of "constrained maximization" among "translucent" players in Gauthier (1986) (see Binmore 1994, ch. 3.2.2). Gauthier considers players who turn a blind eye to the earnings from exploitation and can therefore commit to a strategy of returning the like of their opponent's move. When such constrained maximizers can reliably communicate their commitment to opponents, mutual cooperation becomes possible as a Nash equilibrium outcome (along with mutual defection). Frank takes the idea of translucency to an evolutionary setting and proposes two kinds of signals to convey information about an individual's hard-wired strategy: "sincere manners" and reputation.

[29] The theory of costly signaling was introduced in biology as the 'Handicap Principle' by Zahavi (1975).

The array of symptoms forming a persistently fraud-resistant signal of commitment — sent more easily by cooperators than by defectors — in the "sincere manners pathway to moral sentiments" includes facial expressions, body language, pitch and timbre of voice, and other dimensions of the physiognomy. In the extreme case where cooperative intent manifests itself in symptoms which no defector can display like cooperators do, emotions translate into an appearance and demeanor which indicate a player's type with perfect accuracy. In a similar vein, Frank deals with the "reputation pathway" and argues that a favorable reputation can signal a commitment to cooperate, although its indicative value in the context of altruistic cooperation is by no means evident: To altruistically cooperate means to refrain from cheating in a situation where no detection (and hence future punishment) is possible. As a consequence, nonaltruists by defecting in (only) these situations will by definition never compromise their reputation and could thus perfectly 'fake' the signal. Frank argues that nonaltruists reveal their true spirit nevertheless. Not all interactions in life are 'safe' for defectors (in the sense that cheating simply cannot be detected). Nonaltruists in order to preserve their favorable reputation need to restrain themselves and refrain from defecting in precisely these instances. In this regard, they face an "implementation problem": Psychological findings known as the 'matching law' suggest that when rewards from defecting are not safe yet immediate (as opposed to the rewards from preserving one's reputation which may not arrive until much later in time), subjects find it hard to be patient enough. In the extreme case, it is assumed that nonaltruists can impossibly build a favorable reputation while altruists — who never defect — naturally have it.

The implications of the two signal variants for the evolution of altruistic cooperation are the same. If signalers' preference to stay with their like or, equivalently, cooperators' prudence not to get exploited tells them to beware of defectors (which they know to reveal themselves by not sending the signal in the latter case), strategies are perfectly correlated from the start. Interactions are structured much like from the second generation onward in the haystack model (assuming that at least one homogeneous haystack full of cooperators got colonized), and cooperators drive defectors to distinction. Genuinely altruistic

or not, the survival of cooperation in this case is not very convincing as it depends on mutations to be ruled out: Clearly, a signaler whose type changes for some unexplained reason and her progeny would take over the cooperative milieu. A more general model with partially reliable signals addresses this problem.

In Frank (1987) a signal is displayed by all individuals and takes the form of a continuous variable influenced by both genetic and random components. Players draw their individual signal levels from two different densities such that cooperators by genetic influence have a higher mean signal level than defectors. The two signal densities overlap in a way which makes the signal partially reliable: Values under a lower threshold are only drawn by defectors, while values above an upper one are only drawn by cooperators. Signal values between the two threshold levels are however drawn by cooperators and defectors alike. Signal 'fakes', albeit imperfect, are hence admitted to the picture. Conditional on the signal value of potential opponents, agents in this model can either choose to play their 'hard-wired' strategy in a PD or to take an outside option which here yields the same earnings as two defectors in a PD would receive. While defectors' expected earnings are higher in the PD than from the outside option irrespective of their opponent's type, cooperators are faced with a trade-off: Playing a cooperator in the PD yields higher earnings than the outside option, but victims of exploitation by a defector receive less than the outside option. Frank shows that cooperators who account for the exploitation risk in the PD by requiring a minimum signal from opponents they play with can end up with the same average earnings[30] as defectors in a stable polymorphic equilibrium, depending on the parameters on the model.

[30]Which, as has been argued, does not rule out within-group altruism in the 'mixed' encounters. Frank (1994) acknowledges the usefulness of the multi-level perspective; originally he viewed selection as operating only at the individual level. See Field (2001, ch. 4) for a discussion.

## 4.2 Moral Sentiments, Cognitive Misers, and Embeddedness

Both "detection" (of possibly misleading signals of commitment) and "projection" (i. e. ascribing one's own intention to others) are explored by Macy & Skvoretz (1998) in their agent-based model of cooperation among strangers. Yet rather than taking these mechanisms for granted as in Frank (1988, 1987) and in Orbell & Dawes (1991)[31], respectively, Macy and Skvoretz ask whether they can come to evolve in competitive settings.

Interactions in this model are "embedded" in the sense that agents belong to one of several "neighborhoods". Each agent pairs with one of her neighbors in a certain fraction of encounters and with a "stranger" in the rest. Agents decide in each encounter to either engage in a PD game or resort to an outside option (valued equal to or greater than the earnings from mutual defection). Agents are represented as bit strings (like in Ashlock *et al.* 1996) which encode a wide variety of behavior: In each pairing, agents can (but need not) base their decision to play rather than use the outside option on an assumption about how potential opponents will behave in case the interaction gets consummated. This assumption (if applicable) may involve projection of one's own intentions onto others, as is assumed in Orbell and Dawes, or the opposite, i. e. expecting that potential opponents' intentions of defecting or cooperating run contrary to one's own. are about to make the opposite move behavior in will either behave like themselves if an interaction is consummated. Agents can also base their decision to participate in the game on a "marker" displayed by potential opponents. Yet unlike in Frank's model, where telltale clues by assumption allow inferences about their senders' commitment to cooperate, this marker is available to all agents independent of their cooperative or defective disposition. Another criterion available to agents is membership in their own neighborhood, the decision rule being to play neighbors and run away from strangers or vice versa. Agents

[31] the example given in section 2.5 of how the evolutionary approach encompasses the one-shot case

using more than one of these criteria will engage in a PD with a potential opponent with a probability corresponding to the proportion of favorable pieces among the information received, while agents using none of them will either play in all encounters or not at all. When interactions get consummated, agents are disposed either to unconditionally defect or cooperate or alternatively to defect against strangers only and while cooperating with neighbors.

Strategies in Macy and Skvoretz's model evolve by selection and mutation according to a genetic algorithm which operates after each round of pairings but takes into account a weighted average of both current and previous earnings. More successful strategies spread only by "social contact" in the sense that an encounter (or even a consummated interaction, in some specifications) is required for a relatively wealthy agent to pass on parts of her 'genes' to someone else. Macy and Skvoretz find that agents (starting from a random distribution of strategies) generally prefer an attractive outside option — valued higher than the earnings from mutual defection — over engaging in the PD in the first phase of a typical simulation run. This allows cooperative intentions (towards neighbors and strangers alike) to spread by random drift, in some neighborhoods more so than in others. Agents in a neighborhood where cooperative intentions have become disproportionately frequent can then start successful cooperation and with their wealth serve as role models within their community. In their contacts with strangers, members of the 'cooperative' neighborhood will export cooperative attitudes to other groups and thus create a population-wide climate for the emergence of a willingness to play the one-shot PD with strangers. The transition from parochial (meaning to avoid strangers and transact only with neighbors) to universal cooperation here occurs via either projection and/or detection, which allow 'open-minded' cooperators to find and consummate interactions among themselves. Macy and Skvoretz's simulations indicate that a population with universal cooperation is quite robust against subversion: Mutants who evolve to defect against their neighbors will of course catch on within their community. The unraveling of cooperation caused by a mutant in one neighborhood cannot however spread by contacts among strangers to 'healthy' neighborhoods where no mutation has occurred and wealth

is relatively high, making them insusceptible to a strategy which at home is already eating into the foundations of its own success. On the other hand, agents who evolve to selectively cheat on strangers (e. g. by faking the marker associated with cooperative play) cause transaction between strangers to be ceased population-wide: After "infecting" their own neighborhood due to the exploitation gains earned, all agents who interact with members of this community will get adapted to the danger of being exploited in anonymous encounters by reclaiming the outside option. Selective pressures against agents prepared to cooperate with strangers is hence weakened, which allows these agents to again concentrate in one of the groups by random drift and evolve projecting and detection methods anew.

Macy and Skvoretz conclude that there is no need to outright assume the fraud-resistance of Frank's telltale clues and the projection strategy of Orbell and Dawes's cognitive misers. Both avenues to the evolution of altruism can emerge and be sustained endogenously as a punctuated equilibrium in which the society experiences cooperation break-downs and recoveries alike. What instead drives their model (in addition to a comfortable outside option and their specification of the selection criterion) are socially distinct subpopulations, preferably those which are small relative to the population. As an example they invoke Max Weber's Protestant Sects.

## 4.3 Max Weber's 'Protestant Sects'

Max Weber (1970), when traveling the United States of the early 20th century, observed that affiliation to one of the various churches of the country traditionally played an important role when it came to initiating business relationships:

> "Hardly a generation ago when businessmen were establishing themselves and making new social contacts, they encountered the question: 'To what church do you belong?"' (p. 303)

Referring to Baptists, e. g., Weber reports that "[a]dmission to the congregation is recognized as an absolute guarantee of the moral qualities of a gentleman, especially of those qualities required in business matters" (p. 305). According to Weber, Baptist affiliation serves as a signal because the admission to the Protestant sects in America — unlike churches in e. g. France and Germany — requires applicants to pass "the most careful probation and [...] closest inquiries into conduct going back to early childhood" (p. 305). Moreover, Weber hints at the fact that membership was quite costly also in financial terms:

> "It should be realized, in addition, that church affiliation in the U.S.A. brings with it incomparably higher financial burdens, especially for the poor, than anywhere in Germany." (p. 302)

The importance of belonging to a specific church (or sect) got less important over time so that when Weber visited the country,

> "[...] the kind of denomination [to which one belongs] is rather irrelevant. It does not matter whether one be Freemason, Christian Scientist, Adventist, Quaker, or what not. What is decisive is that one be admitted to membership by 'ballot,' after an examination and an ethical probation in the sense of the virtues which are at a premium for the inner-worldly asceticism of Protestantism and hence, for the ancient puritan tradition. Then, the same effect could be observed." (p. 307)

Weber emphasizes that although these associations also served as a lender of resort in times of financial distress, the signaling value of the emblem used by members in order to identify themselves concerned primarily one's intentions:

> "And hence the badge in the buttonhole meant, 'I am a gentleman patented after investigation and probation and guaranteed by my membership'. Again, this meant, in business life above all, tested *credit worthiness*." (p. 308; his italics)

Boudon (1987) points out how Weber's appreciation of the signaling value of membership in the Protestant sects relates to the special requirements of the U.S. society, which resembled today's large-scale and depersonalized economies already in Weber's days (at least when compared with France and Germany). Boudon summarizes Weber's impressions as follows: "In the United States social and geographic mobility are greater; ethnic heterogeneity is greater; the stratification system is less rigid; and the stratification symbols are less visible and less marked than in France or Germany." The Protestant sects in the U.S. then can be understood as a functional equivalent to e. g. the *légion d'honneur* in France, which served as an important symbol system associated with high entry costs. The lower demand for symbols and the existence of alternatives in Europe also explain why sects with by their careful screening of applicants and imposition of entry costs were qualitatively different from churches, as emphasized by Weber:

> "It is crucial that sect membership meant a certificate of moral qualification and especially of business morals for the individual. This stands in contrast to membership in a 'church' into which one is 'born' and which lets grace shine over the righteous and the unrighteous alike. Indeed, a church is a corporation which organizes grace and administers religious gifts of grace, like an endowed foundation. Affiliation with the church is, in principle, obligatory and hence proves nothing with regard to the member's qualities. A sect, however, is a voluntary association of only those who, according to the principle, are religiously and morally qualified. If one finds voluntary reception of his membership, by virtue of religious probation, he joins the sect voluntarily." (p. 305f)

Weber's observations date back to over one hundred years ago. How do they relate to today's large-scale western economies? Social structure is, of course, based in part on common values, and religion still has its place in the matrix of values in the society. Yet church membership per se is nowadays anything else than obligatory and involves significant costs as well as time (if practiced). Hence it seems problematic to simply apply Macy and Skvoretz's premise — that all members of society firmly and invariably belong to one sect or, for that matter, group — to the contemporary world. Secularism in today's occidental societies makes it even more interesting to investigate the role of religious

involvement in a group selection approach to the evolution of altruistic cooperation. Or, to borrow the term from Schlicht (1995), how can religion (if only accidentally) act as a "seed crystal" for high-altruism groups in mobile societies?

## 4.4 Religious Signaling and the Costly-to-Fake Principle

Religious involvement entails costs (e. g., sacrificing energy and time in order to perform mandatory rituals) as well as returns (in this world or elsewhere). In a standard signaling model, individuals evaluate the net benefit of producing a signal to the effect that agents of one type find it more and agents of the other type find it less worthwhile to do so. In a separating equilibrium, the extent to which an agent produces the signal allows to infer her type.[32] Applied to the evolution of altruism via religious signaling, the story would be that joining a religious congregation is worthwhile for cooperators while it is not for defectors and membership can hence serve as a fraud-resistant signal of their intentions (see e. g. Irons (2001) or Sosis & Alcorta 2003).

It is intuitive to assume that a genuine interest in the activities carried out in the realm of religion and hence religious participation correlates heavily with an individual's commitment to cooperate in the one-shot PD: After all, cooperators in a PD game unmistakably demonstrate their readiness to content themselves with lower-than-possible material wealth, which distinguishes them from defectors. If some consistency in behavior can be presupposed one should expect cooperators to be less deterred than defectors by the time and energy consuming nature of religious rituals which entail no proximate resource gains in return. Yet the question is of course more involved: If cooperators tend to cluster in the religious milieu and achieve high levels of productivity, nonbelieving defectors can be supposed to take notice. Sosis (2003) develops a model to address precisely the question

[32] The classical reference in economics is, of course, Spence (1973) who shows that the amount of education taken by an individual can serve as a signal for her abilities.

of why defectors do not to engage in religious activity despite the benefits associated with the more productive climate in religious communities. In his example of the Hutterites — a religious group living in rural North America which enjoys "extraordinary reproductive success" — it becomes obvious that being part of a religious group in some way or another entails positive *net* fitness advantages over nonmembers. From an evolutionary point of view one would hence expect believing cooperators as well as nonbelieving defectors to join the group regardless of the costly rituals mandated. Sosis resolves this puzzle by arguing that individuals' *subjectively perceived* costs of these rituals are central to the joining decision rather than type-independent resource costs. Subjectively perceived costs can in fact differ across types because an individual's actions and her beliefs are not independent from each other: Self-attribution and cognitive dissonance theories[33] suggest that defectors incur psychological costs when performing rituals which believing cooperators do not: While believers' actions and attitudes are in line, defectors suffer from a conflict between the two. Sosis in his static model concludes that religious communities can adjust their ritual requirements to a level where the representative cooperator joins the community whereas the representative defector stays out.

## 4.5 Summary and Discussion

How can group selection come into play in human societies, especially in those with high degrees of both geographic and social mobility? John Maynard Smith's mice dwell and interact in isolated haystacks by definition, some of which contain larger proportions of cooperative strategies than others. This chapter has picked up the idea that such assortation can obtain in human societies when a player's cooperative intent is discernible

[33] In short, self-attribution theory is about how (overtly or covertly) performed actions create and mould attitudes which allow an individual to make sense of herself. Cognitive dissonance theory maintains that when actions are in contradiction with attitudes, individuals (consciously or unconsciously) strive to reduce the resulting discomfort by changing their values or actions. For a demonstration of how these (and other) findings from psychology advance our understanding of the economy see, e. g., Schlicht (1998).

from the outside. When potential opponents can base their actions on the display of a signal such as Frank's telltale clues, cooperators have at their disposal a means of finding like-minded cooperators in otherwise anonymous sporadic interactions. In order to take the telltale clues avenue to group selection seriously one should of course be prepared to respond to the critique of assuming away the problem. The rhetoric of "green beards" is due to Dawkins (1989), who disguised the assumption of cooperators monopolizing a signal of cooperative intentions as illegitimate. Frank's defense is on empirical grounds, not only by referring to physiognomic and psychological regularities but also by pointing to experimental results (see Frank *et al.* 1993). These findings do not however make it less worthwhile to explore the possibility of telltale clues evolving in competitive environments rather than taking them as givens (the more so as there exist 'manuals' like Ekman (2002) for improving one's ability to detect lies, which might in turn help potential liars improve on their lying skills). Findings from the agent-based model by Macy and Skvoretz suggest that reliable telltale clues can in fact obtain endogenously when interactions are "embedded" to a sufficiently high extent. As pointed out by Max Weber, small-scale religious communities lend themselves as examples of where embeddedness might come from. In modern societies, however, religious affiliation can hardly be taken as granted. Instead it is natural — at least from an economist's perspective — to view religious participation as creating a costs-benefit tradeoff. A static model concerned with such a tradeoff has been reported, which motivates the need for a dynamic model of "religious signaling".

# Chapter 5

# Religion as a Seed Crystal for Altruism: An Ecological Model

Turning to an evolutionary account of within-group altruism and hence the competition for resources earned in the PD, the implications of Sosis's model are clear: When ritual requirements can be made so intense that all defectors prefer to stay among themselves, cooperators in the religious milieu can not only flourish but will eventually crowd out defectors. The population would in the long run converge to an all-Hutterite state as their "extraordinary reproductive success" would be indefinitely sustained. Such dynamics driven by homogeneous groups are however not entirely satisfying, as mutations would drastically change the picture: When a firmly nonbelieving defector (resisting cognitive dissonance) gravitates towards the excellent opportunities for exploitation in the religious milieu, the community will get undermined from within. The signaling value of religion reduces to zero as evolution proceeds, and the long-run equilibrium population will consist entirely of (nonsignaling) defectors.

In the Sosis model, cooperators and defectors arrive at opposite decisions about joining the community or not purely because of the psychological costs imposed on defectors:

Material benefits as well as costs are taken to be the same for both types. A natural way to extend the model and integrate the idea of religious signaling into a full-fledged ecological model would be to endogenize the resource (opportunity) costs and benefits of religious involvement on the basis of forgone and gained earnings in the PD interactions. Such a model will proposed in the following sections. Differential survival and reproduction on the basis of earnings and the divergent characteristics of believers' and skeptics' involvement in religion will simultaneously be taken into account by drawing on the indirect evolutionary approach.

## 5.1 The Indirect Evolutionary Approach

Sociobiology takes behavior as hard-wired, i. e. unaffected by conscious choice, and makes predictions about which behavior will be observed in the long run according to some selection criterion. A natural extension in the social sciences is the indirect evolutionary approach, introduced in Güth & Yaari (1992), which puts preference-based strategic choices to an evolutionary test: Players in indirect evolutionary analyses are supposed to make rational decisions according to some subjective preference relation over outcomes, where preference orderings are not ex-ante restricted to be monotonic in the earnings from interactions as given by the earnings matrix. Competitive pressures then operate on different preferences as embraced by different individuals, which allows to rank preferences with respect to their relative success and thus to assess their evolutionary viability in the given context.

The indirect evolutionary approach has of course been applied to the one-shot Prisoners' Dilemma game. Yet introducing an "extra value" for cooperative behavior into some players' utilities, as in Ockenfels (1993) and Guttman (2000), does not *per se* facilitate the evolution of altruism: As shown by these authors, the evolutionary viability of altruism requires players who derive extra utility from cooperating to have reliable information (at least in a probabilistic sense) about their opponent's preference type in order to flourish.

This finding relates nicely to the argument advanced by Robert Frank (see section 4.1), but do not speak to anonymous interactions and is therefore of limited relevance in the context taken up here.

To set the scene for what will be proposed in the next sections it is instructive to note that the models by Ockenfels and by Guttman have all players act rationally, the only difference between cooperators and defectors being cooperator's transformed utility functions which include extra values for cooperative behavior. In fact, most indirect evolutionary analyses share the logic that all players are taken to make rational choices in the sense of maximizing expected utility, with surprisingly few exceptions (like e. g. Banerjee & Weibull (1994) and Sethi 1996). The model to be developed here adds to the number of these exceptions as it accommodates agents of the 'rational' and hard-wired type alike. By allowing for informed decision-making in parts of the population and for programmed actions in others, the model will allow to simultaneously consider learning and imitation on the one hand and differential survival and reproduction on the other. Both mechanisms imply that behaviors which earn above average grow at the expense of those with below-average performance, and are usually perceived by economists as substitutes (in motivating the axiom of earnings monotonicity for utility functions) rather than complements (in allowing for a richer model of social interactions). Anthropologists, in contrast, have begun to argue that what happens in people's minds and what happens to people's genes should be seen as a coevolutionary process (see the summary of recent work in this area provided by Richerson & Boyd 2005). In other words, cultural and natural selection interact.

The model to be proposed in the following section relates to the idea that cultural and natural selection should be seen as complements. Yet unlike in the literature on cultural group selection, very little reference will be made to the possibly different *characteristics* of the cultural transmission and natural selection — see e. g. Henrich (2004) for examples of how the specific properties of cultural selection *per se* can foster cooperation by reducing the within-group and increasing the between-group components of behavioral

variance. Considering both natural selection and purposeful imitation in an indirect evolutionary model with a mix of maximizing and hard-wired types will rather be taken as an opportunity to assign each of the two selection mechanisms a specific *domain* of operation. This is important because religious involvement by believers will be argued to differ in a specific regard from religious involvement by nonbelievers. The activity concerned is of course the same, yet a difference resides in agents' motivation to partake in religious activities: Believers, by definition, engage in religion for its own sake, while for nonbelievers religion is an instrument they 'use' in order to attain some other ultimate goal. Put shortly, believers are 'hard-wired' to engage in religion and hence evolve by natural selection, while nonbelievers 'culturally' arrive at their maximization goal by imitation and learning.

## 5.2 The Model

Consider a society where a large number of players meet in pairs to engage in symmetric PDs of the repeated one-shot type. Each period is long enough for individuals to engage in the PD *twice* or spend half of their time in a religious community. Religious involvement yields no earnings, so that in the latter case earnings are derived from interacting in the PD once. The economic and the religious spheres are closely tied to each other in that PD interactions occur only among individuals who are part of the religious community on the one hand and among individuals who are not on the other hand, but not across groups.

### 5.2.1 Agents

There are two dimensions to agents in this society: Their motivation with regard to religious involvement and their (pure) strategy played in the PD. Agents are grouped into

three categories with regard to their motivation for religious involvement:

1. *Believers* who see religious activity as an end in itself and partake in religious activity independently of any possible (proximate or ultimate) material rewards. In other the words, these agents are "intrinsically motivated" (Deci & Ryan 1985) to be involved in religion.

2. *Opportunists* inclined to use religious involvement as a means to an end, more precisely as a way to achieve higher earnings in the PD interactions. These agents have "extrinsic motivation".

3. *Ignoramuses* who are repelled by religion so much that they avoid any possible contact. With regard to religion they are nonmotivated.

Agents are genetically conditioned to either cooperate or defect in the PD as in sociobiology. It has been argued above with reference to consistency in behavior that an (individually costly) commitment to cooperate and a belief in (time-consuming yet materially unrewarding and hence costly) religious teachings and activity will plausibly coincide. Believers are therefore taken to be cooperators — an assumption which is in line with recent experimental results by Tan & Vogel (2005) who find that participants' (self-reported) religiosity correlates heavily with both their trustworthiness and their trusting behavior. The important qualification that religious involvement may after all be rewarding in resource terms due to its possible signaling value is accounted for by imputing a defective strategy to opportunists. Ignoramuses can safely be assumed to be defectors.[34]

[34] As in many evolutionary analyses, the types (or, in this context, combinations of religious motivation and PD strategies) *not* to be included in the model are at least as important for the results obtained as the strategies which are. The crucial point here is of course that defecting believers are not admitted. Leaving these types out of the model can be defended with reference to self-perception and cognitive dissonance theories as in Sosis's model (see section 4.4). It is even more legitimate to do so as the presence of opportunists will ensure that results will not be driven by homogeneous groups.

### 5.2.2 Timing

The model is in discrete time. More precisely, each generation $g \in [1, \infty)$ comprises a finite number of $N > 1$ interaction periods indexed by $t \in [1, N]$ as illustrated in figure 5.1.

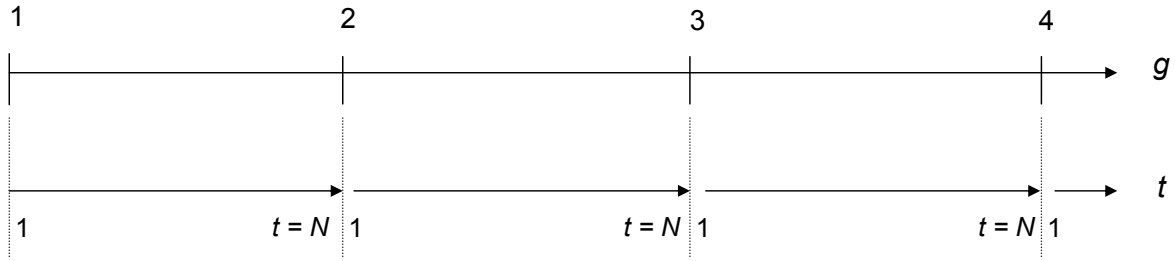


Figure 5.1: Dual Time Axes

The breakdown of generations to interaction periods is essential to the model as it allows to assign separate domains to natural selection on the one hand and to conscious learning on the other: Conscious learning and imitation are taken to go on from period to period (within generations), while natural selection operates from generation to generation. This amounts to the (intuitive) conjecture that evolution by learning is faster than by natural selection.

### 5.2.3 Earnings of Individuals

Let $p_{j,g}$ denote the proportion of individuals of type $j$ present in generation $g$ and consider the interaction period indexed by $t \in [1, N]$ in this generation. Individuals' earnings depend, of course, on which group they are in and on the composition of groups. While believers (type 1) and ignoramuses (type 3) are unambiguously inside and outside of the religious community during the entire generation, respectively, let $p_{21,t,g}$ and $p_{23,t,g} = p_{2,g} - p_{21,t,g}$ denote the population proportions of opportunists who are inside and outside of the religious community in interaction period $t$ of generation $g$, respectively. This distinction is necessary because not all opportunist have joined the

community by any given period within a generation. Unlike with (intrinsically motivated) believers, two prerequisites must instead be fulfilled: Being in the religious group must be worthwhile for an opportunist in resource terms and, equally important, she must have made up her mind to become aware of that fact. Average earnings of individuals with this distinction read

$$\begin{aligned} E_{1,t,g} &= \left(\frac{p_{1,g}}{p_{1,g}+p_{21,t,g}}\right)R+\left(\frac{p_{21,t,g}}{p_{1,g}+p_{21,t,g}}\right)S, \\ E_{21,t,g} &= \left(\frac{p_{1,g}}{p_{1,g}+p_{21,t,g}}\right)T+\left(\frac{p_{21,t,g}}{p_{1,g}+p_{21,t,g}}\right)P, \\ E_{3,t,g} &\equiv E_{23,t,g}=2P \quad \forall\, t,g. \end{aligned}$$

(Notice that $\forall\, p_{1,g}, p_{21,t,g}$, $E_{21,t,g} > E_{1,t,g}$ reflecting the dominant-strategy property of the PD where $T > R > P > S$.)

### 5.2.4 Opportunists' Within-Generation Dynamics

In the model it is assumed that natural selection operates only from generation to generation. What evolves within a generation is the population ratio of opportunists who have joined the religious group by the interaction period indexed by $t$, $p_{21,t,g}$ (recall that believers are assumed to be in the religious group at all times within a generation just like ignoramuses are outside).

A plausible equation of motion reflecting the dynamics which arise from opportunists' decision-making in period $t$ will be argued to be

$$\begin{aligned} p_{21,t=1,g} &= 0 && \forall\, g, \\ p_{21,t+1,g} &= \min\left\{p_{21,t,g}+\alpha\left(p_{1,g}+p_{21,t,g}\right), p_{2,g}\right\}, \quad \alpha\in(0,1) && \text{if} \quad E_{21,t+1,g} > E_{23} \equiv E_3 = 2P, \\ p_{21,t+1,g} &= p_{21,t,g} && \text{otherwise.} \end{aligned}$$

These dynamics, of course, merit discussion with respect to at least two aspects. First,

it is advocated that opportunists — unlike believers who engage in religious activity for the sake of its own — are not present in the religious group at the beginning of each generation. In fact, why should they? For opportunists to enter the religious milieu in view of the profit opportunities reflected by $E_{21,t+1,g} > E_{23}$, they must be aware of the high-altruism qualities of the community. This information is taken to be not inherited from opportunists in $g-1$. This is crucial for the model but in line with the finding in evolutionary biology that characteristics *acquired* during an animal's lifetime is not inherited by its offspring. 'Lamarckian inheritance' e. g. was disproved by genetics, see e. g. Futuyma (1998, p. 18f).[35] Opportunists must as a consequence learn about the exploitation gains from religious involvement in each generation anew, $p_{21,t=1,g} = 0 \quad \forall g$.

The second issue relates to the speed of opportunists' inflow into the religious community. Here it is natural to assume that the larger (and hence more visible) the church — measured by $p_{1,g} + p_{21,t,g}$ — the more opportunists learn of the religion's signaling value in each period. The earnings differential $E_{21,t+1,g} - E_{23}$ is taken to be relevant only with regard to its sign: Opportunists who get aware of the religion-altruism link will enter the church as long as the gains to be achieved through such a move are positive, independent of their magnitude. Opportunists' inflow into the religious community terminates when a generation ends, when an additional dose of opportunists would push their within-church earnings below the level to be earned outside or when all opportunists have joined the community, depending on what happens first.

Figures 5.2 - 5.4 all give an impression of the initially convex shape which results when plotting $p_{21}$ against the within-generation time axis $t$. In figure 5.2, a generation lasts sufficiently long and the gains from exploitation remain sufficiently high for all opportunists to make their way into the religious community, whereas the generation in figure 5.3 ends before all opportunists get involved in religion. In figure 5.4, opportunists have enough

[35] On a side note, the author knows of parents in the process of raising children who strongly agree (the parents, not the children). See however e. g. Jablonka & Lamb (1995) for the concept of "epigenetic inheritance" of non DNA-encoded characteristics by cells.

time on their hands but elect not to exhaust it: Due to the small fraction of believers in this plot, the cooperative milieu deteriorates to an extent that lets opportunists' inflow run dry before the generation ends.

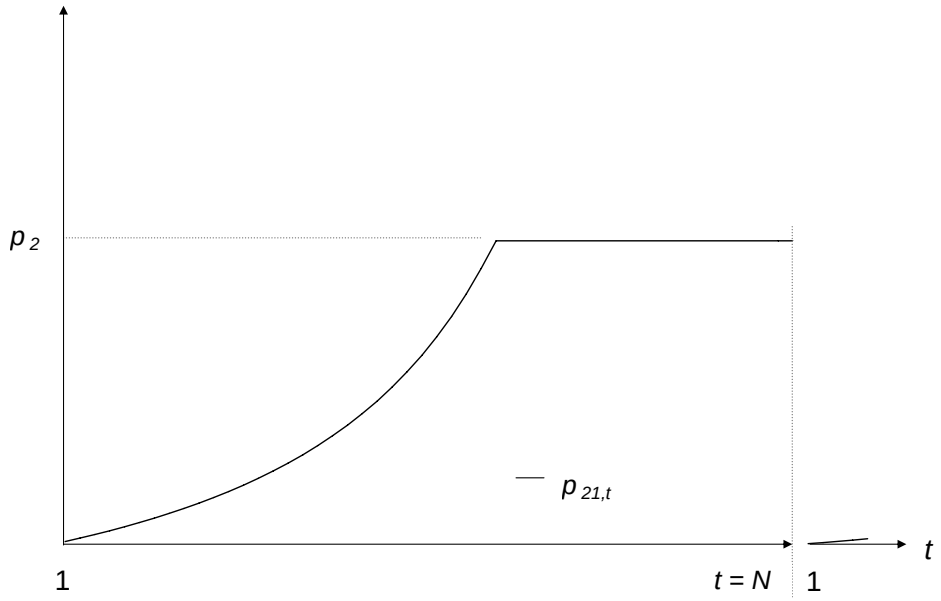


Figure 5.2: Opportunists' Within-Generation Dynamics ("Long" Generation)

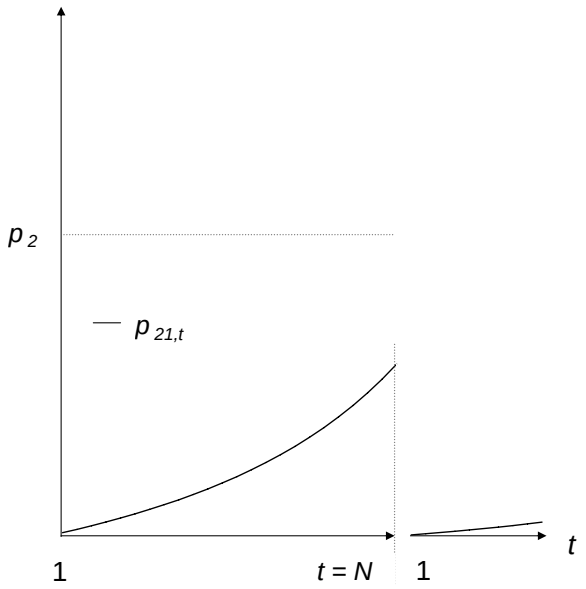


Figure 5.3: Opportunists' Within-Generation Dynamics ("Short" Generation)

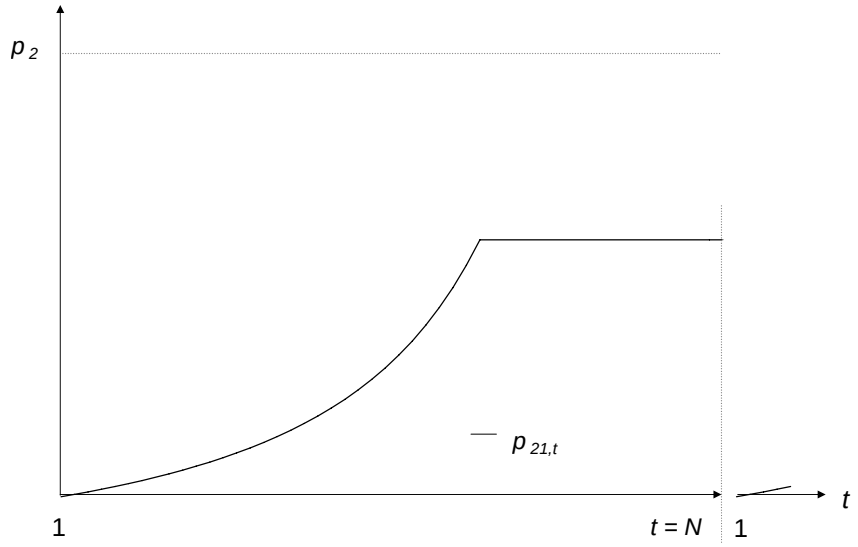


Figure 5.4: Opportunists' Within-Generation Dynamics ("Lack" of Believers)

Figures 5.2 - 5.4 each depict a single generation. Only the early interaction periods of the subsequent generation are indicated. This is because generations in the convergence process differ from one another with regard to population ratios. The next section details.

### 5.2.5 Natural Selection

Natural selection by differential survival and reproduction intervenes at the end of each generation, i. e. after $N$ interaction periods. A convenient 'accounting level' for fitness differences in the society is *types'* (cumulated) earnings. While per-period earnings of types are identical to individuals' earnings for believers and ignoramuses, a typical opportunist's earnings in period $t$ must take into account the fraction of opportunists who are within and who are without the community. Types' per-period earnings hence read

$$\begin{aligned} \hat{E}_{1,t,g} &\equiv E_{1,t,g}, \\ \hat{E}_{2,t,g} &= \left(\frac{p_{21,t,g}}{p_{2,g}}\right) E_{21,t,g} + \left(\frac{p_{23,t,g}}{p_{2,g}}\right) E_{23}, \quad \text{and} \end{aligned}$$

$$\hat{E}_{3,t,g} \equiv E_{3,t,g} = 2P \quad \forall\, t, g.$$

Types' cumulated earnings $\Pi_{i,g}$ in generation $g$ can then be computed as

$$\Pi_{j,g} := \sum_{t=1}^{N} \left( \frac{1}{1+\delta} \right)^{t-1} \hat{E}_{j,t,g}, \quad j = 1,2,3,$$

where $\delta \geq 0$ is a discount rate reflecting the possible fact that earnings realized earlier in life play a disproportionately important role.

Natural selection's workings on types are modeled by standard discrete-time replicator dynamics (see e. g. Weibull 1995, p. 123):

$$p_{j,g+1} = p_{j,g} \cdot \frac{\Pi_{j,g}}{\bar{\Pi}_g} \quad , j = 1,2,3,$$

where $\bar{\Pi}_g := \Sigma_{j=1}^{3} p_{j,g} \Pi_{j,g}$ is the weighted population average of types' cumulated earnings in generation $g$.

## 5.3 Simulation Studies

The society outlined above is currently implemented as a computational model in Microsoft® Visual Basic for inductively analyzing its dynamical properties. In particular, the evolution of the population proportion of types, $p_{j,g}$, can be studied for different initial conditions $p_{j,g=1}$ (of course satisfying $\Sigma_{j=1}^{3} p_j = 1$) and different parameter values such as $N$, $\alpha$, $\delta$, and for different earnings matrices. Some simulation results are reported in what follows.

Concerning the evolution of within-group altruistic cooperation, the most important finding is that believers and opportunists can coexist in long-run equilibrium provided that the effect of believers' initial advantage of an intermittently intact religious milieu is strong enough. For a 'benchmark' case, payoffs are taken to be $T = 13, R = 11, P = 3, S = 1$,

which implies a cost-benefit ratio of cooperation of $\frac{c}{b} = 0.2$. Moreover, $N = 50$, $\alpha$ =0.025, and $\delta = 0.01$.

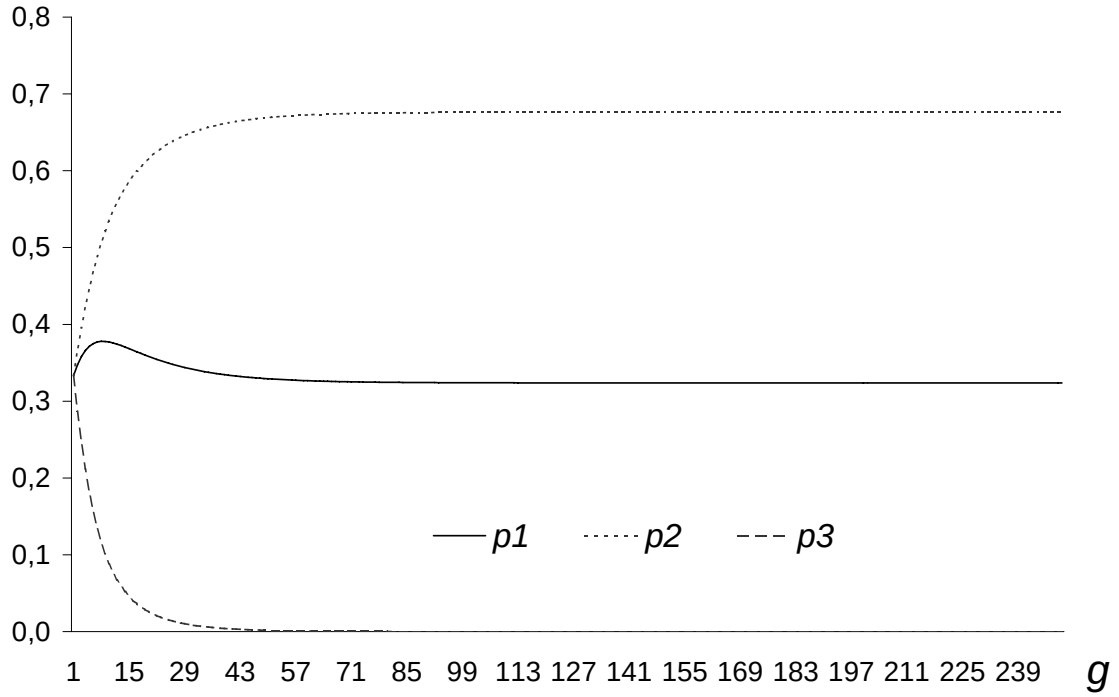


Figure 5.5: $T = 13$, $R = 11$, $P = 3$, $S = 1$, $N = 50$, $\alpha$ =0.025, $\delta = 0.01$

From an initially even distribution of types in the population, the society evolves to a long-run equilibrium where believers and opportunists coexist while the frequency of ignoramuses reduces to zero.

An important question to ask at this point relates of course to the robustness of these findings with regard to changes in initial conditions. A simulation run with different initial proportions was therefore conducted. In Figure 5.6, $p_{1,g=1} = 0.1$, $p_{2,g=1} = 0.8$ and $p_{3,g=1} = 0.1$.

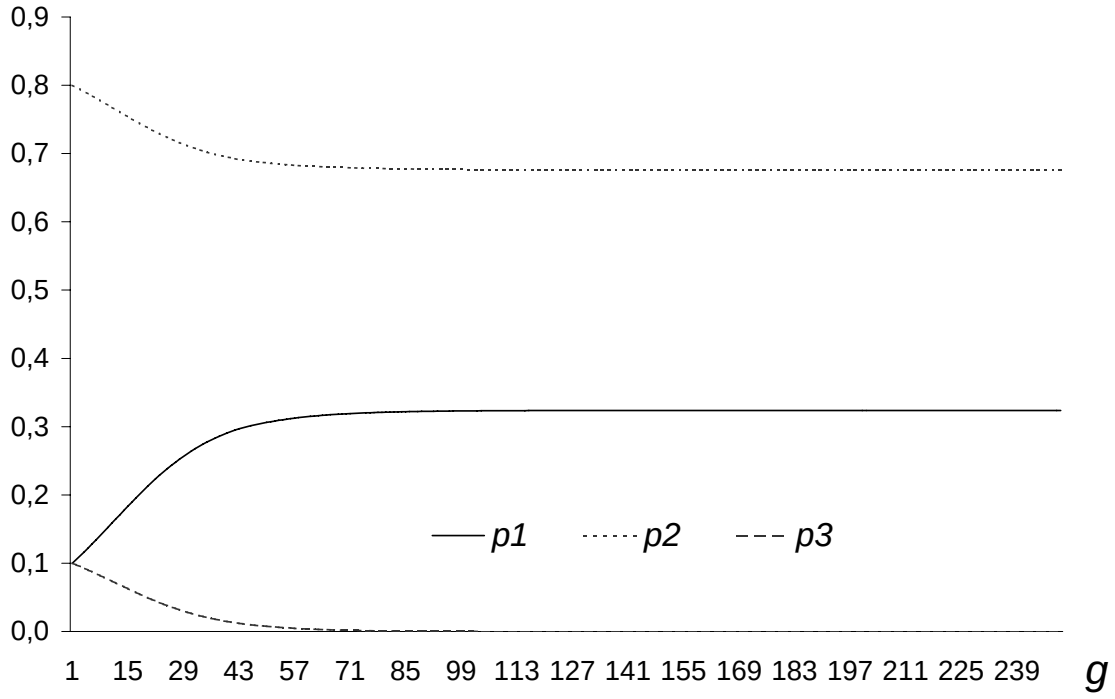


Figure 5.6: Irrelevance of Initial Conditions

Inspection of figure 5.5 and 5.6 suggests that the long-run distribution of types is independent of initial conditions. As for the convergence process, all types evolve monotonically in figure 5.6 whereas believers initially spread (for a short while) and subsequently decline towards their long-run proportion in figure 5.5. The non-monotonic shape in this figure results from the fact that believers' initial proportion is higher than their long-run ratio in this exercise on the one hand and from the replicator dynamics on the other: Opportunists as long as present in the population push down the average earnings to an extent that both believers along with opportunists earn above average in early stages of the simulation. Believers can hence for a short while grow in proportion before dropping (slightly) below their initial level.

Simulation studies make it possible to trace not only the evolution of population ratios with regard to types but equally the shape of opportunists' inflow into the religious community — which had been illustrated for single generations in figures 5.2 - 5.4 — across multiple generations.

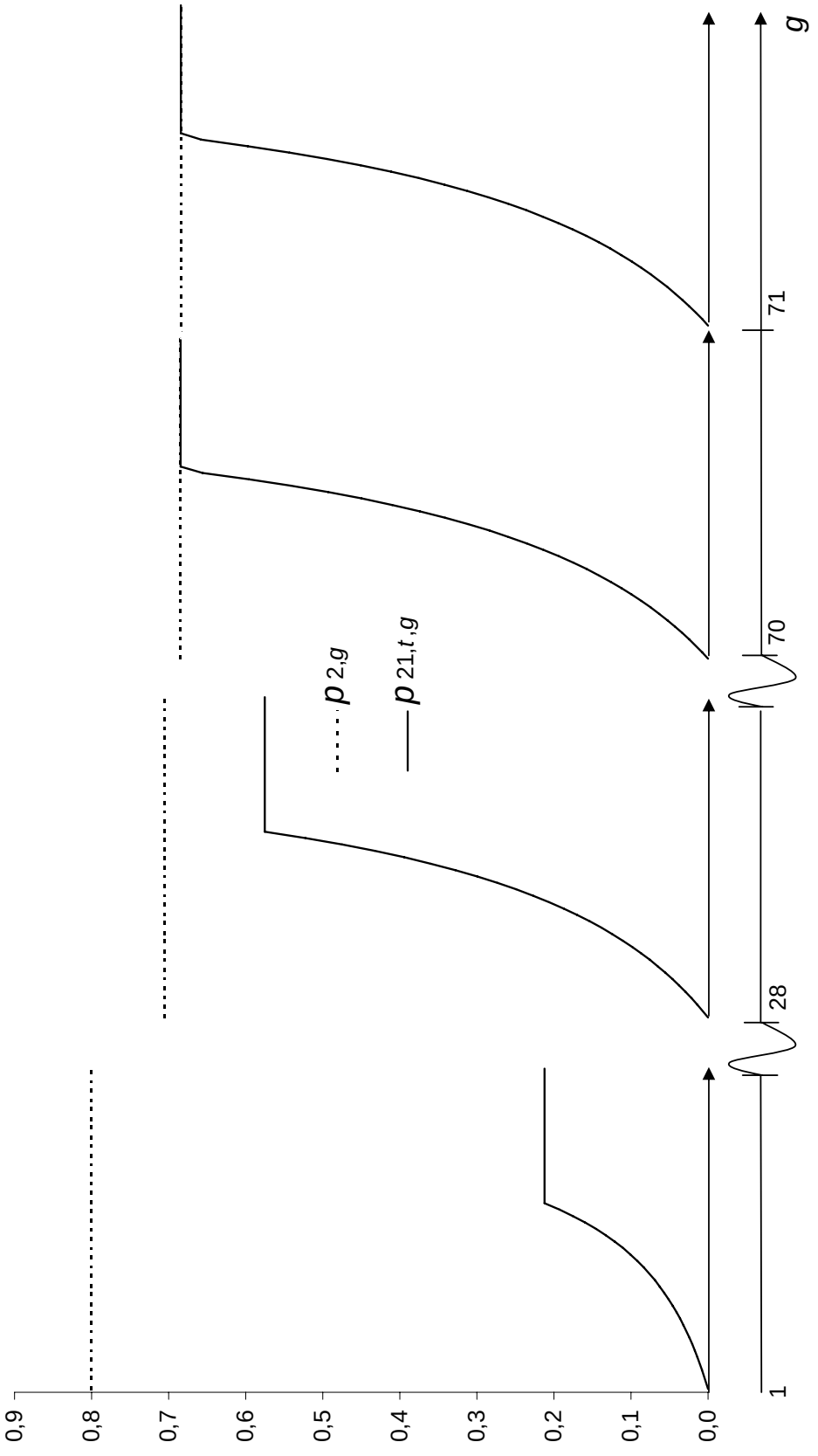


Figure 5.7: Opportunists' Inflow into Religious Community

Figure 5.7 depicts how opportunists make their way into the church for (initially) non-consecutive generations such that the differences between early and later generations become visible: Early in the simulation, opportunists flow into the religious community rather slowly, and the inflow runs dry before a generation ends. Both the moderate speed and the premature stop of opportunists' inflow are due to a small initial population proportion of believers, which amounts to only $p_{1,g=1} = 0.1$. Believers can spread at the cost of opportunists, whose aggressiveness is initially moderate. Due to the expanding size of the church, opportunists can in later generations catch up as they flow into the community more and more quickly and for longer periods of time. The process eventually equilibrates in a state where opportunists flow into the church at a speed which in this case allows all opportunists to spend part of a generation in the church.

How does the model respond to changes in parameter values? Consider the next simulation run where all parameters are as in Figures 5.5 and 5.12, expect for the "length" of a generation which is taken to be 60 instead of 50 interactions.

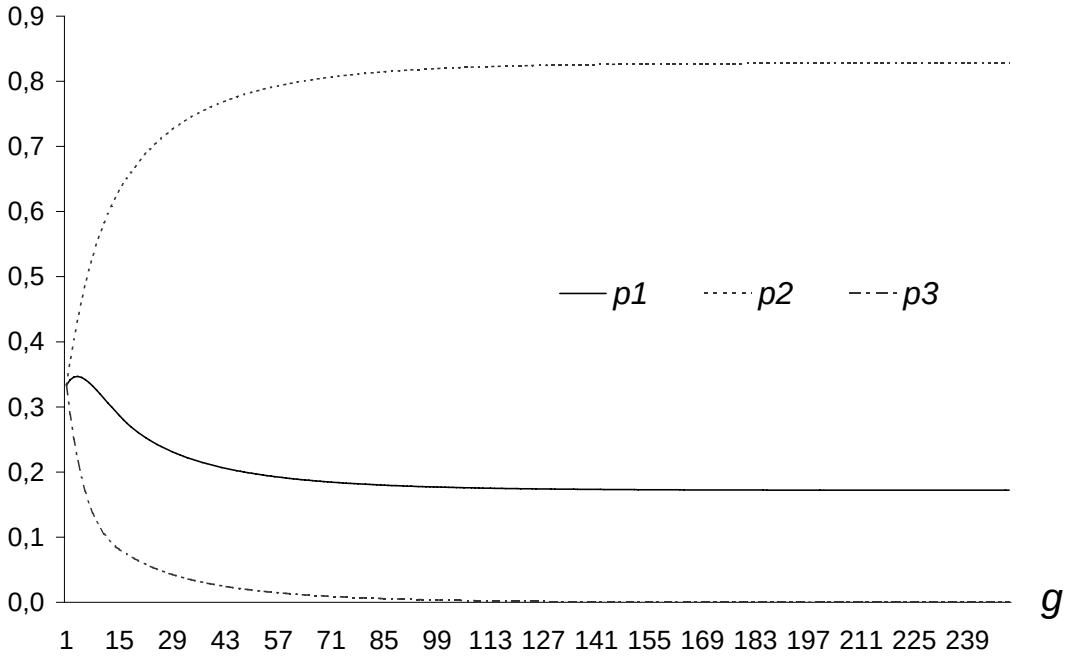


Figure 5.8: $T = 13$, $R = 11$, $P = 3$, $S = 1$, $N = 60$, $\alpha$ =0.025, $\delta = 0.01$

As expected, the long-run equilibrium proportion of believers is lower than in the previous

figures because opportunists have more time to catch up on believers' initial advantage in each generation.

The discount rate $\delta$ plays an intuitive role here. When repeating the two previous exercises with $\delta = 0.02$ instead of 0.01 the equilibrium proportions of believers are $p_1 = 0.46$ for $N = 50$ and $p_1 = 0.37$ for $N = 60$, which both exceed the values for the cases where $\delta = 0.01$. A higher discount rate clearly favors believers.

The next run illustrates the significance of opportunists' learning speed. Parameters differ from the 'benchmark' case in that $\alpha = 0.04$. In this case, opportunists' learning occurs at a speed such that believers are asymptotically crowded out. In this situation, opportunists are behaviorally indistinguishable from the ignoramuses who survive in the long run.

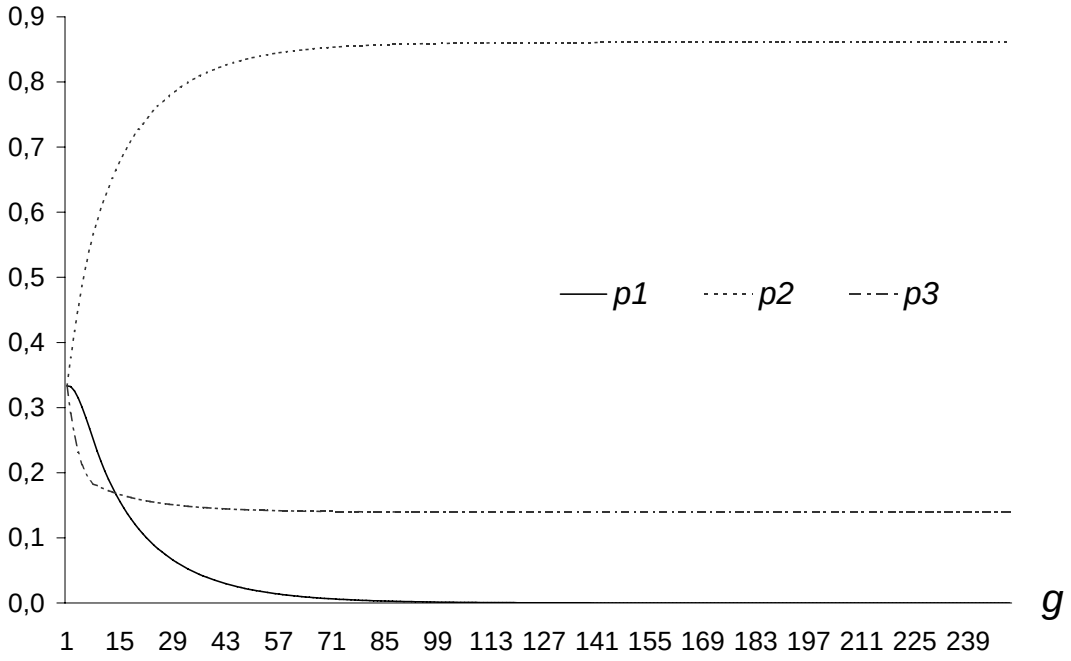


Figure 5.9: $T = 13$, $R = 11$, $P = 3$, $S = 1$, $N = 50$, $\alpha$ =0.04, $\delta = 0.01$

Unlike long-run outcomes where believers survive along with opportunists, the outcome in figure 5.9 is associated with its initial conditions of equal representation of types in the population: Simulations of the same parameters yet with different initial population proportions (not shown) suggest that the long-run proportion of opportunists increases

with *ceteris paribus* a higher initial proportion of believers (to feed on during convergence) on the one hand and with a lower initial proportion of ignoramuses on the other hand.

Believers are crowded out in figure 5.9 because opportunists learn too fast given the other parameters. One way to illustrate the interplay of all model parameters is to increase the discount factor to $\delta = 0.03$, leaving all other parameters constant (including $\alpha = 0.04$). This exercise results in long-run population proportions of $p_1 = 0.22$, $p_2 = 0.78$, $p_3 = 0$ (not shown).

The model can similarly be inspected for its response to changes in the earnings matrix. In the above simulations, earnings were $T = 13$, $R = 11$, $P = 3$, $S = 1$. As expected, results do not change when all earnings are multiplied by some positive number. This is not true, however, in the case of a constant being *added* to (or deducted from) all earnings. Consider the next diagram which depicts a simulation run where — as in the 'benchmark' case — $N = 50$, $\alpha = 0.025$, $\delta = 0.01$ but where earnings are $T = 12.5$, $R = 10.5$, $P = 2.5$, $S = 0.5$. This transformation preserves the cost-benefit ratio of $\frac{c}{b} = 0.2$, yet the long-run distribution of types differs from the 'benchmark' case:

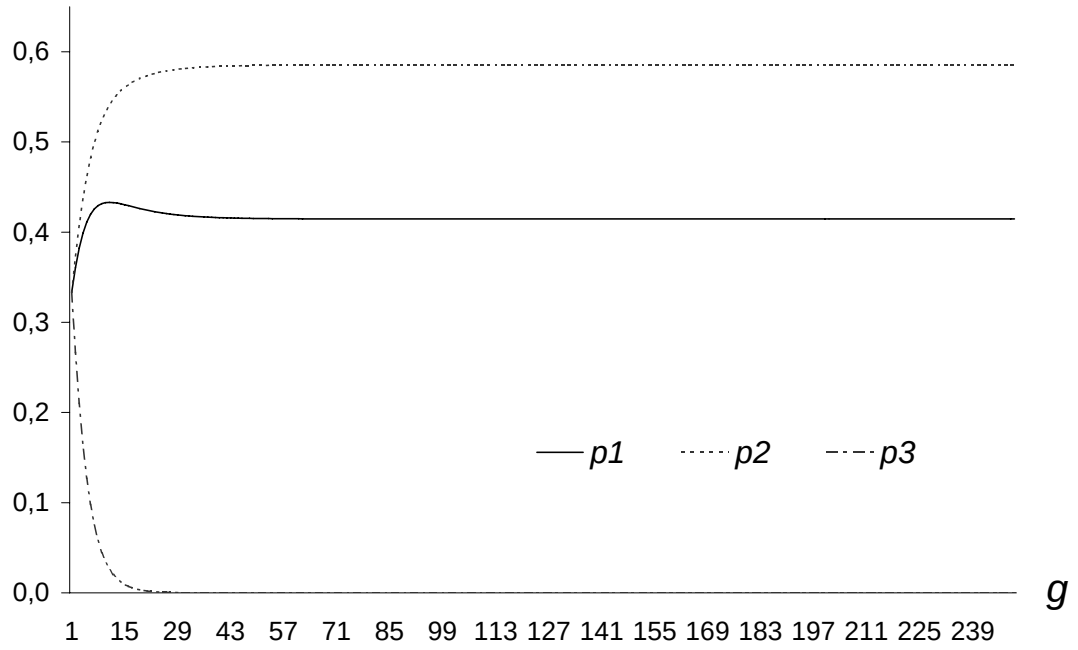


Figure 5.10: $\frac{c}{b} = 0.2$, $N = 50$, $\alpha = 0.025$, $\delta = 0.01$ as in 'Benchmark' Case but Higher $\frac{R}{P}$

The transformed earnings matrix — which has a higher ratio of the earnings from mutual cooperation to those from mutual defection than in the 'benchmark' case — favors cooperators. In contrast, believers disappear in the case (not shown) $T = 17$, $R = 15$, $P = 7$, $S = 5$. In this case, the cost-benefit ratio of cooperation is again 0.2 as in the 'benchmark' case, yet the low ratio $\frac{R}{P}$ disfavors believers.

As for a change in the cost-benefit ratio $\frac{c}{b}$, the model reacts as expected. Consider the last simulation run to be reported, where $T = 12.5$, $R = 11$, $P = 3$, $S = 1.5$. With these earnings $\frac{R}{P} = \frac{11}{3}$ as in the 'benchmark' case, yet the cost-benefit ratio of cooperation is only $\frac{c}{b} \simeq 0.158$ and hence smaller than in the 'benchmark' case. This clearly favors believers, as is obvious from figure 5.11.

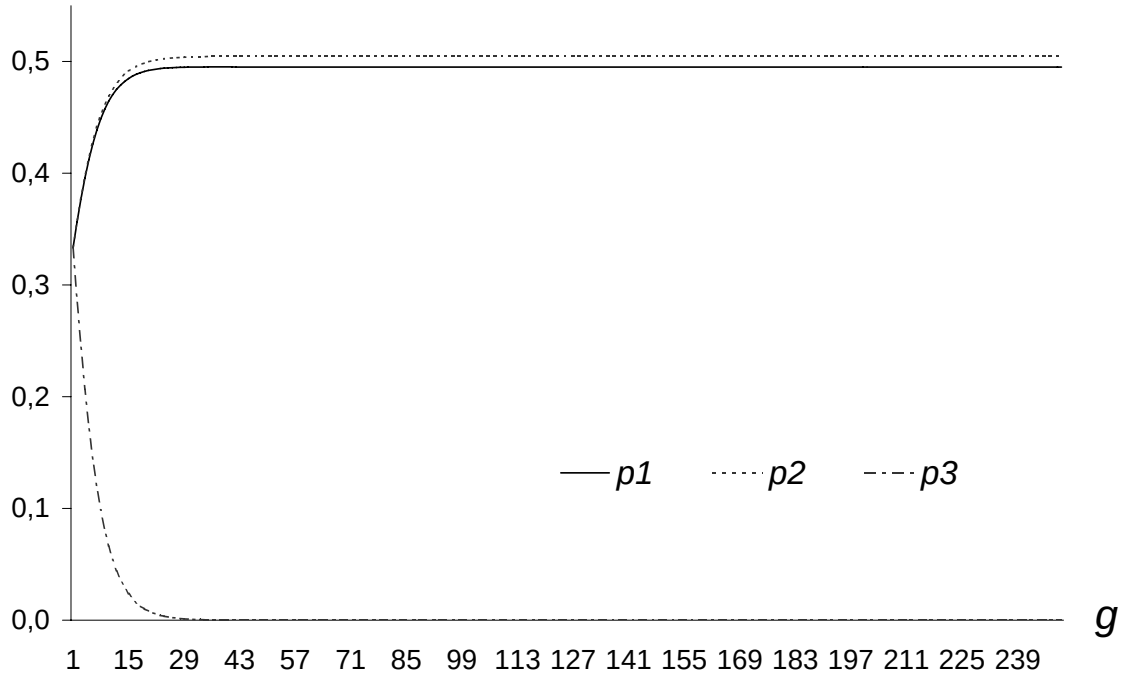


Figure 5.11: $\frac{R}{P} = \frac{11}{3}$, $N = 50$, $\alpha = 0.025$, $\delta = 0.01$ as in 'Benchmark' Case but Lower $\frac{c}{b}$

## 5.4 Exploring the Model

The simulation studies reported in the previous section indicate that the society under study accommodates basically two types of long-run equilibria, depending on the relative importance of believers' initial advantage within generations: Coexistence of either believers and opportunists or (behaviorally indistinguishable) opportunists and ignoramuses. This section further explores the nature of the first, and more interesting, equilibrium of believers and opportunists. A robustness check with regard to the specification of opportunists' within-generation dynamic will be performed afterwards.

### 5.4.1 Equilibrium Properties

Figures 5.5 and 5.6 suggest that long-run population ratios in outcomes with believers and opportunists do not depend on initial conditions. As is evident from figure 5.12, such invariance extends to the case where opportunists rather than getting crowded out by believers and opportunists in the course of time are absent right from the start.

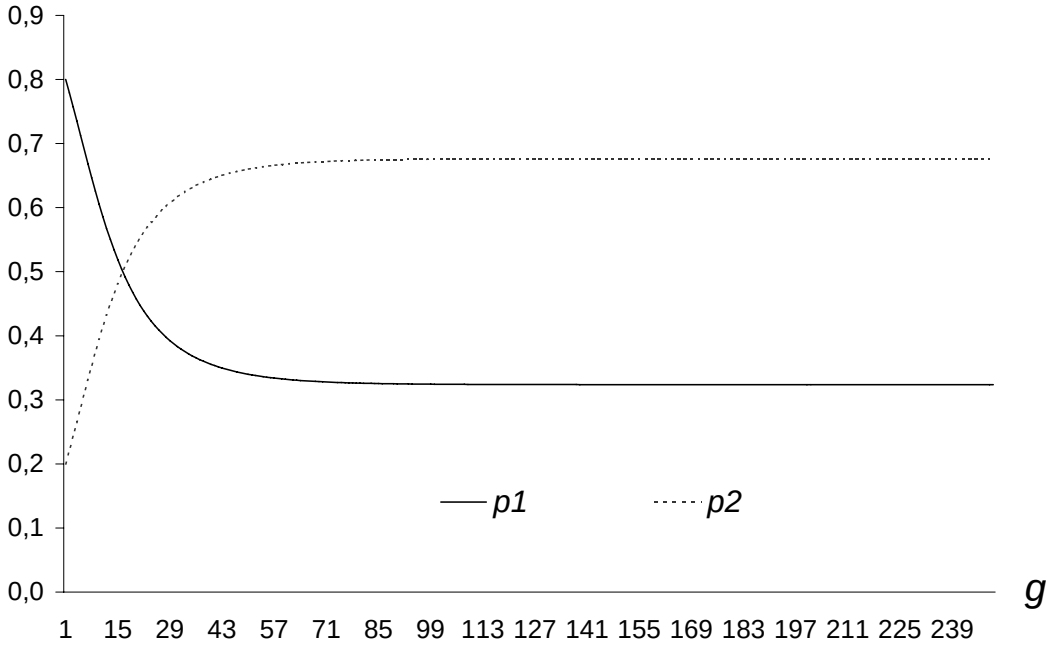


Figure 5.12: 'Benchmark' Case as in Figures 5.5 and 5.6 but $p_{3,g=1,...} = 0$

The presence of ignoramuses is hence irrelevant for the long-run distribution of types in equilibria with believers and opportunists. Uniqueness and stability of the two types' long-run distribution are illustrated in figure 5.13, which provides a phase diagram — disregarding ignoramuses — for the 'benchmark' case with $p_{3,g=1} = 0$.

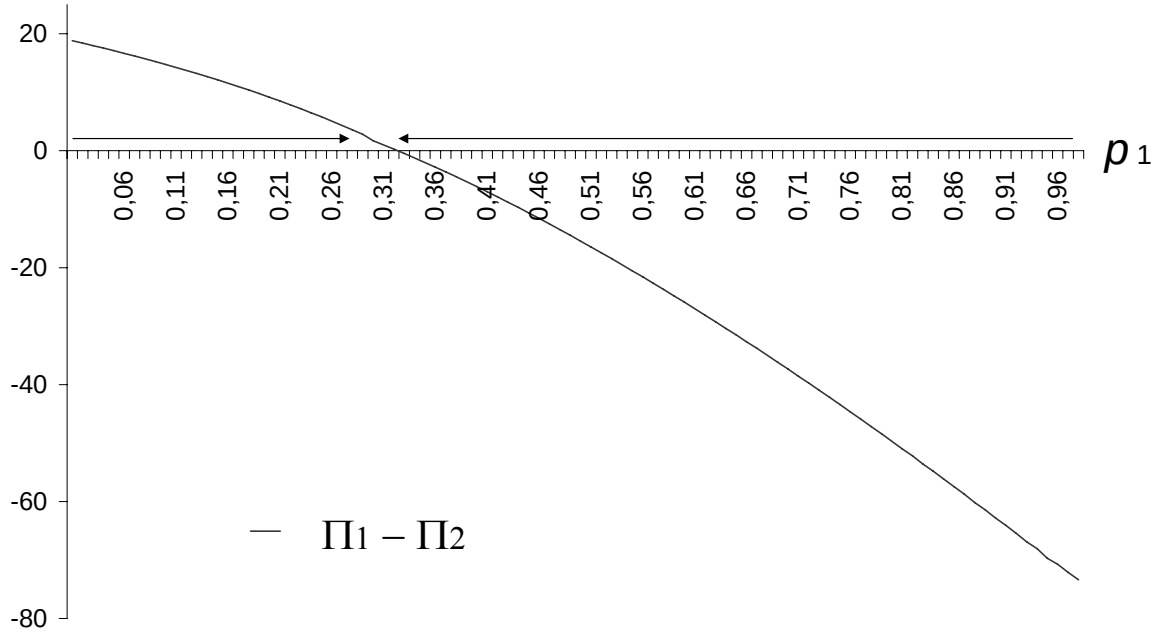


Figure 5.13: Phase Diagram for 'Benchmark' Case ($p_{3,g=1,...} = 0$)

Plotting believers' advantage in cumulative discounted earnings against their population proportion, the graph intersects the abscissa once and from above, thereby manifesting uniqueness and global[36] stability. The phase diagram can also be inspected with regard to parameter changes in order to corroborate the comparative statics exercises reported in the previous section: Plotting the graph for different parameter settings reveals that *ceteris paribus* increases in $N$, $\alpha$, and $\frac{c}{b}$ move the stationary value of $p_1$ inward. Decreases in these variables just like increases in $\delta$ or $\frac{R}{P}$ on the contrary move the stationary value of $p_1$ to the right. As it turns out, drastic parameter changes disfavoring believers can push the entire graph below the abscissa and provide for an equilibrium with opportunists and ignoramuses. Moving the entire graph above the abscissa is however beyond the power of

[36] Of course, global here refers to an initial proportion of believers which satisfies $0 < p_{1,g=1} < 1$.

believer-friendly changes, leaving no hope for a monomorphic equilibrium with believers only.

What is the nature of this finding? In figure 5.7 a flavor was given as for the equilibrating qualities of the particular dynamics governing opportunists' inflow into the religious community: Starting from a population with a relatively small proportion of believers and hence poor visibility of the church, believers can spread in the population because of opportunists' limited aggressiveness. As the fraction of believers grows (and the church becomes more visible due to its increasing size from the beginning of each generation), opportunists' inflow speed increases and hence narrows down the gap in cumulative earnings. Figures 5.14 - 5.16 confirm this intuition in terms of resources by plotting types' per-period earnings within a generation for different population ratios of believers and opportunists (parameters correspond again to the 'benchmark' case).

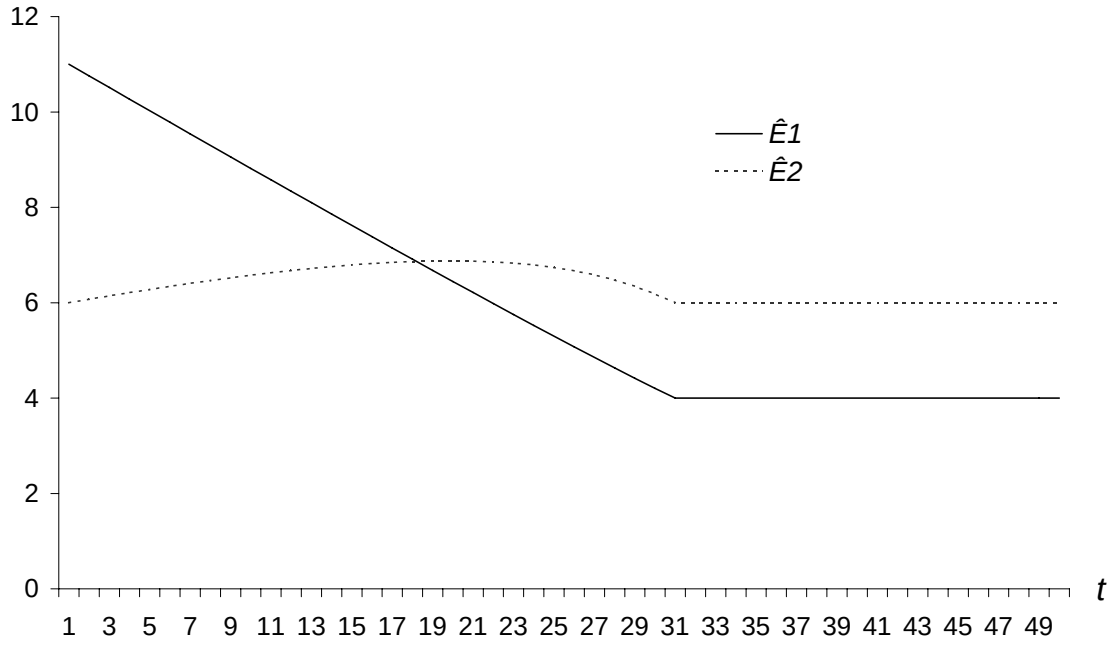


Figure 5.14: Per-Period Earnings of Types, $p_1 = 0.3$, $p_2 = 0.7$

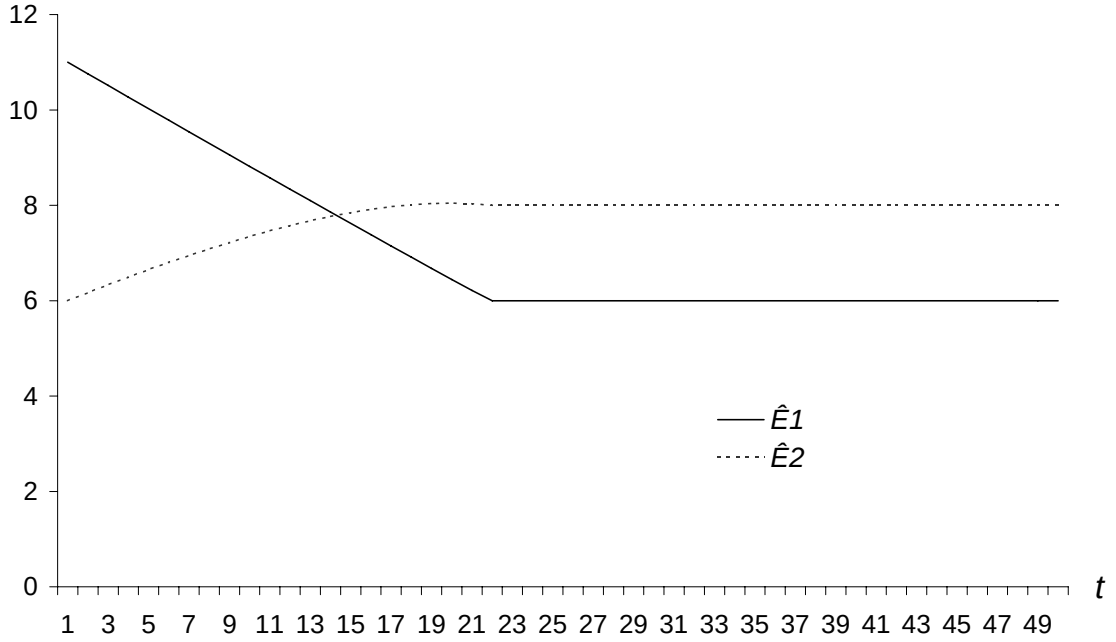


Figure 5.15: Per-Period Earnings of Types, $p_1 = 0.5$, $p_2 = 0.5$

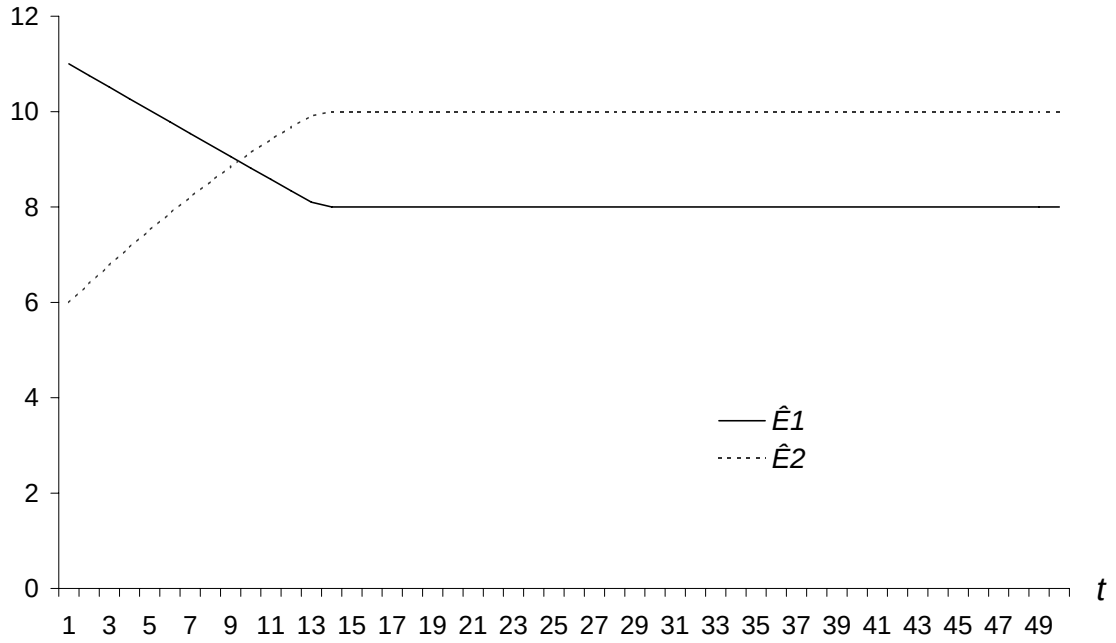


Figure 5.16: Per-Period Earnings of Types, $p_1 = 0.7$, $p_2 = 0.3$

For relatively small $p_1$ (such as in figure 5.14), believers' earnings stay above opportunists' type-level earnings until roughly $t = 19$. With growing $p_!$ the point of intersection of believers' and opportunists' earnings is shifted towards the beginning of generations:

While the (nonlinear, although not obvious from the diagrams) declining portion of the earnings trajectory for believers is unaffected by changing population ratios, type-level earnings for opportunists increase at a higher rate the larger the church, which compromises believers' initial advantage.

### 5.4.2 Model Robustness

The specification of opportunists' within-generation dynamics advocated in section 5.2.4 clearly stabilizes the model and contributes to the uniqueness of equilibria with believers and opportunists. This section explores the role played by the specific dynamics of opportunists' inflow into the religious community by adopting a more parsimonious specification for the purpose of comparison: Rather than having opportunists' inflow speed depend positively on the size of the church, opportunists are taken to learn about the exploitation gains to be made in the religious community (as long as applicable) at a constant rate. Opportunists' within-generation dynamics then become

$$\begin{aligned} p_{21,t=1,g} &= 0 && \forall g, \\ p_{21,t+1,g} = \min\left\{p_{21,t,g} + \beta, p_{2,g}\right\}, \quad \beta \in \left(0, p_{2,g}\right) & && \text{if} \quad E_{21,t+1,g} > E_{23} \equiv E_3 = 2P, \\ p_{21,t+1,g} &= p_{21,t,g} && \text{otherwise.} \end{aligned}$$

The implications of this modification become clear when plotting a phase diagram with $p_{3,g=1} = 0$ for again the 'benchmark' parameter values (except for $\beta = 0.003$). Figure 5.17 reveals that coexistence of believers and opportunists remains a (if only locally) stable equilibrium outcome also with these dynamics and hence does not depend on the specification advocated in section 5.2.4. Yet a second (unstable) equilibrium of the same type arises along with a third (locally stable) long-run outcome where believers get crowded out. Simulation studies indicate that the stable equilibrium with believers is reached equally for $p_{3,g=1} > 0$ as long as $p_{1,g=1}$ is larger than the value where the graph in figure 5.17 crosses the abscissa from below.

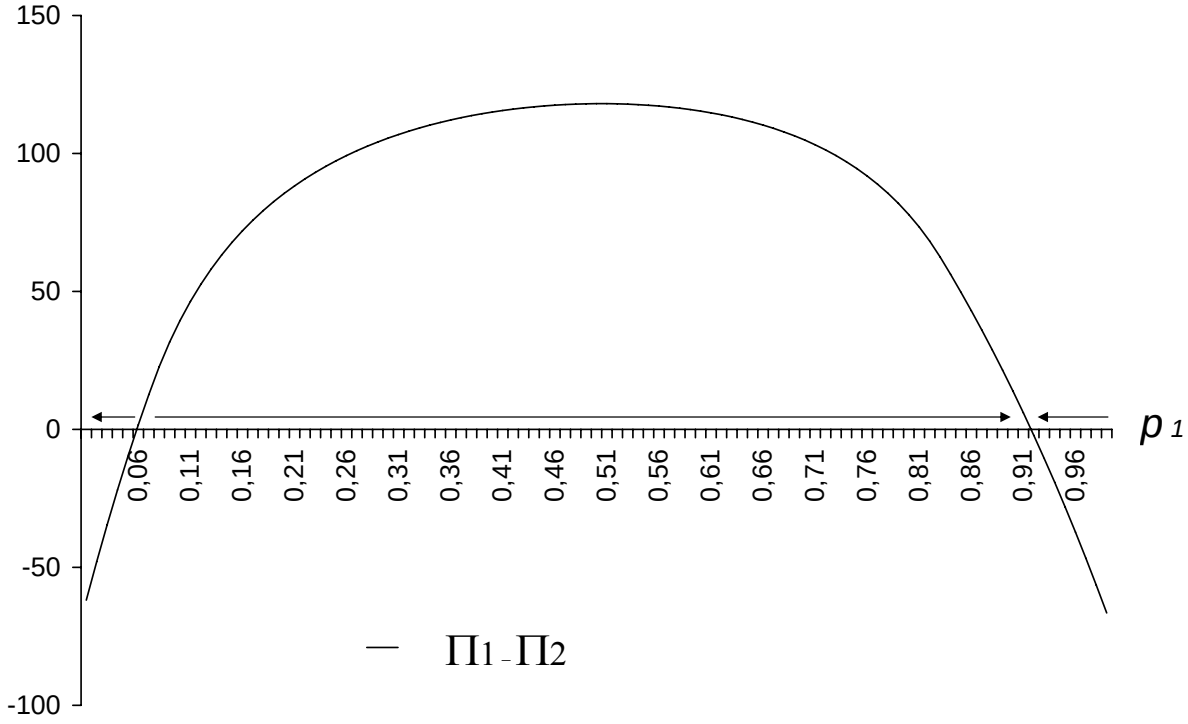


Figure 5.17: Phase Diagram for Alternative Specification ($p_{3,g=1,...} = 0$)

A different modification of opportunists' within-generation dynamics is to deprive them of their capacity to steer clear of the religious community when it no longer pays to join, i. e. to consider the specification

$$p_{21,t=1,g} = 0 \qquad \forall g,$$

$$p_{21,t+1,g} = \min\left\{p_{21,t,g} + \alpha\left(p_{1,g} + p_{21,t,g}\right), p_{2,g}\right\}, \quad \alpha \in (0,1).$$

This modification is innocuous when parameter values provide for an equilibrium where each and every opportunist finds it worthwhile to join the religious community anyway, like in figure 5.7. Consider however the simulation run of the modified dynamics in figure 5.18.

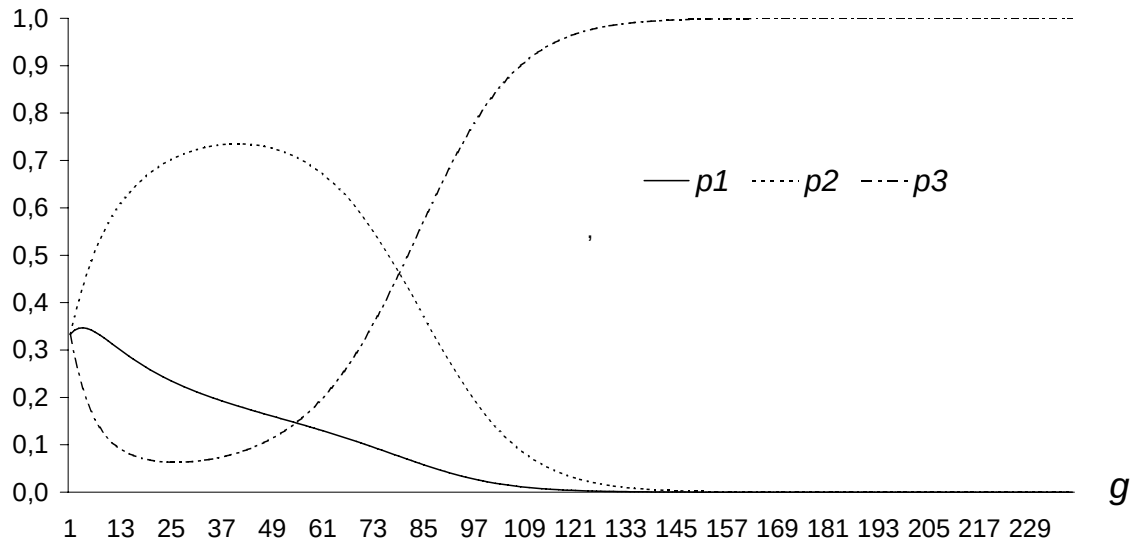


Figure 5.18: Ill-informed Opportunists (Parameters as in Figure 5.8)

Figure 5.18 has the same parameter values as figure 5.8, where believers and opportunists coexist in the long run. Yet by flowing into the religious community not only up to but also beyond the point where religious involvement no longer pays, 'ill-informed' opportunists pave the road to ruin first for believers and then for themselves. The long run sees ignoramuses take over the population. Given parameter constellations like this one, believers hence benefit from opportunists' artfulness. From believers' point of view, opportunists had better live up to their name and be smart if opportunism must be dealt with in the first place.

## 5.5 Summary and Discussion

Signaling by religious involvement is proposed as a group selection mechanism to sustain altruistic cooperation as modeled by the one-shot Prisoners' Dilemma. The main finding is that cooperators (*believers* in the model) and defectors (*opportunists*) can coexist in long-run evolutionary equilibrium. The crucial point consists in believers' material advantage over opportunists which derives precisely from the non-material nature of their

motivation for religious involvement. Defectors can mimic the signal sent by cooperators, but — borrowing the term from Frank (1988, p. 60) — only with *delay*. In this model, the delay results from their opportunistic motivation to use religion instrumentally: Unlike believers who see religious involvement as an end in itself, they have to learn of the exploitation gains from religious involvement during a generation before attaining them. The dual time axes of interaction periods and generations introduced in the model relate to the "two-tiered dynamic process" envisioned in the indirect evolutionary analysis in Güth & Kliemt (1998). While these authors assume that "overall evolutionary success is in the last resort determined by the payoffs earned after adaptation has taken place and is not distorted by what happens in the adaptation phase itself", opportunists' within-generation dynamics are crucial to the model at hand: If opportunists' delay is sufficiently important, believers and opportunists coexist in the long run. Simulation studies suggest that the model responds to parameter changes in intuitive ways.

To avoid misunderstandings and possible offense it should be emphasized that no theory of religion is intended here, although researchers since Azzi & Ehrenberg (1975) have inquired about the economics of religion both theoretically and empirically (see Iannaccone (1998) for a survey and Barro & McCleary (2003) for recent empirical work on the interaction of religion and economic growth). Neither does the model address the evolutionary origins of religion as such or its possible adaptive value (see Boyer (2001) and Wilson (2002) for such analyses). Religion is rather taken as illustration of an activity which (i) does not offer a *proximate* (net) reward in *resource* terms but is (ii) appealing for *non-material* (i. e. not convertible into earnings) reasons of its own. Volunteering, honorary offices and charity also come to mind.

# Chapter 6

# Epilogue

This piece of work is about the evolutionary viability of cooperative play in one-shot Prisoners' Dilemma situations. The precarious nature of cooperation has kept evolutionary biologists occupied for ages, and economists — as soon as they leave the cosy world of general equilibrium with ideally functioning markets — naturally share their concern. After all, both disciplines tend to operate on a maximization hypothesis which implies that it is successful rather than unsuccessful behavior which will prevail in competitive settings, and cooperating in the Prisoners' Dilemma puts the individual in the latter category. Yet the task is maybe easier for economists, who deal with human societies. Humans' capacity to remember faces and prudently deliberate over costs and benefits allows economists to invoke the "shadow of the future" and argue that mutual cooperation can be sustained as an equilibrium outcome in the iterated version of the game.

Such reasoning was shown to be problematic on both theoretical and empirical grounds in Chapter 2: On the one hand, it was argued that classical game theorists' treatment of backward induction and coordination problems leaves much to be desired. On the other hand, the repeated-game argument has no bite when it comes to sporadic interactions. Yet modern market economies lean heavily towards depersonalization and tend to rely on

anonymous transactions in — possibly imperfect — markets. Whenever the goods and services traded in these markets do not live up to the neoclassical ideal of perfect homogeneity (the same goes of course for the integrity of market participants), transactional problems arise and take the form of the one-shot Prisoners' Dilemma.

Both the lessons to be learned from the repeated-game approach and the special interest in the one-shot Prisoners' Dilemma were taken on board in Chapter 3, along with biologists' approach towards the problem. Biologists refer to cooperation in the one-shot Prisoners' Dilemma as an act of altruism and wonder at the evolutionary survival of a behavior which benefits competitors in the fitness race at a cost to the donors. Group selection was offered as an approach to this seemingly paradoxical phenomenon. Group (or multi-level) selection clears the view to appreciate fitness penalties incurred by altruists within subpopulations regardless of altruists' evolutionary viability at the population level. If group selection hence provides a theoretical framework for meaningful analyses of (within-group) altruism, the question remained of how to make the group effect work and be sustained in plausible ways, especially in human societies where canonical haystack models are out of place.

Chapter 4 took up the task and highlighted the possible role of signaling in achieving the between-group variance in behavior which is necessary for the group selection effect. "Telltale clues" were reviewed and reported to endogenously acquire and retain their indicative value in competitive settings when interactions are ex-ante embedded to a sufficiently high extent. In search of a possible factor of such embeddedness, sources were followed which focus on membership in one of Max Weber's Protestant Sects, or on the signaling value of religious involvement in general.

The central message of this piece of work, developed in Chapter 5, is the following: The religion example of signaling lends itself as an avenue to the evolution of altruism in competitive settings also when church membership is viewed as the outcome of a cost-benefit trade-off accessible to cooperators and defectors alike. Cooperators, taken

to believe in religion and view religious involvement as an end in itself, benefit from and during the time lead they have with regard to religious involvement. The time lag disfavoring opportunistic defectors, who in fact notice of the instrumental qualities inherent in religious involvement when it comes to maximizing their wealth, follows naturally from the fact that imitation and mimicry refer to behaviors which already exist. This mechanism depends of course on the type of agents which is not admitted to the picture: Those who see religion as an end in itself and simultaneously defect in the Prisoners' Dilemma. Psychological regularities such as self-attribution and cognitive dissonance theories were invoked for a justification of leaving such 'split personalities' out.

The argument developed in Chapter 5 is of course highly stylized, and religion should only be understood as an illustration, as emphasized in section 5.5. What has been proposed is a group selection mechanism to accommodate the evolution of (within-group) altruism. Such arguments are called for if one does not wish to follow the example of evolutionary biologist John Haldane, whose "personal view of the world was that altruism *is* rare in humans and other animals, so he was content to drop the subject." (Sober & Wilson 1998, p. 65, their italics). This seems unnecessary. After all, the evidence to the contrary is immense. Fortunately.

# Bibliography


ALCHIAN, ARMEN. 1950. Uncertainty, Evolution, and Economic Theory. *Journal of Political Economy*, **58**(3), 211–222.

ALCHIAN, ARMEN, & DEMSETZ, HAROLD. 1972. Production, Information Costs, and Economic Organization. *American Economic Review*, **62**, 777–795.

ALEXANDER, RICHARD D. 1987. *The Biology of Moral Systems.* New York: Aldine de Gruyter.

ARROW, KENNETH. 1974. *The Limits Of Organization.* New York and London: W. W. Norton & Co.

ASHLOCK, DAN, SMUCKER, MARK D., STANLEY, E. ANN, & TESFATSION, LEIGH. 1996. Preferential Partner Selection in an Evolutionary Study of Prisoner's Dilemma. *BioSystems*, **37**, 99–125.

AUMANN, ROBERT J. 1959. Acceptable Points in General Cooperative q-Person Games. *Pages 287–324 of:* LUCE, R. D., & TUCKER, A. W. (eds), *Contributions to the Theory of Games IV.* Annals of Mathematics Studies, vol. 40. Princeton NJ: Princeton University Press.

AXELROD, ROBERT. 1984. *The Evolution of Cooperation.* New York: Basic Books.

AZZI, CORRY, & EHRENBERG, RONALD. 1975. Household Allocation of Time and Church Attendance. *Journal of Political Economy*, **83**(1), 27–56.

BANERJEE, ABHIJIT, & WEIBULL, JÖRGEN W. 1994. Evolutionary Selection and Rational Behavior. *Chap. 12, pages 343–363 of:* KIRMAN, ALAN, & SALMON, MARK (eds), *Rationality and Learning in Economics.* Oxford: Basil Blackwell.

BARRO, ROBERT, & MCCLEARY, RACHEL. 2003. *Religion and Economic Growth.* NBER Working Paper No. 9682.

BENDOR, JONATHAN, & SWISTAK, PIOTR. 1997. The Evolutionary Stability of Cooperation. *American Political Science Review*, **91**(2), 290–307.

BERG, JOYCE, DICKHAUT, JOHN, & MCCABE, KEVIN. 1995. Trust, Reciprocity, and Social History. *Games and Economic Behavior*, **10**, 122–142.

BERGSTROM, THEODORE C. 2002. Evolution of Social Behavior: Individual and Group Selection. *Journal of Economic Perspectives*, **16**, 67–88.

BERGSTROM, THEODORE C. 2003. The Algebra of Assortative Encounters and the Evolution of Cooperation. *International Game Theory Review*, **5**, 1–18.

BICKEL, PETER J., HAMMEL, EUGENE A., & O'CONNELL, J. W. 1975. Sex Bias in Graduate Admissions: Data From Berkeley. *Science*, **187**, 398–404.

BINMORE, KEN. 1994. *Game Theory and the Social Contract. Volume 1: Playing Fair.* Cambridge, Mass.: The MIT Press.

BOUDON, RAYMOND. 1987. The Individualistic Tradition in Sociology. *Chap. 1, pages 45–70 of:* ALEXANDER, JEFFREY C., GIESEN, BERNHARD, MÜNCH, RICHARD, & SMELSER, NEIL J. (eds), *The Micro-Macro Link.* Berkeley: University of California Press.

BOYER, PASCAL. 2001. *Religion Explained: The Evolutionary Origins of Religious Thought.* New York: Basic Books.

COHEN, DAN, & ESHEL, ILAN. 1976. On the Founder Effect and the Evolution of Altruistic Traits. *Theoretical Population Biology*, **10**, 276–302.

COOK, KAREN S., TOSHIO, YAMAGISHI, COYE, CHESHIRE, COOPER, ROBIN, MATSUDA, MASAFUMI, & MASHIMA, RIE. 2005. Trust Building via Risk Taking: A Cross-Societal Experiment. *Social Psychology Quarterly*, **68**, 121–142.

COOPER, BEN, & WALLACE, CHRIS. 2004. Group Selection and the Evolution of Altruism. *Oxford Economic Papers*, **56**, 307–330.

DARWIN, CHARLES. 1874. *The Descent of Man; and Selection in Relation to Sex*. 2nd edn. New York: Crowell.

DAWKINS, RICHARD. 1989. *The Selfish Gene*. 2nd edn. Oxford: Oxford University Press.

DECI, EDWARD L., & RYAN, RICHARD M. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior*. New York: Plenum Press.

EKMAN, PAUL. 2002. *Telling Lies: Clues to Deceit in the Marketplace, Marriage, and Politics*. 3rd edn. New York and London: W. W. Norton & Co.

FARRELL, JOSEPH, & WARE, ROGER. 1989. Evolutionary Stability in the Repeated Prisoner's Dilemma. *Theoretical Population Biology*, **36**, 161–166.

FIELD, ALEXANDER J. 2001. *Altruistically Inclined? The Behavioral Sciences, Evolutionary Theory, and the Origins of Reciprocity*. Ann Arbor: The University of Michigan Press.

FLETCHER, JEFFREY A., & ZWICK, MARTIN. 2000. Simpson's Paradox Can Emerge from the N-Player Prisoner's Dilemma: Implications for the Evolution of Altruistic Behavior. *In:* ALLEN, J. K., & WILBY, J. M. (eds), *Proceedings of The WorldCongress of the Systems Sciences and ISSS 2000*. Toronto, Canada: International Society for the Systems Sciences.

FRANK, ROBERT H. 1987. If Homo Economicus Could Choose His Own Utility Function, Would He Want One with a Conscience? *American Economic Review*, **77**, 593–604.

FRANK, ROBERT H. 1988. *Passions within Reason: The Strategic Role of the Emotions*. New York and London: W. W. Norton & Co.

FRANK, ROBERT H. 1994. Group Selection and "Genuine" Altruism. *Behavioral and Brain Sciences*, **17**, 620–621.

FRANK, ROBERT. H., GILOVICH, TOM, & REGAN, DENISE. 1993. The Evolution of One-Shot Cooperation - an Experiment. *Ethology and Sociobiology*, **14**(4), 247–256.

FRIEDMAN, JAMES W. 1971. A Non-Cooperative Equilibrium for Supergames. *Review of Economic Studies*, **38**(1), 1–12.

FRIEDMAN, MILTON. 1953. *Essays in Positive Economics*. Chicago: The University of Chicago Press.

FUTUYMA, DOUGLAS J. 1998. *Evolutionary Biology*. 3rd edn. Sunderland, Mass.: Sinauer Associates.

GALE, DAVID, & SHAPLEY, LLOYD S. 1962. College Admission and the Stability of Marriage. *American Mathematical Monthly*, **69**, 9–15.

GAUTHIER, DAVID. 1986. *Morals by Agreement*. Oxford: Oxford University Press.

GÜTH, WERNER, & KLIEMT, HARTMUT. 1998. The Indirect Evolutionary Approach: Bridging the Gap between Rationality and Adaptation. *Rationality and Society*, **10**(3), 377–399.

GÜTH, WERNER, & YAARI, MENAHEM. 1992. Explaining Reciprocal Behavior in Simple Strategic Games: An Evolutionary Approach. *Pages 23–34 of:* WITT, ULRICH (ed), *Explaining Process and Change - Approaches to Evolutionary Economics*. Ann Arbor: The University of Michigan Press.

GUTTMAN, JOEL M. 2000. On the Evolutionary Stability of Preferences for Reciprocity. *European Journal of Political Economy*, **16**, 31–50.

HAMILTON, WILLIAM D. 1963. The Evolution of Altruistic Behavior. *American Naturalist*, **97**, 354–356.

HAUK, ESTHER. 2001. Leaving the Prison: Permitting Partner Choice and Refusal in Prisoner's Dilemma Games. *Computational Economics*, **18**, 65–87.

HENRICH, JOSEPH. 2004. Cultural Group Selection, Coevolutionary Processes and Large-Scale Cooperation. *Journal of Economic Behavior & Organization*, **53**(1), 3–35.

HOLLAND, JOHN H. 1975. *Adaptation in Natural and Artificial Systems.* Ann Arbor: The University of Michigan Press.

HRUSCHKA, DANIEL J., & HENRICH, JOSEPH. 2006. Friendship, Cliquishness, and the Emergence of Cooperation. *Journal of Theoretical Biology*, **239**, 1–15.

IANNACCONE, LAURENCE R. 1998. Introduction to the Economics of Religion. *Journal of Economic Literature*, **36**, 1465–1496.

IRONS, WILLIAM. 2001. Religion as a Hard-to-Fake Sign of Commitment. *Chap. 13, pages 292–309 of:* NESSE, RANDOLPH M. (ed), *Evolution and the Capacity for Commitment.* New York: Russell Sage Foundation.

JABLONKA, EVA, & LAMB, MARION J. 1995. *Epigenetic Inheritance and Evolution.* Oxford: Oxford University Press.

JAMES, HARVEY S. JR. 2002. The Trust Paradox: A Survey of Economic Inquiries into the Nature of Trust and Trustworthiness. *Journal of Economic Behavior & Organization*, **47**, 291–307.

KREPS, DAVID. 1990. Corporate Culture and Economic Theory. *Pages 90–143 of:* ALT, JAMES, & SHEPSLE, KENNETH (eds), *Perspectives on Positive Political Economy.* Cambridge: Cambridge University Press.

LORBERBAUM, JEFFREY P., BOHNING, DARYL E., SHASTRI, ANANDA, & SINE, LAUREN E. 2002. Are There Really No Evolutionary Stable Strategies in the Iterated Prisoner's Dilemma? *Journal of Theoretical Biology*, **214**, 155–169.

MACY, MICHAEL W. 1998. Social Order in Artificial Worlds. *Journal of Artificial Societies and Social Simulation*, **1**(1), available online at `http://www.soc.surrey.ac.uk/JASSS/1/1/4.html`.

MACY, MICHAEL W., & SKVORETZ, JOHN. 1998. The Evolution of Trust and Cooperation between Strangers: A Computational Model. *American Sociological Review*, **63**, 638–660.

MARINOFF, LOUIS. 1990. The Inapplicability of Evolutionarily Stable Strategy to the Prisoner's Dilemma. *British Journal for the Philosophy of Science*, **41**, 461–472.

MAYNARD SMITH, JOHN. 1964. Group Selection and Kin Selection. *Nature*, **201**, 1145–1147.

MAYNARD SMITH, JOHN, & PRICE, GEORGE R. 1973. The Logic of Animal Conflict. *Nature*, **246**, 15–18.

NORTH, DOUGLASS. 1984. Government and the Cost of Exchange in History. *The Journal of Economic History*, **44**(2), 255–264.

NOWAK, MARTIN A., & SIGMUND, KARL. 1998. Evolution of Indirect Reciprocity by Image Scoring. *Nature*, **393**, 573–577.

OCKENFELS, PETER. 1993. Cooperation in Prisoners' Dilemma - An Evolutionary Approach. *European Journal of Political Economy*, **9**, 567–579.

ORBELL, JOHN, & DAWES, ROBYN M. 1991. A "Cognitive Miser" Theory of Cooperators' Advantage. *American Political Science Review*, **85**(2), 515–528.

RADNER, ROY. 1980. Collusive Behavior in Noncooperative Epsilon-Equilibria of Oligopolies with Long but Finite Lives. *Journal of Economic Theory*, **22**, 136–154.

REEVE, HUDSON KERNE. 2000. Multi-Level Selection and Human Cooperation. Review of *Unto Others: The Evolution and Psychology of Unselfish Behavior* by Elliott Sober and David Sloan Wilson. *Evolution and Human Behavior*, **21**, 65–72.

RICHERSON, PETER J., & BOYD, ROBERT. 2005. *Not by Genes Alone: How Culture Transformed Human Evolution.* Chicago and London: The University of Chicago Press.

SCHLICHT, EKKEHART. 1995. Economic Analysis and Organised Religion. *Chap. 3, pages 111–162 of:* JONES, ERIC, & REYNOLDS, VERNON (eds), *Survival and Religion: Biological Evolution and Cultural Change.* New York: John Wiley & Sons.

SCHLICHT, EKKEHART. 1998. *On Custom in the Economy.* Oxford: Clarendon Press.

SCHLICHT, EKKEHART. 2002. Reflections and Diffractions. Schlicht Replies to His Critics. *American Journal of Economics and Sociology*, **61**(2), 571–594.

SCHÜSSLER, RUDOLF. 1989. Exit Threats and Cooperation Under Anonymity. *Journal Of Conflict Resolution*, **33**(4), 728–749.

SETHI, RAJIV. 1996. Evolutionary Stability and Social Norms. *Journal of Economic Behavior and Organization*, **29**, 113–140.

SIMPSON, E. H. 1951. The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society*, **B13**, 238–241.

SKYRMS, BRIAN. 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

SOBER, ELLIOTT, & WILSON, DAVID SLOAN. 1998. *Unto Others. The Evolution and Psychology of Unselfish Behavior*. Cambridge, Mass: Harvard University Press.

SOSIS, RICHARD. 2003. Why Aren't We All Hutterites? Costly Signaling Theory and Religious Behavior. *Human Nature*, **14**(2), 91–127.

SOSIS, RICHARD, & ALCORTA, CANDACE. 2003. Signaling, Solidarity, and the Sacred: The Evolution of Religious Behavior. *Evolutionary Anthropology*, **12**, 264–274.

SPENCE, MICHAEL. 1973. Job Market Signaling. *Quarterly Journal of Economics*, **87**, 355–374.

TABELLINI, GUIDO. 2005. *Culture and Institutions: Economic Development in the Regions of Europe*. CESifo Working Paper No. 1492.

TAN, JONATHAN H. W., & VOGEL, CLAUDIA. 2005. *Religion and Trust: An Experimental Study*. European University Viadrina Frankfurt (Oder), Department of Business Administration and Economics Discussion Paper No. 240.

TESFATSION, LEIGH. forthcoming. Agent-Based Computational Economics. *In:* LUNA, FRANCESCO, PERRONE, ALLESSANDRO, & TERNA, PIETRO (eds), *Agent-Based Theories, Languages, and Practices*. London and New York: Routledge Publishers.

TRIVERS, ROBERT L. 1971. The Evolution of Reciprocal Altruism. *Quarterly Review of Biology*, **46**, 35–57.

TULLBERG, JAN. 2003. Rationality and Social Behavior. *Journal of Theoretical Biology*, **224**, 469–478.

WEBER, MAX. 1970. The Protestant Sects and the Spirit of Capitalism. *Pages 302–322 of:* GERTH, HANS H., & MILLS, C. WRIGHT (eds), *From Max Weber: Essays in Sociology*. London: Routledge & Kegan Paul.

WEIBULL, JÖRGEN W. 1995. *Evolutionary Game Theory*. Cambridge, Mass.: The MIT Press.

WILLIAMS, GEORGE C. 1966. *Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought*. Princeton: Princeton University Press.

WILLIAMS, GEORGE C., & WILLIAMS, D. C. 1957. Natural Selection of Individually Harmful Social Adaptations amongs Sibs with Special Reference to Social Insects. *Evolution*, **11**, 32–39.

WILLIAMSON, OLIVER E. 1993. Calculativeness, Trust, and Economic Organization. *Journal of Law and Economics*, **36**(1), 453–486.

WILSON, DAVID SLOAN. 1987. Altruism in Mendelian Populations Derived from Sibling Groups: The Haystack Model Revisited. *Evolution*, **41**, 1059–1070.

WILSON, DAVID SLOAN. 2002. *Darwin's Cathedral: Evolution, Religion, and the Nature of Society*. Chicago and London: The University of Chicago Press.

WRIGHT, SEWALL. 1945. Tempo and Mode in Evolution: A Critical Review. *Ecology*, **26**, 415–419.

WYNNE-EDWARDS, VERO C. 1962. *Animal Dispersion in Relation to Social Behaviour*. Edinburgh: Oliver & Boyd.

ZAHAVI, AMOTZ. 1975. Mate Selection - Selection for a Handicap. *Journal of Theoretical Biology*, **53**, 205–214.

ZYWICKI, TODD J. 2000. "Was Hayek Right About Group Selection After All?" Review Essay of *Unto Others: The Evolution and Psychology of Unselfish Behavior,* by Elliott Sober and David Sloan Wilson. *Review of Austrian Economics,* **13**, 81–95.

# Lebenslauf

**Persönliche Angaben**

| | |
|---|---|
| | Wolfgang Pfeuffer |
| | geboren am 23. Juli 1974 in Ansbach |

**Schule**

| | |
|---|---|
| 1980-1984 | Grundschule in Estenfeld |
| 1984-1993 | Städt. Schönborn-Gymnasium in Würzburg |
| | Abschluss: Allgemeine Hochschulreife |

**Zivildienst**

| | |
|---|---|
| 1993-94 | Paritätischer Wohlfahrtsverband |
| | Bezirksverband Unterfranken, Geschäftsstelle Würzburg |

**Studium**

| | |
|---|---|
| 1994-1995 | Julius-Maximilians-Universität Würzburg |
| | Rechtswissenschaft (ohne Abschluss) |
| 1995-2001 | Otto-Friedrich-Universität Bamberg |
| | Studiengang Europäische Wirtschaft |
| | Abschluss: Diplom-Volkswirt Univ. (Europa-Studiengang) |

**Promotion**

| | |
|---|---|
| seit 2002 | Ludwig-Maximilians-Universität München |
| | Promotionsstudium (bis 2006) an Volkswirtschaftlicher Fakultät |
| | sowie wissenschaftlicher Mitarbeiter |
| | am Seminar für Theorie und Politik der Einkommensverteilung |

München, 31. August 2006 Wolfgang Pfeuffer