DISS. ETH Nr. 19968

The evolution of fairness preferences, altruistic punishment, and cooperation

A dissertation submitted to
ETH ZURICH

for the degree of
DOCTOR OF SCIENCES

presented by
PHILIPP MORITZ HETZER
Dipl. Inform.-Wirt., Universität Karlsruhe
born 30. January 1980
citizen of Germany

accepted on the recommendation of
Prof. Dr. Didier Sornette, examiner
Dr. Charles Efferson, co-examiner

2011

Summary

The evolution of prosocial behavior and, in particular, of cooperation is still considered as one of the 25 major unsolved questions in science (Pennisi, 2005). Any prosocial behavior seems to contradict Darwin's principle of "the survival of the fittest" and the widely accepted assumption of a ubiquitous rational and selfish actor. Nevertheless, an enormous level of large-scale cooperation among humans and other forms of life can be observed.

As a consequence, researchers from various disciplines have started to investigate the puzzle of cooperation. Among these fields are evolutionary biology (Robinson, Fernald, and Clayton, 2008), neuroscience (Singer, Seymour, O'Doherty, Stephan, Dolan, and Frith, 2006; Donaldson and Young, 2008), anthropology (Henrich, 2004; Burkart, Hrdy, and Van Schaik, 2009), sociology (Coleman, 1998; Elster, 2007), and economics (Fehr and Gachter, 2000; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Camerer, 2003). This resulted in the development of theories and models of reciprocity (Cox, Friedman, and Gjerstad, 2007; Nowak, 2006), other-regarding behavior (Rabin, 1993; Fehr and Schmidt, 1999), and social coherence (Bernheim, 1994; de Hooge, Zeelenberg, and Breugelmans, 2007; Henrich, Mcelreath, Barr, Ensminger, Barrett, Bolyanatz, Cardenas, Gurven, Gwako, Henrich, Lesorogol, Marlowe, Tracer, and Ziker, 2006). In addition, laboratory experiments and field studies have been carried out to analyze the prosocial behavior of humans (Fehr and Gachter, 2000, 2002; Hamlin, Wynn, and Bloom, 2007), animals (Brosnan and de Waal, 2003; Silk, Brosnan, Vonk, Henrich, Povinelli, Richardson, Lambeth, Mascaro, and Schapiro, 2005; Jensen, Call, and Tomasello, 2007b,a; de Waal, Leimgruber, and Greenberg, 2008; Range, Horn, Viranyi, and Huber, 2008), and even insects (Nowak, Tarnita, and Wilson, 2010). In sum, this diverse body of literature suggests that our prosocial behavior is deeply rooted in our genetic and cultural heritage.

The co-evolution of culture and genes represents the fundamental assumption underlying this thesis. Applying methods from complex systems science combined with approaches from biology, evolutionary psychology, sociology and behavioral economics, we have developed two models that help to understand the emergence of fairness preferences, altruistic punishment and cooperation

in an evolutionary competitive and resource-limited world. In particular, we focus on the behavior of subjects in a public goods problem scenario which is considered to reflect many real life situations.

In the first part of this thesis, we develop an analytical framework that reflects the interactions of agents playing a public goods game with punishment under evolutionary dynamics. We compare the results with the empirical observations obtained in three previously conducted laboratory experiments. This leads to the following two results. First, the perception of unfairness in combination with the maximization of one's relative fitness explains quantitatively the observed altruistic punishment behavior among humans: the behavior of subjects in the experiments seems to be driven by an aversion against disadvantageous inequitable outcomes. Second, a disadvantageous inequity aversion preference is evolutionary dominant and stable in an evolutionary environment when compared to purely self-regarding behavior.

In the second part of this thesis, we complement our analytical model by numerical simulations. This allows us to relax the assumption of a homogeneous population that was required in the analytical model. We are able to verify that disadvantageous inequity aversion inevitably leads to the emergence of altruistic punishment in a heterogeneous population of multiple interacting agents. Furthermore, we show that an aversion against disadvantageous inequitable outcomes dominates essentially all other variations of other-regarding preferences in an evolutionary environment.

In the third part of the thesis, we focus on the effect that punishment has on the level of cooperation among agents who play a public goods game. We do this empirically with an analysis of the micro-level data from the three previously conducted experiments and by using our numerical simulation model. The empirical observations suggest that punishment acts as a coordination mechanism in one-shot interactions. Also, the simulation results show that punishment only sustains a preexisting level of cooperation but cannot explain its evolutionary emergence.

In the last part of this thesis, we first show that punishment can promote cooperation if the population of agents is sufficiently heterogeneous in the cooperation behavior. Then, we investigate different mechanisms of multi-level

selection and show that they are able to generate and to maintain heterogeneity among the agents even in the presence of punishment. The combination of the aversion against disadvantageous inequitable outcomes and the resulting altruistic punishment behavior together with the heterogeneity induced by multi-level selection processes ultimately explains the evolutionary emergence of cooperation.

Kurzfassung

Die Entstehung von prosozialem Verhalten und insbesondere von Kooperation wird immer noch als eine der 25 grossen unbeantworteten Fragen der Wissenschaft angesehen (Pennisi, 2005). Jegliche Art prosozialen Verhaltens steht im Widerspruch zu Darwins Grundsatz des "Überleben des Stärkeren" und der weithin verbreiteten Annahme eines immer rational und egoistisch handelnden Menschen. Nichtsdestotrotz kann ein weit verbreitetes und hohes Mass an kooperativem Verhalten zwischen Menschen und auch bei anderen Lebensformen beobachtet werden.

Als Konsequenz daraus haben Wissenschaftler vieler verschiedener Forschungsrichtungen angefangen, die Grundlagen der Kooperation zu ergründen. Darunter befindet sich unter anderem die Evolutionsbiologie (Robinson et al., 2008), Neurowissenschaften (Singer et al., 2006; Donaldson and Young, 2008), Anthropologie (Henrich, 2004; Burkart et al., 2009), Soziologie (Coleman, 1998; Elster, 2007) und Ökonomie (Fehr and Gachter, 2000; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Camerer, 2003). Dies führte zur Entwicklung von Theorien und Modellen der Reziprozität (Cox et al., 2007; Nowak, 2006), des Gruppen- und Umfeld bezogenen Verhaltens (Rabin, 1993; Fehr and Schmidt, 1999) und der sozialer Kohärenz (Bernheim, 1994; de Hooge et al., 2007; Henrich et al., 2006). Zusätzlich wurden Labor Experimente und Feldstudien durchgeführt, die das prosoziale Verhalten von Menschen (Fehr and Gachter, 2000, 2002; Hamlin et al., 2007), Tieren (Brosnan and de Waal, 2003; Silk et al., 2005; Jensen et al., 2007b,a; de Waal et al., 2008; Range et al., 2008) und auch Insekten (Nowak et al., 2010) untersuchen. Die zusammenfassende Betrachtung der Erkenntnisse oben genannter Disziplinen lässt darauf schliessen, dass unsere Neigung zu prosozialem Verhalten tief in unserem genetischen und kulturellen Erbe verwurzelt ist.

Die Koevolution von Kultur und Genen bildet eine grundlegende Annahme dieser Arbeit. Dabei entwickeln wir mit Hilfe von Methoden komplexer Systeme in Verbindung mit Denkansätzen aus der Biologie, der evolutionären Psychologie, der Soziologie und der Verhaltensökonomie zwei Modelle, die dabei helfen, die Entstehung von Fairness Präferenzen, altruistischer Bestrafung und Kooperation in einem evolutionären und kompetitiven Umfeld und unter einer

beschränkten Anzahl an Ressourcen zu erklären. Insbesondere betrachten wir das Verhalten von Subjekten im Rahmen eines Public Goods Problem Szenarios, welches eine Abstraktion vieler Situationen des alltäglichen Lebens darstellt.

Im ersten Teil dieser Arbeit entwerfen wir ein analytisches Modell, welches die Interaktionen und die evolutionäre Dynamik von Agenten abbildet, die ein Public Goods Spiel mit Bestrafungsmöglichkeit spielen. Wir vergleichen die Ergebnisse des Modells mit den empirischen Beobachtungen aus drei Laborexperimenten, die von Fehr, Gächter und Fudenberg, Pathak durchgeführt wurden (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009). Dies führt zu den folgenden zwei Resultaten: Unser Begriff von Fairness, respektive Unfairness, im Zusammenhang mit unserer Neigung, stets unsere relative Fitness gegenüber anderen Individuen zu maximieren, erklärt quantitativ exakt das beobachtete altruistische Bestrafungsverhalten; das Verhalten der Probanden in den Experimenten scheint eindeutig durch eine Abneigung gegenüber eines für sie selbst nachteiligen Spielergebnisses bestimmt zu sein. Zweitens können wir zeigen, dass diese Abneigung gegenüber eines nachteiligen Ausgangs eine evolutionär stabile und gegenüber einem rein egoistischen und selbstzentrierten Handeln dominante Strategie darstellt.

Im zweiter Teil dieser Arbeit komplementieren wir das zuvor eingeführte analytische Modell mit Hilfe von numerischen Simulationen. Diese Methode ermöglicht es uns, die Homogenitätsannahme des analytischen Modelles zu lockern. Mit Hilfe der Simulationen verifizieren wir, dass die unterbewusste Aversion gegenüber Situationen, die sich für einen selbst als nachteilig erweisen, auch in heterogenen Population zu einem Verhalten altruistischer Bestrafung führt. Desweiteren zeigen wir, dass die Abneigung gegenüber nachteiligen Situationen im wesentlichen alle anderen Varianten von Fairness Präferenzen innerhalb einer evolutionär kompetitiven Umgebung dominiert.

Im dritten Teil der Arbeit konzentrieren wir uns darauf, wie Bestrafung sich auf den Grad der Kooperation zwischen Agenten auswirkt, welche innerhalb eines Public Goods Spiel interagieren. Dazu führen wir eine detaillierte Analyse der zuvor von Fehr, Gächter und Fudenberg, Pathak empirisch beobachteten individuellen Verhaltensmuster durch. Die empirischen Beobachtungen unter-

stützen die Annahme, dass Bestrafung lediglich als Koordinationsmechanismus zwischen Probanden dient, welche nur einmalig miteinander interagieren. Unsere Simulationsergebnisse zeigen zusätzlich, dass Bestrafung nur ein zuvor bereits existierendes Mass an Kooperation erhalten kann, jedoch nicht eine Entstehung dessen erklären.

Im letzten Teil dieser Arbeit zeigen wir zunächst, dass Bestrafung die Entstehung von Kooperation begünstigen kann, wenn die Population der Agenten über die Zeit hinweg hinreichend heterogen ist. Anschliessend analysieren wir verschiedene Varianten von Multi-Level Selektion und zeigen, dass diese auch in der Gegenwart einer koordinierenden Bestrafung in der Lage sind, ein geeignetes Mass an Heterogenität in der Population zu erhalten. Wie kann die Entstehung von kooperativem Verhalten schlussendlich erklärt werden? Zum einen ist es unsere zutiefst innere Abneigung gegenüber nachteiligen und unfairen Situationen. Zum zweiten die daraus resultierende Neigung zu altruistischer Bestrafung und dessen koordinierende Funktion. Zum dritten ist es das Zusammenspiel der daraus resultierenden Koordinationswirkung und der durch die Gruppenselektion bedingten Heterogenität in der Population.

Acknowledgement

I would like to express my gratitude to my advisor Didier Sornette for his excellent academic coaching and the freedom and support he gave me to pursue my research in different directions and with full personal responsibility. Didier has always been available for discussions of new ideas and solving problems which was highly appreciated given it is a rare and invaluable characteristic for someone with such busy schedule.

Special thanks go to Dr. Charles Efferson for his co-supervision, remarks and comments that contributed substantially to the quality of this thesis and to the resulting papers.

Parts of this thesis have been supported by the Zürcher Kantonalbank. The received financial support is gratefully acknowledged. In particular, I want to thank Paolo Vanini.

Working at the Chair of Entrepreneurial Risks has always been a very enjoyable, creative and stimulating environment. Special thanks go to Heidi Demuth for her support in all administrative tasks and for being the reliable and constant social hub in the group. I want to extend my thank to Mirko Birbaumer, Peter Cauwels, Riley Crane, Gilles Daniel, Maroussia Favre, Vladimir Filimonov, Georges Harras, Andreas Hüsler, Thomas Maillart, Dirk Martignoni, Moritz Müller and Ryan Woodard for all the intense, stimulating and fruitful discussions and the great time together at ETH Zurich. I also want to thank the HPC-cluster team of ETH around Olivier Byrde for their outstanding support in technical issues.

I am grateful to the financial engineering department of the Zürcher Kantonalbank, especially to Bernhard Maeder, Sara Nogly, Carla Schneider, Constantin Schrafl, Nikola Snaidero, Silvan Spross, Juerg Syz, Roman Würsch and Ulf and Ernie for their great support and the unbeatable working atmosphere in this fantastic team.

Last but not least I want to thank my mother, Britta Hetzer, for her invaluable support and her guidance on all aspects of life and Alix Wiegand for her encouragement and exhaustless patience with me.

Moritz Hetzer
Zurich, August 2011

Contents

1. Introduction

Living organisms and in particular humans are characterized by their tendency to form social groups, to jointly work for common goods and to cooperate. The evolution of cooperation and prosocial behavior is considered to be one of the big question in science that still remains (partly) unanswered (Kennedy and Norman, 2005). This thesis contributes to a better understanding of the evolution of cooperation and prosocial behavior by taking on a perspective from various disciplines ranging from biology, evolutionary psychology and complex systems to sociology and economics.

1.1 Motivation, aim and scope

During the past centuries research in economics, biology, psychology, sociology and anthropology shared the common objective to analyze, model and ultimately explain the individual as well as the collective behavior of humans. In particular patterns of prosocial behavior have been analyzed and discussed using lab experiments, field studies and theoretical frameworks. At a first view and from an evolutionary perspective these behaviors seem to be puzzling and in contradiction with Charles Darwin's theory of natural selection and the principle of the "survival of the fittest". Moreover, evolution has all too often been interpreted as a goal-oriented process, a thinking that made its way into the different disciplines. Most prominently, economists and game theorists,

first established and (partly) still insists in the paradigm of an ubiquitous rationality that shapes and is shaped by pure self-regarding maximization objectives. Of course, the assumed selfish-optimization behavior is not misplaced in general, however, it often tells only half of the story. If we replace the word "fittest" in Darwin's theory by the concept of "best adapted" we kill two birds with one stone: First, this phrasing more precisely reflects the intended meaning of the initial statement; second, it highlights the relative character of the fitness measure and thus (i) puts the pure self-regarding maximization of self-interest into perspective and (ii) emphasizes the indetermination of the evolutionary process [1].

Behavioral sciences emerged and developed as largely disconnected disciplines in the scientific landscape and were and still are shaped by a history of well established and even better advocated disciplinary boundaries. This is considered to be one of the vital errors in social sciences (Capra, 2004). We believe that thinking in strictly-separated categories and disciplines is not adequate to capture the complexity and interdependencies that are required to understand and explain pro-sociality and its evolutionary emergence. As a consequence, this thesis provides a complex systems approach to explain the evolution of cooperative behavior that integrates perspectives from biology, psychological anthropology, sociology and economics. In this way, we can show how the interplay of mechanism at different scales, e.g. genetic or cultural, can promote the emergence of prosocial behavior in the form of cooperation and its supporting mechanisms such as altruistic punishment of non-cooperators.

Within our framework we define cooperative behavior as follows: cooperation is a joint action of multiple individuals to achieve a common purpose or mutual benefit. The second term that is of relevance for our work is "altruism". The French philosopher August Comte first brought up the definition of "vivre pour altrui", that is "live for others", which later on led to the term altruism. In this thesis we focus on the act of "altruistic punishment" that is defined as the punishment of defectors (non-cooperators) at own costs and without material benefit. Depending on the context, cooperation and altruism are closely tied. The distinguishing factor between a cooperative and an altruistic act is rooted

[1] In section 2.2 we will provide a more detailed discussion about the relative character of fitness in evolutionary systems.

in the definition of the underlying evolutionary fitness measure. In terms of relative fitness measure, cooperation is equivalent to altruism. Here, the cooperative behavior can reduce the fitness of the cooperator relative to the fitness of the recipient(s) in an otherwise competitive environment and thus can be considered to be altruistic. However, by the strict definition of altruism, an altruistic act requires to reduce the absolute fitness of the altruist and to increase the absolute fitness of the recipient(s) (Wilson, 1977; McElreath and Boyd, 2007). In our competitive evolutionary environment, the fitness of an agent is always defined relative to her context, i.e. to other agents. We are aware of the fact that within a relative definition of fitness cooperation can be promoted by a much wider range of mechanisms that are subject to less strict requirements. However, this assumption does not constitute a strict relaxation of the problem of cooperation to our approach: We will show that the emergence of cooperation requires the existence of altruism in the strict sense, namely in the form of altruistic punishment.

In our approach, we focus on the cooperation in voluntary contribution mechanisms, also known as the common- or public goods problem. In 1968, Hardin formulated this type of a social dilemma in his paper "The Tragedy of the Commons" and highlighted its relevance for the history and future of mankind (Hardin, 1968). The social dilemma in public goods has become prominent in the context of climate change and the depletion of natural resources by the human species. Throughout our analysis, we focus on the classical public goods problem that has been analyzed and discussed in many studies, most prominently in the context of economics and game theory. Most studies however share the limitations that come along with studying the problem only from the perspective of one discipline. A non-exhaustive list of exceptions is the work of Herbert Gintis, Joseph Henrich, Charles Efferson and others.

In the last two centuries, research has been shaped largely by a prototype of an exact and quantifiable science with Physics leading the way as opposed to the soft "social" sciences. This has led to the tendency and the perception that everything, even in social sciences, should be modeled analogously to the Newtonian mechanic in Physics, which shaped the thinking in many fields of science, most prominently in economics (Capra, 2004). However, the dynamics in social systems, such as the interaction in economies, are fundamentally

different to those applied in Physics which mostly base on constant and exactly defined natural phenomenons (except for extensions of the newtonian model in the field of quantum physics). In contrast to the Newtonian mechanical dynamics in physics, biology provides a framework of evolutionary dynamics that bases on non-deterministic processes which are characterized by random events in the form of mutations and the recurring recombination of a finite pool of genes. However, when looking at socio-economic systems, as we do in this thesis, it becomes evident that the evolution of these systems occurs on average in much shorter periods than in the biological context. Moreover, the evolution of an economic system inextricably reciprocates with the evolution of its underlying society. In turn, a society co-evolves along with a system of social conventions and values such as norms, culture and religion (Capra, 2004). It becomes obvious that the evolution of socio-economic systems can neither be characterized by unique objectives nor by clearly defined ideals and moral concepts. As a response to the constantly changing environment, either endogenously by social events or exogenously e.g. by (natural) disasters, the process of cultural evolution constantly develops new structures within our value system. This highlights the demand for a more integrative and transdisciplinary approach, something that we try to provide with this thesis.

In the remaining part of this chapter we discuss cooperation and pro-sociality viewed from different disciplines and at different scales. In particular, we look into the biological aspects of prosocial behavior and take on a perspectives of evolutionary psychology, sociology, economics and philosophy.

1.2 Biology and neuroscience

Biology deals with the characteristics of living organisms including their formation, growth and evolution. All biological organisms are subject to evolutionary dynamics in form of selection, replication and mutation. Most of them are also subject to adaptation dynamics, i.e. organism react and adapt to their environment during lifetime, and to cross-over replication in form of reproduction by mating of two or even more organisms. These fundamental forces apply to all living organisms and continually modify, shape and let them evolve across time.

From a biological perspective selection, cross-over replication and mutation occurs on the level of the genotype, i.e. by the continuous evolution of the organism's blueprint. The genotype consists of DNA molecules. These molecules encode and store genetic instructions. In the process of gene transcription the genetically stored information of the DNA is interpreted and transformed into RNA and protein molecules. This process of *gene expression* defines the starting point of the transition from the genotype to the observable characteristics of an organisms, the phenotype. The phenotype characterizes an organism by its biochemical properties, the physiological appearance and development and ultimately its behavior.

In general, evolutionary replication occurs either by plain copying of DNA structures or by cross-over replication through the combination of the DNA structure of two different organisms, e.g. by sexual reproduction. Besides the genetic variation in the process of cross-over replication, genetic diversity is induced and maintained in form of mutations that originate e.g. from DNA copying-errors. Parts of the DNA may also mutate as a consequence of environmental influences, e.g. by radiation damages. Thus, the phenotype of an organism is determined by its genotype and by the reciprocal interaction of the organism with its environment across time. Interaction occurs e.g. by adaptation to specific conditions while at the same time the environment is altered, e.g. by specific actions. The influence and interaction of the environment on the organism controls for the selection for well- vs. badly adapted organisms. This selection mechanism determines the genetic diversity of a population of organisms.

Research in ethology and neuroscience shows that social information patterns, e.g. in form of hierarchy structures among animals and social stimuli such as communication patterns in birds are shown to have an impact on the brain circuits (Mello, Vicario, and Clayton, 1992). Genes do not directly determine the behavior of individuals but encode the structure of molecules and therefrom direct the formation and development of neural circuits in the brain. The composition of these brain structures finally provides the substructure of the individually expressed behaviors.

Evidence for short-term updates of gene expressions in the brain structure have been verified across species (Donaldson and Young, 2008; Mello et al., 1992; Jarvis, Scharff, Grossman, Ramos, and Nottebohm, 1998; Clayton, 2000; Goodson, Evans, and Wang, 2006; Cummings, Larkins-Ford, Reilly, Wong, Ramsey, and Hofmann, 2008). Here, shifts in the neurogenomic state of the brain that base on the encoded information in the genotype are triggered by social stimuli. In particular, social information and social cognition have a non-negligible effect on brain structures. E.g. songbirds update gene expressions in the brain when exposed to unknown vocal sequences in order to adapt to a changing social environments and to detect potential unknown invaders. This happens even within the short time scale of hours. Worker bees change their behavior from brood care to pollen collectors which is triggered by an alternation of genes expressions in their brains. This alternation of the DNA, in turn, is controlled by the lack of specific repressive pheromones, i.e. RNA, that correlate inversely with the population's need for additional foraging (Grozinger, Sharabash, Whitfield, and Robinson, 2003). Another example of the influence of social information on behavior and genome modifications is the social hierarchy in groups that controls the access of individuals to common resources. The degree up to which resources can be accessed and utilized determines the fitness and the reproduction rate Whitfield, Cziko, and Robinson (2003); Grosenick, Clement, and Fernald (2007) and thus controls the propagation of specific genes.

Vice versa, an influence of genes and proteins on the social behavior has also been manifested. For instance, specific neural circuits of the brain can be associated with (social) behavior and thus the genotype might include a behavioral component (Fehr and Camerer, 2007; de Quervain, Fischbacher, Treyer, Schellhammer, Schnyder, Buck, and Fehr, 2004; Glimcher and Rustichini, 2004; Soares, Bshary, Fusani, Goymann, Hau, Hirschenhauser, and Oliveira, 2010). For example neural messengers account for the social cognition among members of the same species. Similarly, the reproductive behavior is influenced by peptises and their encoded gene expressions (Dickson, 2008).

A good overview of the interrelation of genes, brain structures and social behavior is presented in (Robinson et al., 2008).

1.3 Evolutionary Psychology

Evolutionary psychology is an interdisciplinary field of research that interlinks the biological aspects of human evolution with the psychological mechanisms underlying human behavior. Evolutionary psychology assumes that the behavior of humans is influenced by inherited factors both on a biological level as well as on a cultural level. This deduces from the fact that humans are nothing else than animals and are subject to the same evolutionary processes, i.e. humans are the product of nature, biology, nurture and culture. In particular, the human brain is made up of neural circuits that have been shaped by natural selection and thus every behavioral traits results from an adaptive reason. This leads to the perspective of a "modular mind" which describes our brain in terms of specialized (cognitive) modules, e.g. the module for language or the ability to recognize faces, which have evolved along adaptive problems within our evolutionary history. Even though we diverged from our most recent common primate ancestors 65 million years ago, the last 10.000 years most probably provided a most formative environment for the evolutionary adaptiveness of our minds. During this period humans started to group in hunter-gatherer societies, a development that still shapes our modern psychology and social behavior. In particular, we developed methods that allowed us to externalize knowledge and to teach this knowledge to our offsprings. In combination with an increasing population density and the associated higher rate of interactions, this gave rise to the emergence of a cumulative culture. The evolution of culture distinguishes us from other animals, although primates such as chimps and orang-utans developed relatively large brains that also enabled them to evolve initial indications and features of a culture. However, as the brain size correlates negatively with the reproduction rate of a species due to energetic constraints, hominids soon reached an evolutionary barrier. The human species successfully side-lined this barrier by the development of cooperative breeding structures and the organization in social group sizes and structure that supported the evolution of larger brains (Dunbar, 1998; Silk, Alberts, and Altmann, 2003; Zhou, Sornette, Hill, and Dunbar, 2005; Burkart et al., 2009; Burkart and van Schaik, 2010). Today, evolution occurs mainly by means of cultural adaptation and less on the level of biology. Cultural evolution has started to invent technologies and to establish social

institutions that enabled us to sideline the biological aspects of evolution and to establish certain behavioral patterns that are associated with pro-sociality. However, natural selection was replaced by other mechanisms and fitness measures that now determined the evolutionary process on the cultural level. The most prominent representative of this development is the advent of money.

1.4 Social Psychology

The interdisciplinary field of social psychology bridges the gap between the fields of psychology and sociology; while psychology focuses on the situational aspects of feelings, thoughts and decision making and the resulting behavior of individuals, sociology provides explanations for the collective decision making and behavior of groups and societies and how institutions and cultures form. In contrast, social psychology studies the effects of the interplay among humans on the feelings, decision making and the behavior of the single individual, either in the form of direct social interactions or indirectly by imagined or implied social influences (Allport, 1985); in other words: social psychology studies the human behavior in a social context, i.e. it focus on the single individual within the group. The field of social psychology mainly addresses questions in the following three domains:

- **social cognition**, i.e. how we perceive and interpret social objects and actions,
- **social influence**, i.e. how our behavior and attitudes are influenced and caused by others and
- **social interaction**, i.e. how interaction takes place in a social environment.

Laboratory and field experiments provide an important instrument in social psychology in which the effect of one or multiple altered variables that determine a specific situation are tested against other fixed variables. For example the violation of social vs. moral norms and its implication on the feeling of the observer (contempt vs. anger) and the violator (shame vs. guilt) was analyzed in experiments (de Hooge et al., 2007). Other areas such as the

differences between rational choices based on beliefs and desires compared to emotional/heuristic decisions (Gigerenzer and Selten, 2002; Plous, 1993), the effect of money in the social contexts (Vohs, Mead, and Goode, 2006) and the interplay of emotions and fairness norms (Reuben and van Winden, 2005) were investigated by tools of social psychology. In summary, social psychology is a very large, diverse and dynamic field of research on prosociality that goes far beyond the short overview in this thesis.

1.5 Sociology

Sociological theory focuses on the collective behavior of a society that results from the socialization of individuals by the internalization of social norms. Therefore a role/actor model has been introduced (Goffman, 1959). The process of socialization itself is characterized by the internalization of norms that occurs as a transmission of norms and moral values between successive generations (Parsons, 1967). The concept of norms has been widely discussed in the sociological theory resulting in various different and even sometimes inconsistent definitions. Essentially, the following three definitions of norms can be distinguished (Elster, 2007):

- **Social norms**, which affect the behavior of the actor as a result of the fact that this behavior can or will be observed by other individuals.
- **Moral norms**, which in general are unconditional and have a proactive character. This means that the behavior of the actors is not affected as a results of the presence of other individuals or as a reaction to them.
- **Quasi-moral norms**, which describe a reactive behavior that is triggered by the fact that the actor can observe the others' behavior.

In general, roles are characterized by one or multiple types of the above described norms. This normative commitment of roles imposes an intrinsic expectation on the actor's behavior in the form of moral virtues and ethical values (Gintis, 2009). Violations of the normative commitments associated with a role are expressed as emotions such as anger, guilt or shame. These universal emotions sustain the normative commitments, i.e. the continuity of

social norms. One example is the perception of fairness and the reaction to unfair behavior.

The set of internalized norms defines, characterizes and keeps a society together by attaching expectations, duties and obligations to the specific roles in the society. The roles may vary geographically as a result of different evolutionary trajectories. Beside the normative commitments of norms that are associated with a role, motivations of material interest also play an important aspect in sociology. The private payoff of an actor who holds a specific role in society might be in conflict with the public expectation or even with the public payoffs coming along with this role. Thus the wrong personal commitment of an actor to a role can lead to socially inefficient outcomes (Gintis, 2009). These kind of situations have been the objects of study in other disciplines, e.g. in experimental economics by analyzing social dilemmas such as voluntary contribution mechanisms and public goods problem. The coordinative features of social norms can help to overcome this conflict of interests and can help to promote prosocial behavior (Gintis, 2009).

1.6 Economics and Game Theory

Microeconomic theory focuses on the decision making of individuals which are thought of to strictly pursue only their private interests. A large body of the economic theory bases on the rational actor model and assumes that agents always seek to maximize their utility, i.e. to optimally achieve their desired preferences. Preference can be represented by a set of discrete choices or in the form of continuous preference relations. Various extensions to the standard utility framework have been formulated. Most known among them is the expected utility theory framework that adds the possibility for stochastic outcomes in decision settings. The definition of the preference or utility function plays a crucial role in the modeling of agents' behavior, in particular, when it comes to prosocial behavior. Three main classes of pro-social economic models have been identified in (Meier, 2006):

- **Outcome-based pro-social preferences:** This type of models account for other-regarding preferences in the utility function. Characteristic examples of outcome-based prosocial preferences are inequity or

inequality-aversion models, e.g. those presented in (Fehr and Schmidt, 1999; Rabin, 1993; Bolton and Ockenfels, 2000).

- **Reciprocity models:** Reciprocity models consider a time-dimension in the social interaction between agents. This induces a conditionality into the decision making of agents, keeping it with the motto "I scratch your back if you scratch/scratched mine". One example for prosocial reciprocity is the costly punishment of unfair behavior. Therefore agents need to have a common sense of what is perceived as being "unfair", which directly leads back to the concepts of norms, moral and culture. E.g. Falk and Fischbacher introduced a theory of reciprocity in (Falk and Fischbacher, 2006). Different forms of reciprocal rules have been defined, among them kin selection, direct and indirect reciprocity, network reciprocity and types of group selection (Nowak, 2006).

- **self-identity models:** Self-identity models are a mixture of the previous two types of models: They include other-regarding preferences, namely the second-order perception of the self-identity: Agents care for the advantageous perception of their reputation in the eyes of others. In turn, maintaining a "reputation" requires to conform to the social norms of the associated reference group. Thus, self-identity models indirectly include a reciprocal aspect, as norms do not emerge out of thin air, but moreover evolve from reciprocal interactions.

An increasing number of experiments and field studies are conducted with the aim to better understand the human decision process with respect to prosocial behavior and to verify the assumption made in the theoretical models. Evidence for prosocial preferences has been manifested across various scales in lab experiments and field studies. The design of most lab experiments introduces either a distinct competition- or coordination-problem, such as the competition in voluntary contribution-, market- and bargaining-games or the coordination problem in the various versions and modifications of the prisoner dilemma. In particular, pro-social behavior in the form of altruism and cooperation is the subject of numerous empirical investigations. Following our definition in section 1.1, cooperation is a joint action of multiple individuals to pursue a common purpose or mutual benefit. This definition provides the

basis for the social dilemma known as the public goods problem. This has been the object of study in many empirical and theoretical economic papers. In the public goods problem, a group of individuals can invest an effort, e.g. money, into a common public good which yields an amount back to the group that is larger than the sum of the individual contributions. However, the return generated by the public good is defined in a way that the per capita gain for an agent can become negative if only a small fraction of agents contributes to the public goods while others free-ride. Thus, the public goods problem defines a social dilemma situation that is comparable to a n-person prisoners dilemma and thus is susceptible to material self-interest. A wide range of decision settings in real life can be characterized as public goods problems (Meier, 2006) ranging from group work in seminars up to credit crunch on the interbanking market or the tragedy of the commons (Hardin, 1968). For that reason the public goods problem will play a central role in this thesis.

In addition to the perspective of utility theory, game theory provides a toolbox for studying strategic interactions among individuals. It provides a mathematical coherent framework to solve the strategic decision making problem that individuals are facing when the outcome is characterized by simultaneous or dependent decisions between actors. Game theory is considered as a logical extension of evolutionary theory as strategic behavior is often equated with being evolutionary stable (Gintis, 2009). However, short term adaptation and learning might be fundamentally different to the resulting dynamics and outcome of long-term co-evolutionary processes. Evidence for this will be shown later on in this thesis.

1.7 Philosophical perspective

Explaining prosocial behavior, such as cooperation and altruism, inevitably requires to discuss the definition of moral behavior. There is no truth about what is moral and what is good. Truth is always a question of scientific findings, whereas morality is a question of experience (Precht, 2010). Morality is always defined on basis of subjective criterions that vary among cultures, religions, locations and people; to keep it with Albert Einstein: “Morality is not God-given, but moreover a fully human affair”. It is the result of group communication

based on a joint background (Precht, 2010). The human brain developed the ability to formulate and think about very abstract and high-level concepts of morality. However, many decisions in everyday life are driven by heuristics and emotions that originate from different brain functions and regions than the abstract reasoning about the construct of morality. This ultimately prevents the high-order principles of morality to be applied in many decision settings (Gigerenzer, 2010; Precht, 2010). Already the Scottish philosopher David Hume differentiated between reasons and passions: Rational reasons alone cannot cause actions, but the subliminal emotions, i.e. the "slave of the passions" can do so. Thus, the pure rational part of decision making alone cannot explain moral behavior as this relies on social intuition. Social intuition is inevitably associated with emotions and instinctual feelings such as anger, fear, love, respect, shame and many more (Precht, 2010). Coming back to the "tragedy of the commons" (Hardin, 1968) and its inherent public goods problem, the question of "how cooperation emerges" receives an even more important aspect because the emotional basis accounting for our prosocial behavior is limited to our proximate social environment. To (partly) overcome this problem mechanisms of cultural evolution, such as the emergence of institutions and social norms, started to play a vital role among humans.

While neuroscience and biological processes can explain the way we perceived and experience specific situations on the basis of biochemical processes, other forms of consciousness can develop and occur only on higher levels of abstraction. Human beings are most probably the only species that is able to formulate assertions about their self, i.e. we can think of what we are and how we should be (Precht, 2010). This results in the existence of different levels of self-consciousness. For example, the interaction with other individuals and, in particular, the forming of beliefs about the others' perception about oneself (second, third,... -order beliefs) opens a different dimension to the concept consciousness and the perception of the self. We organize and compare everything that we experience not only with our interests but also with the perception of the self. Thus we rely on the attention and recognition of others (Precht, 2010). This leads to one of the essential aspects analyzed in this thesis: other-regarding preferences, -beliefs and -behavior which set the own-decision making and acting in relation to the beliefs and behaviors of oth-

ers. Making decisions by taking the behavior of other into account essentially creates a "social chess game" (Precht, 2010), which requires certain cognitive abilities such as e.g. language, memory or computational skills. These neural abilities are subject to evolutionary processes on the biological scale and consequently the "social chess game" is indispensably interlinked with our biological heritage. In chapter 2 and 3 we show that other-regarding preferences and the perception of unfairness are not necessarily an abstract cultural agreement between humans but moreover are deeply rooted in our evolutionary history. In the end our nature is characterized by a set of conflicting and diverse mechanisms of decision making that evolve and operate on different scales and levels of self-consciousness.

1.8 Structure of the thesis

The thesis is structured as follows: In chapter 2 we present an analytical framework to explain the emergence of fairness preferences and altruistic punishment behavior among agents who interact through a public goods problem. The framework combines ideas from expected utility theory and from mechanisms of evolutionary dynamics. Simplifying assumptions are made about the population structure to ensure the mathematical solvability of the problem. Chapter 3 addresses the same questions of fairness preferences and altruistic punishment as in chapter 2, however, we present a numerical approach in order to be able to account for heterogeneity in the population structure. Chapter 4 in detail investigates the level of cooperation among subjects who interact by means of a public goods game. Furthermore, it reveals the effect of altruistic punishment that emerges as a result of the evolutionary dominant fairness preferences, on the level of cooperation in a public goods game. In the last chapter 5 we analyze which evolutionary mechanisms are required to promote the emergence of cooperation in a competitive and resource limited environment that is susceptible to material self-interest. In particular, we show how the interplay of fairness, altruistic punishment and multi-level selection leads to high levels of cooperation. In this way, we try to provide a comprehensive and integrated picture to help understanding the evolutionary puzzle of cooperation. In the conclusion section 6 we finally provide a critical review of our work and give an outlook of potentially interesting extensions of our research.

2. A theory of evolution, fairness and altruistic punishment

In this chapter we identify and explain the mechanisms that account for the emergence of fairness preferences and altruistic punishment in voluntary contribution mechanisms using an analytical framework. In particular, we combine an evolutionary perspective together with an expected utility model. In order to cope with the complexity of the evolutionary dynamics and the n-player characteristics of the analyzed public goods game, we use common methods and assumption that are often applied in game theoretical frameworks to achieve a best trade off between computational tractability and representative results. Our approach is motivated by previous findings on other-regarding behavior, the co-evolution of culture, genes and social norms, as well as bounded rationality. Our first result reveals the emergence of two distinct evolutionary regimes that force agents to converge either to into a defection state or to a state of coordination, depending on the predominant set of self- or other-regarding preferences. Our second result indicates that subjects in public goods experiments coordinate and punish defectors as a result of an aversion against disadvantageous inequitable outcomes. Our third finding identifies disadvantageous inequity aversion as an evolutionary dominant and stable strategy in a heterogeneous population of agents that initially consists only of self-regarding and selfish-acting agents. We validate our model using

previously obtained results from three independently conducted experiments of public goods games with punishment.

2.1 Introduction

Why do we maintain moral attitudes, display other-regarding behavior, have a distaste for unfairness, act prosocially and, at times, even behave altruistically towards others? How is this behavior compatible with the predominant theories of rational choice, selfish utility maximization and, in particular, with Darwin's principle of the survival of the fittest? This chapter presents an evolutionary utility framework of fairness, altruistic punishment and cooperation. It presents quantitative arguments supporting the hypothesis that the key to understanding the ostensibly mysterious patterns of human behavior is deeply rooted in our evolutionary history.

Prosocial behavior in humans has been studied in many laboratory experiments throughout the world. One key finding is the evidence for altruistic punishment behavior in humans, i.e. the punishment of non-cooperators and norm violators at own costs without direct or indirect material benefit (Bochet, Page, and Putterman, 2006; Nikiforakis and Normann, 2008; Nikiforakis, 2010; Anderson and Putterman, 2006; Brandts and Fernanda Rivas, 2009; Fehr and Gachter, 2002; Fudenberg and Pathak, 2009; Gachter, Renner, and Sefton, 2008; Egas and Riedl, 2008; Masclet, Noussair, Tucker, and Villeval, 2003). To allow for this pro-social behavior that is often marked as "irrational", economists shifted from purely self-regarding assumptions to theories that incorporated other-regarding preferences (Camerer, 2003). In particular, analytical frameworks of fairness, reciprocity and cooperation have been formulated that consolidate individual utility maximization with inequality and inequity aversion (Rabin, 1993; Cox, Friedman, and Sadiraj, 2008; Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999; Falk and Fischbacher, 2006; Englmaier and Wambach, 2010; Andreoni and Miller, 2002). In this way, results from experimental economics have been rationalized and aligned with the predominant rational choice theory of pure self-interest.

Besides these equilibrium-based and time-independent utility theories, a second class of models emerged that focuses on the evolutionary origin of al-

truistic punishment and cooperation (Axelrod and Hamilton, 1981; Bowles, 1998; Imhof, Fudenberg, and Nowak, 2005; Sigmund, De Silva, Traulsen, and Hauert, 2010; Jensen, 2010; Gaechter, Herrmann, and Thoeni, 2010; Berger, 2010). These models are often motivated from a biological perspective including arguments from evolutionary psychology, anthropology and sociology. Although the emergence of pro-social behavior in settings which are subject to material self-interest seems to contradict rational choice theory and the principle of the survival of the fittest, one can show that altruistic punishment and other-regarding behavior can originate, emerge and be sustained in a competitive, resource-limited environment even in the presence of evolutionary dynamics.

This chapter presents a combination of both approaches: an expected utility framework that allows for other-regarding preferences, and which is subject to standard evolutionary dynamics. In particular, we show that the interplay of natural selection and selfish utility maximization inevitably results in the emergence of other-regarding preferences in the form of disadvantageous inequity aversion. The term "disadvantageous" implies a relaxation from the concept of inequity aversion and fairness preferences: Subjects only dislike situations in which the inequity is to their disadvantage. Consequently, no a priori stipulated modeling assumptions about altruistic, self-discriminating behavior are embodied. The aversion against inequitable outcomes causes altruistic punishment behavior to emerge, even in social dilemma situations that are subject to material self-interest. We argue that the bare individual survival needs of our ancestors induced an inherent predisposition to unfairness aversion that persists in our behavior up to this day.

This argument might sound farfetched given that human beings are probably the most successful species in eluding or manipulating natural selection by continuous enhancing, e.g., via improvements of health care and medical engineering. However, at the same time, our cultural evolution developed higher, more abstract levels of selection mechanisms that operate e.g. as monetary, bargaining and market competition, and led to hierarchical structures of power and of social standing. In other words, the natural selection that was previously affecting and operating on our hunter-gatherer ancestors has substantially been replaced in our modern societies by social institutions, most

notably by the advent of money and the measures of economic power. Our primal instinct to unfairness aversion is still subliminally active and can be triggered by this high-order social and cultural selection mechanisms. In consequence, the corresponding reactions to unfair behavior can be observed today even though we are in most situations not directly affected in our biological viability.

The analysis of our expected utility model, in combination with the underlying evolutionary dynamics, allows us to identify and explain the origin and the emergence of other-regarding preferences and, ultimately, enables us to quantitatively explain the degree of altruistic punishment that is observed in lab experiments. As a result, our approach complements and extends other utility frameworks, e.g. the Fehr/Schmidt model (Fehr and Schmidt, 1999), Bolton/Ockenfels (Bolton and Ockenfels, 2000) and Rabin (Rabin, 1993) by adding the too often neglected but, in fact, indispensable evolutionary perspective to the problem of explaining prosocial behavior. Unlike other approaches, our model does not assume ex ante the existence of other-regarding preferences, but instead demonstrates their co-evolutionary emergence along with the emergence of altruistic punishment behavior. The design of our model is inspired by previous findings about the co-evolution of culture, norms and genes, the effect of other-regarding behavior as well as bounded rationality. We motivate our model by the psychological predisposition of individuals to maximize their expected utility together with subliminal disposition to follow social norms (Gintis, 2009; Bernheim, 1994; Messick, 1999; Bardsley and Sausgruber, 2005; Henrich, 2004). Both mechanisms are closely related in the process of gene-culture co-evolution.

The following section 2.2 describes the model in detail and explains the interplay of agents that maximize their expected utility under the effects of natural selection and competitive evolutionary dynamics. Then, section 2.3 presents empirical tests of the theory. Section 2.4 establishes the evolutionary dominance of the specific other-regarding preference in the form of disadvantageous inequity aversion. Section 2.5 concludes.

2.2 The model

2.2.1 General framework

We take an evolutionary utility maximization approach as a starting point to construct our model. The fitness of an agent is considered to be equivalent to her realized cumulative payoff, i.e. to the monetary units (MU) that the agent gains over time. Each agent i is characterized by one or multiple traits. The traits of an agent determine her behavior and correspond to a pure strategy denoted by s_i. Traits are passed on as fitness weighted values to the offspring in the process of evolutionary reproduction. The population thus is determined by the set of pure strategies $S \subset \mathbb{R}^x$. In an evolutionary competitive environment, agents are subject to natural selection which affects their viability and fertility. While viability selection accounts for removing poor performing agents from the population, fertility selection enables more successful agents to spread and to promote their genetic and cultural heritage in the population. This process corresponds to the standard evolutionary challenge of survival and reproduction. Following the Darwinian principle of the survival of the fittest, both selection mechanisms are defined relative to the environment of an agent. This means that the fitness of an agent is determined relative to the performance of the remaining population that she is exposed to and interacts with. In an evolutionary environment, the success of an agent and of its strategies defines the fitness of the agent and thus determines the proportional change of the strategies (traits) in the population over time.

The set of strategies S that characterizes a population of agents is specified by a probability measure P^t that quantifies the frequencies of the single strategies $s_i \in S$ in the population at time t. In the two player case the payoff of an agent who plays a pure strategy $s \in S$ against another agent who plays the pure strategy $\hat{s}$ is denoted by $f(s, \hat{s})$. Both, s and $\hat{s}$ are defined in the x-dimensional continuous strategy space $S \subset \mathbb{R}^x$. For the n-player case, the average payoff of an agent who plays a strategy s at time t against a population characterized by the probability measure P^t over the strategy space S is defined by

$$E(s, P^t) = \int_S f(s, \hat{s}) P^t(d\hat{s}) \ . \tag{2.1}$$

The total average payoff of the entire population at time t is defined by

$$E(P^t, P^t) = \int_S \cdots \int_S f(s, \hat{s}) P^t(d\hat{s}) P^t(ds) \ . \tag{2.2}$$

The success of a strategy s is given by the difference of equations (2.1) and (2.2) as shown e.g. in (Oechssler and Riedel, 2001; Cressman and Hofbauer, 2005; Hofbauer, Oechssler, and Riedel, 2009):

$$\begin{aligned} \Phi(s, S) &= E(s, P^t) - E(P^t, P^t) \\ &= \int_S f(s, \hat{s}) P^t(d\hat{s}) - \int_S \cdots \int_S f(s, \hat{s}) P^t(d\hat{s}) P^t(ds) \end{aligned} \tag{2.3}$$

The dynamics of a specific strategy s in the population are defined by the ordinary differential equation

$$\frac{\partial P^t(ds)}{\partial t} = \int_S \Phi(s, \hat{s}) \cdot P^t(ds) \ . \tag{2.4}$$

By writing the utility of an agent in the form of an evolutionary measure of success, we obtain the utility function of agent i as the sum of the experienced payoff differences between the own monetary payoff f_i and the monetary payoff of the remaining individual group members f_j:

$$u_i(f_1, ..., f_n) = \sum_{j=1..n, j \neq i} (f_i - f_j) \tag{2.5}$$

The utility of an agent is thus not defined in an absolute way but relative to her environment, by putting the payoff of agent i in relation to the payoff of the remaining population. This form of the utility function describes a population of agents that is exposed to evolutionary dynamics. Positive values of $u_i(f_1, \cdots, f_n)$ are desirable, because they are associated with a higher fertility and a lower mortality. Negative values of $u_i(\cdots)$ should be avoided in order to prevent the evolutionary extinction of the own traits.

2.2.2 The public good games with punishment

In the following, we model the behavior of agents playing a standard one-shot-interaction public goods game with punishment as presented in (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009). Agents are pooled in groups of size n. Each agent i is characterized by a strategy $\hat{s}_i = [m_i, k_i]$ that is defined by two traits. The first trait m_i corresponds to the amount of MUs an agent contributes to the common group project (the public good) and thus reflects the agent's willingness to cooperate. The second trait k_i reflects the agent's propensity to punish defectors in the group. In the first stage of the game, agent i contributes m_i monetary units (MUs) to a common public good which yields a return of g MUs per invested MU. The return from the public good is equally redistributed among the n group members. Agents then learn about the contributions of the other group members. In a second stage, they are provided with the opportunity to punish other group members. Punishment comes in the form of additional costs for both the punisher as well as the punished agent: for each MU spent by the punisher, the return that the punished agent obtained from the public goods game is reduced by r MUs. Given the one-shot-interaction characteristic of the game, punishment does not result in a direct or indirect material benefit and is often considered in the literature to be an altruistic act.

2.2.3 Modeling assumptions

We make the following assumptions about the behavior of agents and the evolutionary environment:

- Agents are assumed to be self-interested and to act rationally given their available information and computational capabilities (von Neumann and Morgenstern, 2007; Simon, Egidi, Viale, and Marris, 2007; Arthur, 1994; Gigerenzer and Selten, 2002). In particular, agents are involved in one-shot interactions only and have no ex-ante information about the others' actions at the time they take their decisions.

- Agent i is assumed to punish agent j according to a function that is linearly increasing with the negative deviation between j's and i's contributions. Specifically, if $m_j - m_i < 0$, agent i punished agent j with

$k{\cdot}(m_i{-}m_j)$ MUs, while j suffers a loss of $r{\cdot}k{\cdot}(m_i{-}m_j)$ MUs. We assume this linear dependency because it can frequently be observed in experiments conducted in the western cultural area (Fehr and Gachter, 2000, 2002, 2005; Egas and Riedl, 2008; Fudenberg and Pathak, 2009). Figure 2.1 illustrates this behavioral pattern for data obtained in three public goods games (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009). The factor k describes the propensity to punish negative deviators.

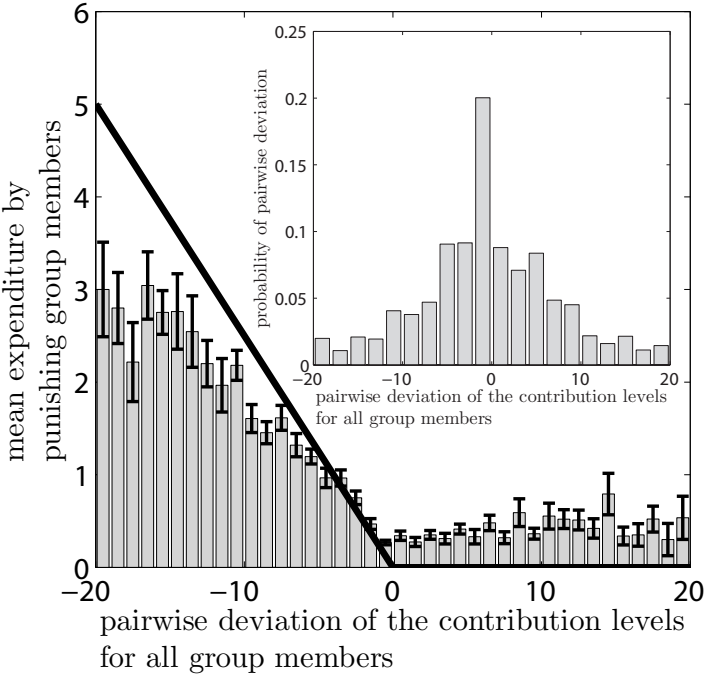


Figure 2.1: Mean expenditure of a given punishing member as a function of the deviation between her contribution and that of the punished member, for all pairs of subjects within a group, as reported empirically (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009). The error bars indicate the standard error around the mean. The straight line crossing zero with a slope of $-k$ shows the average decision rule for punishment. The anomalous punishment of cooperators, corresponding to the positive range along the horizontal axis, is neglected in our model. The inset shows the relative frequency of the pairwise deviations.

- k is assumed to be a common trait or a norm that is shared by all agents within a homogeneous population. It reflects the subjects' genetically and culturally encoded behavior to react to actions that are perceived as being unfair. The interplay of punishment and evolutionary dynamics over hundreds of thousands of years caused the convergence of a previously diverse set of behavioral patterns. This process ultimately led to a common set of behavioral traits which are shared among directly- or indirectly-related and -interacting individuals, e.g. groups originating from the same cultural area. Vice-versa, the prevalent set of behavioral traits determined the anticipated expectations about the behavior of individuals from the same cultural and genetic background. Punishment thus provided the basis for the emergence and manifestation of traits and (social) norms, while simultaneously punishment itself got frequently established as a common trait and norm. In conclusion, humans and our ancestors have converged and evolved to this common norm-enforcing feedback mechanism over hundreds of thousands of years as a result of gene-culture co-evolutionary processes (Henrich, Boyd, Bowles, Camerer, Fehr, Gintis, and McElreath, 2001; Bowles and Gintis, 2004; Gintis, 2003). The subjects' psychological predispositions to render these encoded norms effective ultimately results in the focal action that is observed as a direct and immediate harm towards negative deviators or it acts as a hidden deterrence (Gintis, 2009). Today, lab experiments and field studies such as those of (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009; Henrich et al., 2006; Henrich, Ensminger, McElreath, Barr, Barrett, Bolyanatz, Cardenas, Gurven, Gwako, Henrich, Lesorogol, Marlowe, Tracer, and Ziker, 2010) allow one to sample and observe the statistically stationary characteristics of the common propensity to punish k from subjects originating from a similar cultural background.

- The population of agents is subject to evolutionary dynamics in the form of selection, cross-over and mutation. These three mechanisms affect the viability and fertility of an agent. Viability selection induces a minimal survival condition in the form of a fixed lower value of consumption c_{fix}. This value reflects the basic requirements of an agent, i.e. it defines a lower limit that an agent needs to consume per unit of time in

order to survive. c_{fix} thus constantly absorbs a fraction of the agents' fitness value. Fertility selection accounts for the selection of successful genotypes, i.e. strategies, as opposed to unsuccessful ones. Agents can spread their strategies in the population proportionally to their fitness, e.g. by producing more offsprings. The relative change of the frequency of a trait, i.e. a strategy, is determined by the average success of that trait with respect to the average success of the remaining traits in the population. Cross-over, i.e. the reproduction by mating of two or more agents, accounts for the convergence of the present strategies in the population towards those strategies that are carried by more successful agents. In contrast, mutation induces an additional heterogeneity to the agents' strategy pool and allows the population to explore further the potential strategy space. This ensures that a population of agents is always heterogeneous with respect to the strategies, i.e. $\text{VAR}(m) > 0$ and $\text{VAR}(k) > 0$.

2.2.4 Utility formulation of the public goods game model

We first formulate a utility model assuming complete information. The profit and loss (P&L), i.e. the fitness, of an agent i who plays a public goods game with punishment is determined by the payoff from the game minus the costs of punishing and being punished and minus the contributed effort:

$$\begin{aligned} f_i(m_1, \cdots, m_n) = & -m_i + \frac{g}{n} \cdot (m_i + \sum_{j \neq i} m_j) \\ & - k \cdot r \cdot \sum_{j \neq i} max(m_j - m_i, 0) \\ & - k \cdot \sum_{j \neq i} max(m_i - m_j, 0) \end{aligned} \tag{2.6}$$

The first term in the right hand side of equation (2.6), i.e. m_i, corresponds to the contribution of agent i to the public good. The second term represents the return from the public good. The third and fourth terms quantify the costs of being punished by others and punishing others, respectively. The number of agents in the group is denoted by n, the return from the public good is g per invested MU, and r corresponds to the punishment efficiency factor.

Analogously, the P&L of the remaining agents $j \neq i$ can be written as

$$\begin{aligned} f_j(m_1, ..., m_n) = &-m_j + \frac{g}{n} \cdot (m_i + \sum_{j' \neq i} m_{j'}) \\ &- k \cdot r \sum_{j' \neq j} max(m_{j'} - m_j, 0) \\ &- k \cdot \sum_{j' \neq j} max(m_j - m_{j'}, 0) \ . \end{aligned} \tag{2.7}$$

By substituting equations (2.7) and (2.6) into equation (2.5), we obtain the evolutionary utility of an agent, given by the two-term utility function shown in equation (2.8) below. The first term of (2.8) is defined by equation (2.6): it corresponds to agent i's utility gained from the payoff of the public goods game with punishment. The second term of equation (2.8) defined in (2.7) represents the payoff of the $n-1$ opponents indexed by j. The total utility for agent i is defined by the sum of the differences between all combinations of $f_i(m_1, ..., m_n)$ and $f_j(m_1, ..., m_n)$ with $j \neq i$:

$$u_i(f_1, ..., f_n) = \sum_{j=1..n, j \neq i} (f_i(m_1, ..., m_n) - f_j(m_1, ..., m_n)) \tag{2.8}$$

Consistent with utility theory (even in the presence of bounded rationality) and the underlying evolutionary dynamics, we assume that the agents seek to maximize their utility (von Neumann and Morgenstern, 2007). Obviously, the maximum of the utility function (2.8) can only be calculated in the hypothetical case of complete information about the others' contributions. However, information about the individual contributions $\overrightarrow{m} = (m_1, ..., m_j)$ is not available ex ante, because agents decide about their contributions simultaneously. It follows that agents are required to make assumptions, i.e. to form their first-order beliefs, about the others' contributions. We model this by transforming the utility model in equation (2.8) into a subjective expected utility model.

Therefore, we introduce the subjective probability measure $P_i(m_j)$ that represents agent i's (first-order) belief about the contributions of the other agents. $P_i(m_j)$ quantifies the likelihood as perceived by agent i that another agent

j will contribute m_j MUs[1]. Using $P_i(m_j)$, agent i can form her expectation (Savage, 1972) about the average of the other agents' contributions:

$$E_i[m_j] = \int m_j \cdot P_i(m_j) dm_j \ . \tag{2.9}$$

Similarly to the propensity k to punish, $E_i[m_j]$ can be interpreted as the expected norm-conforming behavior of the population that has co-evolved, learned and internalized across time in a population of interacting agents.

The utility model defined in equation (2.8) is transformed into an expected utility model using the subjective expectations $E_i[m_j]$. Rewriting $f_i(m_1, ..., m_n)$ and $f_j(m_1, ..., m_n)$ by replacing each value $m_j \in [m_1, .., m_{i-1}, m_{i+1}, .., m_n]$ with agent i's subjective expectation $E_i[m_j]$ on m_j gives the following equations:

$$\begin{aligned} E_i[f_i(m_i)] = & -m_i + \frac{g}{n} \cdot m_i \\ & + \frac{g}{n} \cdot (n-1) \cdot \int_0^{\infty} m_j \cdot P_i(m_j) dm_j \\ & - (n-1) \cdot k \cdot r \cdot \int_{m_i}^{\infty} (m_j - m_i) \cdot P_i(m_j) dm_j \\ & - (n-1) \cdot k \cdot \int_0^{m_i} (m_i - m_j) \cdot P_i(m_j) dm_j \end{aligned} \tag{2.10}$$

$$\begin{aligned} E_i[f_j(m_i)] = & -\int_0^{\infty} m_j \cdot P_i(m_j) dm_j + \frac{g}{n} \cdot m_i \\ & + \frac{g}{n} \cdot (n-1) \cdot \int_0^{\infty} m_j \cdot P_i(m_j) dm_j \\ & - k \cdot r \int_0^{m_i} (m_i - m_j) P_i(m_j) dm_j \\ & - k \cdot \int_{m_i}^{\infty} (m_j - m_i) P_i(m_j) dm_j \end{aligned} \tag{2.11}$$

Note that, in the formation of the expectation by agent i of the others' utility functions, agent i's own contribution m_i is obviously known to her, hence the term $\frac{g}{n} \cdot m_i$ appears without averaging.

[1] In the one-shot game version studied here, all agents $j \neq i$ are indistinguishable from the point of view of an agent i, i.e., agent i has no information on any preference, trait or specific characteristics of the other agents.

As in the case of complete information, agents seek to maximize their relative fitness, i.e. the sum of the differences between their own P&L, $f_i(m_1, ..., m_n)$, and the others' P&L. Putting all this together, we obtain the expected utility function $u_i(E_i[f_i(m_i)], E_i[f_j(m_i)])$ of agent i as shown in equation (2.12).

$$u_i(E_i[f_i(m_i)], E_i[f_j(m_i)]) = (n-1) \cdot (E_i[f_i(m_i)] - E_i[f_j(m_i)]) \tag{2.12}$$

We start our analysis by a classical utility optimization problem. Agents maximize $u_i(E_i[f_i(m_i)], E_i[f_j(m_i)])$ with respect to their contribution m_i:

$$m_i \in \arg\max_{m_i} \quad u_i(E_i[f_i(m_i)], E_i[f_j(m_i)]) \tag{2.13}$$

The first order condition of problem (2.13) reads

$$\frac{\partial u_i(E_i[f_i(m_i)], E_i[f_j(m_i)])}{\partial m_i} \stackrel{!}{=} 0 \ , \tag{2.14}$$

with

$$\begin{aligned} \frac{\partial u_i(E_i[f_i(m_i)], E_i[f_j(m_i)])}{\partial m_i} &= (n-1) \cdot \Big(\frac{\partial f_i(m_i, P_i(m_j))}{\partial m_i} - \frac{\partial f_j(m_i, P_i(m_j))}{\partial m_i} \Big) \\ &= \Big(-1 - k \cdot (1 + r - n \cdot r) \int_{m_i}^{\infty} P_i(m_j) dm_j \\ &\quad + k \cdot (1 - n + r) \cdot \int_0^{m_i} P_i(m_j) dm_j \Big) \cdot (n-1) \ . \end{aligned} \tag{2.15}$$

The second-order condition for a local maximum of (2.13) holds for any reasonable assignment of the problem parameters, i.e.

$$\frac{\partial^2 u_i(E_i[f_i(m_i)], E_i[f_j(m_i])}{\partial m_i^2} < 0, \forall k > 0, n > 0, g > 0, r > 0, 0 < m_i < \infty \ .$$

The cumulative distribution function of the contributions m_j of the other agents, as anticipated by agent i, is defined by $CDF_i(m_i) \equiv \int_0^{m_i} P_i(m_j) dm_j$.

The term $\int_{m_i}^{\infty} P_i(m_j)dm_j$ in equation (2.15) corresponds to the survival function of the subjective expected distribution of contributions in the population:

$$a_i(m_i) := 1 - CDF_i(m_i) = P_i(\{m_j > m_i\}) = \int_{m_i}^{\infty} P_i(m_j)dm_j \tag{2.16}$$

Substituting $a_i(m_i)$ as defined in equation (2.16) into equation (2.15) yields:

$$-1 + \frac{g}{n} + k \cdot (n-1) \cdot (a_i(m_i) \cdot r + a_i(m_i) - 1) \stackrel{!}{=} 0 \tag{2.17}$$

Equation (2.17) describes a functional relation between the predetermined parameters of the public goods game, i.e. the group size n, the public good yield factor g and the punishment efficiency r, as well as the variable traits of agent i, i.e. the propensity k to punish and her subjective expectation (first-order belief) about the fraction $a_i(m_i)$ of her group fellows who contribute more than her own contribution m_i.

As we are interested in the agents' evolutionary optimal punishment behavior, we solve equation (2.17) for k and obtain:

$$k_i^{\star} = \frac{1}{1 - n + r + a_i(m_i) \cdot (n-2) \cdot (1+r)} \tag{2.18}$$

$k_i^{\star}$ depends on m_i via the agent i's subjective (first-order) belief embodied in $a_i(m_i) \in [0, 1]$ that the other agents will contribute more than herself. The value $k_i^{\star}$ can be interpreted as the value that makes agent i better off not to deviate negatively or positively from her willingness to contribute m_i MUs to the public good, given she believes a number of $N = n \cdot a_i(m_i)$ of other group fellows contribute more than her own contribution m_i. Equation (2.18) thus determines a strategy profile $s^{\star} = [m_i, k_i]$ that represents a Nash equilibrium.

In the following subsection, we add evolutionary dynamics to our model.

2.2.5 The evolutionary dynamics

The evolutionary dynamics of agents, who face a social dilemma situation in the form of a public goods game with punishment, can be captured by the variations of the P&L as a function of the deviation in the contribution level $m_i(t)$ and in the population's propensity to punish k. If agent i starts to deviate from her current level of cooperation $m(t)$ by a value of $\Delta m = m(t+1) - m(t)$, the absolute change of the P&L for the agent as a function of Δm and k is defined as follows:

$$\Delta \mathrm{P\&L}_i(\Delta m, k) = \begin{cases} -g \cdot \dfrac{\Delta m}{n} + \Delta m - (n-1) \cdot k \cdot \Delta m \cdot r, \ \Delta m \leq 0 \\ g \cdot \dfrac{\Delta m}{n} - \Delta m - (n-1) \cdot k \cdot \Delta m \ , \ \Delta m > 0 \end{cases} \tag{2.19}$$

The deviation of agent i by Δm affects not only her own P&L, but also the P&L of the remaining agents $j = 1 \cdots n, j \neq i$. The absolute change of the P&L of the remaining population as a function of Δm and k reads

$$\Delta \mathrm{P\&L}_j(\Delta m, k) = \begin{cases} -g \cdot \dfrac{\Delta m}{n} - k \cdot \Delta m \cdot, \ \Delta m \leq 0 \\ g \cdot \dfrac{\Delta m}{n} - k \cdot \Delta m \cdot r, \ \Delta m > 0 \end{cases} . \tag{2.20}$$

Putting equations (2.19) and (2.20) together with

$$\widetilde{\Delta} \mathrm{P\&L}_i(\Delta m, k) := \Delta \mathrm{P\&L}_i(\Delta m) - \Delta \mathrm{P\&L}_j(\Delta m) \tag{2.21}$$

yields the relative change of the P&L of agent i with respect to the remaining population:

$$\widetilde{\Delta} \mathrm{P\&L}_i(\Delta m, k) = \begin{cases} \Delta m - (n-1) \cdot k \cdot \Delta m \cdot r + k \cdot \Delta m, \ \Delta m \leq 0 \\ -\Delta m - (n-1) \cdot k \cdot \Delta m + k \cdot \Delta m \cdot r, \ \Delta m > 0 \end{cases} \tag{2.22}$$

The form of equation (2.22) is equivalent to the relative measure of success of a strategy introduced in equation (2.3) with $s := [\Delta m, k]$. As introduced above, the realized P&L from the public goods game with punishment can be interpreted as the fitness of an agent in an evolutionary environment. The fitness, in turn, is associated with the rate of fertility, i.e. the fitter an agent

becomes, the more genetically related offsprings she produces. In this way, traits of agents with a higher realized P&L value tend to spread and to end up dominating the population over time. It thus holds that the traits $[m, k]$ in the population move with time towards values $[\hat{m}, \hat{k}]$ of a subpopulation that on average achieves a higher mean P&L than the average mean P&L of the entire population.
The corresponding replicator dynamics are

$$\begin{aligned} \frac{\partial x(\Delta m)}{\partial t} &= \int_0^\infty \tilde{\Delta}\mathrm{P\&L}(\Delta m, k) \cdot x(\Delta m)\ dk \\ \frac{\partial x(k)}{\partial t} &= \int_{-\infty}^\infty \tilde{\Delta}\mathrm{P\&L}(\Delta m, k) \cdot x(k)\ d\Delta m\ . \end{aligned} \tag{2.23}$$

with $x(\Delta m)$ and $x(k)$ being the proportion of agents deviating by Δm and with a propensity to punish k. The dynamics for the expected group average, $\bar{m}$ and $\bar{k}$, are accordingly defined by

$$\begin{aligned} \frac{\partial E[\bar{m}]}{\partial t} &= \int_{-\infty}^\infty \int_0^\infty \Delta m \cdot \tilde{\Delta}\mathrm{P\&L}(\Delta m, k) \cdot x(\Delta m) \cdot x(k)\ dk\ d\Delta m \\ \frac{\partial E[k]}{\partial t} &= \int_0^\infty \int_{-\infty}^\infty k \cdot \tilde{\Delta}\mathrm{P\&L}(\Delta m, k) \cdot x(\Delta m) \cdot x(k)\ d\Delta m\ dk\ . \end{aligned} \tag{2.24}$$

The sensitivity of $\widetilde{\Delta}P\&L_i(\Delta m, k)$ with respect to the relative change of Δm is defined by the partial derivative

$$\frac{\partial \widetilde{\Delta}\mathrm{P\&L}_i(\Delta m, k)}{\partial \Delta m} = \begin{cases} -1 - k + k \cdot (n-1) \cdot r & ,\ \Delta m \leq 0 \\ \\ -1 - k \cdot (n-1) + k \cdot r & ,\ \Delta m > 0 \end{cases} . \tag{2.25}$$

With the conditions that $n \geq 2$ and $r > 1$, i.e. a game has always two or more players and punishment is less costly to the punisher than to the punished

agent, it holds that for $\Delta m(t) > 0$ the piecewise definition of $\frac{\partial \widetilde{\Delta}\text{P\&L}_i}{\partial \Delta m}$ is always negative and for $\Delta m < 0$ it follows that

$$\begin{aligned} &\text{a)} \quad \frac{\partial \widetilde{\Delta}\text{P\&L}_i(\Delta m, k)}{\partial \Delta m} \leq 0 \ , \quad \forall k \leq \frac{1}{n \cdot r - r - 1}, \ \Delta m < 0 \\ &\text{b)} \quad \frac{\partial \widetilde{\Delta}\text{P\&L}_i(\Delta m, k)}{\partial \Delta m} > 0 \ , \quad \forall k > \frac{1}{n \cdot r - r - 1}, \ \Delta m < 0 \ . \end{aligned} \tag{2.26}$$

This reveals the existence of two distinct evolutionary regimes that are separated by the bifurcation point at

$$k^+ = \frac{1}{n \cdot r - r - 1}. \tag{2.27}$$

- *Defection*: For $k \leq \frac{1}{n \cdot r - r - 1}$ and $\text{Var}(m_j) > 0, \ j = 1, ..., n$, the linear P&L structure of the public goods game with punishment together with the replicator dynamics are responsible for Δm to become more negative over time. It intuitively follows that defection pays out, such that

$$m^a := \lim_{t \to \infty} m(t) \approx \frac{c_{\text{fix}}}{g - 1} \tag{2.28}$$

  results as the evolutionary stable strategy (ESS). Remember that each agent has a minimum cost of living defined by c_{fix}. In order to meet this survival condition, the average minimum contribution of the population is constrained to values of $m > \frac{c_{\text{fix}}}{g-1}$.

- *Coordination*: For $k > \frac{1}{n \cdot r - r - 1}$, a heterogeneous population with $\text{Var}(m_j) > 0, \ j = 1, ..., n$ follows a dynamic that does not converge to a predetermined unique evolutionary attraction point but rather converges to an evolutionary stable set of strategies. As punishment is efficient in this regime, with $\frac{\partial \widetilde{\Delta}\text{P\&L}_i(\Delta m, k)}{\partial \Delta m} > 0$ for values of $\Delta m < 0$, the social dilemma problem transforms into a coordination problem (Fehr and Schmidt, 1999). If punishment is efficient, the utility maximizing strategy is to contribute according to the expected contribution of the remaining group fellows, i.e. to contribute according to the first-order belief. Following Black's theorem, the best estimate for this strategy is the median value $\bar{m}_i$ of the subjective probability measure P_i that is believed to charac-

terize the contributions of the group fellows (Black, 1948; Arrow, 1970; Bernheim, 1994; Selten and Ostmann, 2000). The median value $\bar{m}_i$ of the subjective probability distribution P_i is defined by

$$\int_{\bar{m}_i}^{\infty} P_i(m_j) dm_j = \frac{1}{2} \tag{2.29}$$

Consequently

$$m^b := \lim_{t \to \infty} m(t) = \bar{m} \tag{2.30}$$

results as an ESS in the population.

The population of agents initially consists of uncooperative, non-punishers, i.e. $k_i(0) \simeq 0$ and $m_i(0) \simeq 0$ for $i = 1, ..., n$. The utility maximization problem in equation (2.13) determines the optimal level of punishment as defined in equation (2.18) with

$$k_i^\star = \frac{1}{1 - n + r + a_i(m_i) \cdot (n-2) \cdot (1+r)}$$

It follows that

$$k_i(0) \simeq 0 \ \wedge \ \lim_{t \to \infty} k_i(t) = k_i^\star \quad \longrightarrow \quad 0 \leq k_i(t) \leq k_i^\star \ \ \forall t \ . \tag{2.31}$$

and thus the value range of the propensity to punish is restricted to the interval $k_i(t) \in [0, k_i^\star]$. With the population being initialized at $k_i(0) \simeq 0 \ll k^*$, it follows that agents initially have an incentive to defect as can be inferred from equation (2.26a). In other words, agents have an incentive to contribute less than the amount contributed by the other group fellows. In general, agents have no ex-ante information about the others' contributions at the time they take the decision to contribute m_i MUs. However, agents have beliefs about the others' contribution that is embodied in the subjective probability distribution P_i. This allows them to form their expectations about the group average contribution as defined in equation (2.9). In terms of equation (2.16), "defecting" translates into a probability of one that all m_j values are larger

than the own contribution m_i, i.e. $a_i(m_j) = 1$. With $a_i(m_j) = 1$, it follows that the optimal propensity to punish defined in equation (2.18) becomes

$$\begin{aligned} k^a &= \frac{1}{1 - n + r + (n-2) \cdot (1+r)} \\ &= \frac{1}{(n-1) \cdot r - 1} \ . \end{aligned} \tag{2.32}$$

which is exactly equivalent to the evolutionary threshold value of k^+ defined in equation (2.27). Plugging k^a into equation (2.22) yields

$$\widetilde{\Delta}\mathrm{P\&L}_i(\Delta m, k^a) = \begin{cases} 0 & , \ \Delta m \leq 0 \\ -\frac{\Delta m \cdot (n-2) \cdot (r+1)}{r \cdot (n+1) - 1} < 0 & , \ \Delta m > 0 \end{cases} . \tag{2.33}$$

Together with the replicator dynamics defined in equation (2.24), it follows that for all values of $k \leq k^+$ the population converges towards the evolutionary stable attraction point for m_i that is defined in equation (2.28). Consequently the ESS $s^a = [m^a, 0 \leq k \leq k^a]$ ends up dominating the population.

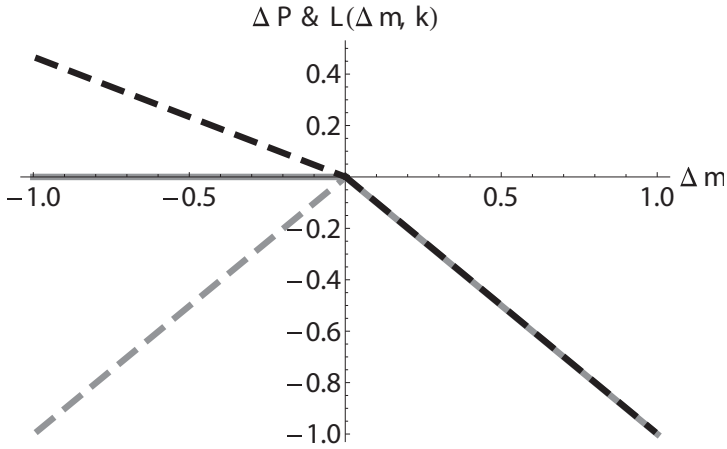


Figure 2.2: Sensitivity of $\tilde{\Delta}\mathrm{P\&L}(\Delta m, k)$ as a function of a relative change Δm of the contributions for a group size of $n = 4$, a punishment efficiency $r = 3$ and a propensity to punish of $k = 0.125$ (grey), $k = \frac{1}{15}$ (black, dashed) and $k = 0.25$ (grey dashed).

In contrast, for values of $k > k^+$, the social dilemma problem turns into a coordination problem. Consequently, an evolutionary stable attraction point m^b emerges that is defined by equations (2.29) and (2.30), respectively. As explained above, m^b corresponds to the median of all m_i values present in the population. The evolutionary attraction point m^b implies that $a_i(m^b) = \frac{1}{2}$, i.e. each agent contributes according to the value that matches the median value of the subjectively expected distribution of populations contributions values m_j. Plugging this into equation (2.18) yields an evolutionary stable strategy for k^b given by

$$k^b = \frac{2}{n \cdot (r-1)} \tag{2.34}$$

Substituting k^b into equation (2.22) results in a ΔP&L profile that is determined by symmetrically downward sloping functions $\widetilde{\Delta}\text{P\&L}_i(\Delta m, k^b)$ centered relative to the maximum at $\Delta m = 0$ with

$$\widetilde{\Delta}\text{P\&L}_i(\Delta m, k^b) = \begin{cases} \frac{\Delta m \cdot (n+2) \cdot (r+1)}{r \cdot (n-1) - 1} < 0 & , \ \Delta m \leq 0 \\ -\frac{\Delta m \cdot (n+2) \cdot (r+1)}{r \cdot (n-1) - 1} < 0 & , \ \Delta m > 0 \end{cases} . \tag{2.35}$$

Consequently, in the presence of the evolutionary dynamics, the population converges to the ESS given by $s^b = [m^b, k^b]$.

Figure 2.2 depicts the structure of equation (2.22) with a punishment efficiency factor of $r = 3$ and a group size $n = 4$ for $k = \frac{1}{15}$ (black, dashed), $k = 0.125$ (grey) and $k = 0.25$ (grey, dashed).

The following subsection analyzes the identified ESSs for a population of agents that is either purely self-regarding and acting selfishly or a population of agents that incorporates other-regarding preferences in their decision process.

2.2.6 The effect of self and other-regarding preferences

First, consider a population of purely self-regarding and selfish acting agents, i.e. agents who try to maximize their utility without e.g. taking into account specific preferences with respect to the P&L and the contributions of the remaining agents in the group. The preferences of self-regarding and selfish agents are simply characterized by the dislike of situations in which their P&L

in current period t is less than their P&L in the previous period $t-1$. This implies that all agents in the population are satisfied if and only if the following expression is fulfilled

$$\begin{aligned} &f_i(m_1(t), \cdots, m_n(t)) \geq f_i(m_1(t-1), \cdots, m_n(t-1)) \quad \wedge \\ &f_j(m_1(t), \cdots, m_n(t)) \geq f_j(m_1(t-1), \cdots, m_n(t-1)) \end{aligned} \tag{2.36}$$

with $f_i(...)$ and $f_j(...)$ being defined in equation (2.6) and (2.7), respectively. Reducing the expression in (2.36) over the domain of reasonable values for the variables $m_j \geq 0 \ \ \forall j = 1, \cdots, n$, $k_i \geq 0$, $0 \leq a_i(m_i) \leq 1$, $n \geq 2$, $0 < g < n$ and $r > 1$ and solving it to the propensity to punish k gives the following condition for k:

$$k^s \geq \frac{n-g}{(n-1) \cdot n \cdot r}. \tag{2.37}$$

For all reasonable values of $n \geq 2$, $g \geq 0$ and $r \geq 1$ and assuming that agents are initially non-punishers, i.e. $k_i(0) \simeq 0$, it holds that the propensity to punish of self-regarding and selfish agents is always less than the bifurcation threshold k^+, defined in equation (2.27). Thus, selfish and purely self-regarding agents are inevitably caught in the *defection* regime, as k^s does not allow to overcome the bifurcation hurdle at k^+. Consequently, the population converges towards the ESS that is defined by $s^a = [m^a, 0 < k^s < k^a]$.

Consider now a population of agents who display other-regarding behavior in the form of disadvantageous inequity aversion. In general, inequity aversion preferences relate the personal utility gained from a public good to the personal contributed effort. If an imbalance exists between the own contributed effort and the personally received payoff compared to the performed effort and the received payoff of other agents in the group, the outcome of the game is perceived as being inequitable or "unfair". Disadvantageous inequity aversion implies that subjects only dislike situations in which the inequity is to their disadvantage. The payoff of an agent i, who plays a public goods game with punishment, is defined by equation (2.6) and the personal effort is equivalent to the contributed amount of MU m_i. An agent with an aversion against disadvantageous inequitable outcomes thus does not like situations in which

- she contributes equally or more than her group fellows ($m_i \geq m_j$) and receives a payoff that is smaller than the average utility received by the remaining group members ($f_i < f_j$) or
- she contributes more to the public good ($m_i > m_j$) and, at the same time, receives a payoff that is smaller or equal to the remaining group's utility ($f_i \leq f_j$).

By implication, the population of agents is satisfied only if at least one of the following three conditions is fulfilled $\forall j = 1, ..., i-1, i+1, ..n$:

$$\begin{aligned} &\text{a) } f_i(m_1, ..., m_n) > f_j(m_1, ..., m_n) \ \wedge \ m_i > m_j \ , \\ &\text{b) } f_i(m_1, ..., m_n) \geq f_j(m_1, ..., m_n) \ \wedge \ m_i = m_j \ , \\ &\text{c) } f_i(m_i, ..., m_n) < f_j(m_j, ..., m_n) \ \wedge \ m_i < m_j \ . \end{aligned} \tag{2.38}$$

Expressing the above conditions (2.38) over the domain of eligible values for the variables $m_j \geq 0 \ \ \forall j = 1, \cdots, n$, $k_i \geq 0$, $0 \leq a_i(m_i) \leq 1$, $n \geq 2$, $0 < g < n$ and $r > 1$ and solving them in terms of the propensity to punish k yields the following inequality

$$k^{ieq} > \frac{1}{a_i(m_i) \cdot (r+1) \cdot (n-2) + r + 1 - n} \ . \tag{2.39}$$

As introduced above, the evolutionary dynamics induce a tendency towards defection in a population that initially consists of uncooperative agents who display no propensity to altruistically punish defectors, i.e. $k_i(0) \simeq 0 \ll k^+$ and $m_i(0) \simeq \frac{c_{\text{fix}}}{g-1}$. In the case of self-regarding agents, the contribution m_i of a given agent i is chosen in a way that it can be expected to be surely less than what the other agents in the group contribute, i.e. $a_i(m_i) = 1$. With $a_i(m_i) = 1$ the condition in equation (2.39) for the optimal level of punishment becomes

$$k^{ieq} > \frac{1}{n \cdot r - r - 1} \ . \tag{2.40}$$

The minimum level of punishment k^{ieq} that is required to satisfy the disadvantageous inequity aversion conditions in equation (2.38) exceeds the evolutionary threshold k^+. Thus, agents are forced to switch from the *defection* regime into the *coordination* regime in order to satisfy their preferences. As described

before, the best response strategy in the coordination regime regarding the level of cooperation m_i is to contribute according to the median value of the subjective probability distribution P_i. By the definition in equation (2.16), it follows that the median value $\bar{m}_i$ of P_i is equivalent to a value of $a_i(\bar{m}) = 0.5$. Plugging $a_i(m_i) = 0.5$ into equation (2.18) yields the following estimate for the optimal propensity to punish

$$k^b = \frac{2}{n \cdot (r-1)} \ . \tag{2.41}$$

k^b is always lager than the evolutionary threshold of k^+ for all reasonable values of $n \geq 2$ and $r > 1$. The population of agents is thus able to maintain a stable level of cooperation at the median value $\bar{m}$ that is determined by the initial distribution P of the contributions. In conclusion, a population of disadvantageous inequity averse agents converges to the ESS that is determined by $s^b = [m^b, k^b]$.

Our first main result can be summarized as follows:

Result 2.1: *In the presence of standard Darwinian evolutionary dynamics, agents' traits (strategies) converge to evolutionary stable strategies, which results in a public goods game with punishment to be either characterized by defection (for weak punishment) or by coordination (for sufficient punishment). Purely self-regarding agents are inevitably caught in the defection regime while disadvantageous inequity averse agents are able to resolve the social dilemma by transforming it into a coordination problem.*

In the following section, we turn to the empirical validation of our model.

2.3 Empirical test of the theory

In this section, we compare the predictions derived from our model with the empirical data obtained in three independently conducted lab experiments and validate our results against the empirical observations.

2.3.1 Description of the empirical data set

We analyze data from three public goods game experiments with punishment (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009), which were carried out independently. In each experiment, groups of $n = 4$ subjects played a two-stage public goods game: at the beginning of stage one, the contribution step, individuals were endowed with 20 monetary units (MUs). Subjects could decide on the amount m_i of MUs to contribute to the public good. The sum of all contributions was compounded by a factor of $g = 1.6$ and subsequently redistributed in equal shares to all group members. Note that this results in a per capita gain of $0.4 < 1$ per contributed MU which induced a distinct social dilemma component. In the second stage, the punishment step, subjects were informed about the contributions of their group mates. Subsequently, they could spend an additional fraction of their endowment to punish other group fellows. Each MU spent by the punisher caused a harm of approximately[2] $r = 3$ MUs to the punished subject.

These two stages were played repetitively either in a stranger or a partner treatment. In the former, group members were reshuffled after each iteration to preserve the characteristics of one-shot interactions, i.e., to control for direct reciprocal effects. In the partner treatment, subjects played continuously with the same group members across all periods. The first experiment was composed of both, a stranger and a partner treatment. Each of them were played for 10 periods. The second and third experiments included only a stranger treatment and were played for 6 and 10 iterations, respectively. In addition, the third experiment differed in the way information about the received punishment was revealed to the punished subjects. In the first one, the so-called observed treatment, subjects were informed immediately after the punishment

[2] In the first experiment, the punishment efficiency factor was determined based on the first stage payoff of the punished individual. However, it can be considered to be approximately equal to the factor 3 as in the remaining two experiments.

stage about the costs of the received punishment, as in experiments one and two. In contrast, in the second treatment, the unobserved treatment, subjects were informed about the costs they had to bear for being punished only after the last period had been played. However, the results of both treatments were found not to be significantly different as the fear of punishment seems to be as effective as the punishment itself (Fudenberg and Pathak, 2009). To obtain a sufficiently large sample size, we pool the observations from all treatments of the three experiments introduced above. The subject pool size amounts to a total of 440 subjects.

2.3.2 Recovering the propensity to punish from the empirical data

The empirical propensity to punish can be calculated by taking the observed deviations $(m_i - m_j) > 0$ between subject i and j and the observed punishment from subject i to j, $p_{i\to j}$. In this way, each pairwise interaction between two subjects provides a realization for the propensity to punish according to the formula

$$k_{i,j} = \frac{p_{i\to j}}{m_i - m_j} \ . \tag{2.42}$$

With the set of all pairwise interactions, we construct the empirical distribution of the propensity to punish, by sampling all realized $p_{i\to j}$ with their corresponding m_i and m_j.

As shown in the first section and also demonstrated in chapter 3, the agents' propensity to punish can be interpreted as a norm-enforcing behavior that has co-evolved over tens and hundreds of thousands of years by gene-culture co-evolution along with the emergence of an aversion to disadvantageous inequitable outcome. The perception of fairness and the reaction to unfair behavior seems to be deeply rooted in our cultural and genetic heritage (Henrich et al., 2001; Gintis, Bowles, Boyd, and Fehr, 2003), as experiments and field studies across different locations and cultural groups suggest (Henrich, 2004; Henrich et al., 2006). We thus consider the propensity to punish k to be a constant on the evolutionary negligible short time-scale of the experiments. This can be substantiated by comparing the results of a two-sample Kolmogorov-Smirnov test between an empirical data set containing only data from the first

period and the corresponding full-sample data set. The null hypothesis that the distributions of the two data sets of $k_{i,j}$ result from the same generating mechanism cannot be rejected with a p-value equal to 0.31. In all three experiments, the observed contributions $m_i \gg 0$ are approximately stable over time, as they do not converge towards full defection. Additionally, the standard deviation of the contributions is on average decreasing over time. Both of these measures indicate that the subjects in the experiments are in the "*coordination*" regime.

2.3.3 Validation of the model prediction for k

We validate the model presented in section 2.2 by asking whether the ESS value k^b of the propensity to punish in the *coordination* regime given by equation (2.34) matches the empirically observed data. The group size n and punishment efficiency r are known parameters in the experiments. The three public goods game experiments with punishment (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009) have been performed with $n = 4$ players and a punishment efficiency factor of $r = 3$, respectively. Plugging both values into equation (2.34) yields

$$k^b = \frac{1}{4} \; . \tag{2.43}$$

As the value given by (2.43) is based on the assumption that subjects contribute according to the median value of their subjective probability distribution about the contributions of their group fellows, k^b corresponds consequently to the median of the distribution of the values $\{k_{i\to j}\}$ of the propensity to punish.

Remarkably, we find an exact match with the median value $\tilde{k}_{\text{emp}}$ estimated from the empirical distribution of the $\{k_{i,j}\}$ values, i.e. $k^\star = \tilde{k}_{\text{emp}} = 0.25$. The standard error of the median of the empirical data is $\hat{\sigma}^k_{\text{med}} = 0.0013$. This corresponds to a one-standard error range given by $\tilde{k}_{\text{emp}} \pm \hat{\sigma}^k_{\text{med}} = [0.2487, 0.2513]$. The corresponding 95% confidence intervals for the sample median values are $CI^\star_{0.95} = [0.2423, 0.2655]$, $CI^+_{0.95} = [0.2486, 0.3336]$, $CI^-_{0.95} = [0.1568, 0.2611]$ and $C^{\text{all}}_{0.95}[0.25, 0.25]$[3].

[3]We used a bootstrap t-method presented in (Efron and Tibshirani, 1994) to estimate the confidence intervals. The superscript on the CI indicates the individual data sets:

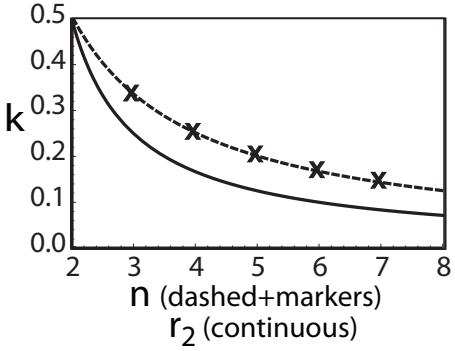


Figure 2.3: Propensity to punish as a function of the punishment efficiency r (continuous line) for a fixed group size $n = 4$ and as a function of the group size n (dashed line with cross markers) for a fixed $r = 3$.

This remarkable agreement between theory and empirical data suggests that subjects act according to the optimization problem defined in (2.13) and that their punishment behavior is dominated by disadvantageous inequity aversion preferences defined in equation (2.38). Again, we argue that in this specific setup the focal action to punish negative deviators by spending roughly a fourth of the negative deviation has emerged as the result of the human's psychological predisposition to render effective the culturally and genetically internalized norms (Gintis, 2009; Hetzer and Sornette, 2010). In this case, these norms are described by disadvantageous inequity aversion.

We can now state our second main result:

Result 2.2: *The level of altruistic punishment that subjects exhibit in public goods game experiments can be explained by a simple aversion to disadvantageous inequitable outcomes together with the individual maximization of the expected utility defined in equation (2.13).*

$\star$ =(Fehr and Gachter, 2000)
$+$ =(Fehr and Gachter, 2002)
$-$ =(Fudenberg and Pathak, 2009)
$^{\text{all}}$=pooled data set of all three experiments

The dependence of the optimal propensity to punish k^b defined in equation (2.34) on the group size n and the punishment efficiency factor r is plotted in figure 2.3. This predicts the potential propensity to punish that should be observed in experiments with differing configurations. In particular, the larger the punishment efficiency r and the group size n, the smaller becomes the optimal propensity to punish. To validate these predictions additional experiments with different groups sizes and punishment efficiency factors have to be performed in future research.

The following section analyzes the co-evolutionary dynamics of agents with disadvantageous inequity aversion compared to agents with purely self-regarding and selfish behavior in a heterogeneous population.

2.4 Evolutionary dominance of other-regarding preferences

The results and findings presented in the previous two sections of this chapter inevitably raise the question about the evolutionary stability and dominance of other-regarding compared to self-regarding preferences. Are agents with other-regarding behavior able to invade a population of initially selfish and self-regarding agents? Can the required conditions for the emergence of altruistic punishment spread in a population of agents that is facing a competitive resource limited environment as described by our model? Is disadvantageous inequity aversion the predominant strategy in a population of agents who face a social dilemma situation that provides the opportunity to punish? This section addresses these questions by providing an analysis of the co-evolutionary dynamics that are at play in a heterogeneous population consisting of a mixture of disadvantageous inequity averse agents and purely self-regarding and selfish-acting agents.

A system that is subject to evolutionary forces is characterized and determined by selection, cross-over and mutation processes. Consequently, the birth and death of agents induce multifaceted and complex co-evolutionary dynamics that are contingent on the states and path dependencies of the individual actors in the system. In view of this complexity, this section presents a simplified

but conclusive analytical representation of the system's dynamics and properties. This is achieved by reducing the assumed heterogeneity in the system and by considering only two groups and types of agents, respectively. An extensive numerical analysis of a population of agents playing a public goods game with punishment that takes into account the full heterogeneity and the full set of evolutionary dynamics and path dependencies is presented in chapter 3.

2.4.1 Conditions for evolutionary dominance

Let us write the evolutionary success of a homogeneous group **A** of agents with size d playing strategy $s_1 = [m_1, k_1]$ that competes with a homogeneous group **B** of size $n-d$ with agents playing strategy $s_2 = [m_2, k_2]$. Using equation (2.3) and the P&L structure of the public goods game with punishment defined in the equations (2.6,2.7), we obtain

$$\begin{aligned}\Phi(d,n,k_1,k_2) =& \sum_{d} f_1(m_1,\cdots,m_n) - \sum_{n-d} f_2(m_1,\cdots,m_n) \\ =& \sum_{d} m_1 + \sum_{n-d} m_2 - \\ & \sum_{d}\sum_{n-d} k_1 \cdot max(m_1 - m_2, 0) + \sum_{d}\sum_{n-d} k_1 \cdot max(m_1 - m_2, 0) \cdot r - \\ & \sum_{n-d}\sum_{d} k_2 \cdot max(m_2 - m_1, 0) \cdot r + \sum_{n-d}\sum_{d} k_2 \cdot (m_2 - m_1, 0) \ .\end{aligned} \tag{2.44}$$

Expression (2.44) can be rewritten by forming the expectations with respect to the evolutionary success Φ and assuming that group **A** randomly varies in the contribution behavior of its agents. Therefore, the contribution m_1 (per agent) of group **A** is assumed to deviate from the contribution m_2 (per agent) of group **B**. The total expected deviation of group **A** is defined by $\Delta\hat{m} = p_1 \cdot (-\Delta m) + (1 - p_1) \cdot \Delta m$ where $\Delta m = |m_1 - m_2|$. Each of the two groups is assumed to be intrinsically homogeneous but differs from each other, not only in the expected contributions, but also with respect to the punishment behavior, i.e. $k_1 \neq k_2$. Agents in group **A** are characterized by the propensity to punish k_1, while group **B** exhibits a propensity to punish that corresponds to k_2. The average evolutionary success (or failure) of group

A with d members who deviate negatively with a given probability p_1 or positively with the probability $1 - p_1$ by a value Δm from the contribution m_2 of group **B** which has a total of $n - d$ members is given by

$$\begin{aligned}\Phi^+(d, n, p_1, k_1, k_2) =& (1 - p_1) \cdot (d \cdot (-\Delta m) + \frac{d^{\,2}\Delta m \cdot g}{n} - \frac{(n-d) \cdot d \cdot \Delta m}{n} - \\ & d \cdot (n-d) \cdot k_1 \Delta m + (n-d) \cdot d k_1 r) + \\ & p_1 \cdot (d \cdot \Delta m \frac{d^{\,2}\Delta m \cdot g}{n} - \frac{(n-d) \cdot d \Delta m \cdot g}{n} - \\ & d \cdot (n-d) \cdot k_2 \cdot \Delta m \cdot r + (n-d) \cdot d \cdot k_2 \cdot \Delta m) \; .\end{aligned} \tag{2.45}$$

The measure Φ^+ defines a relation between the relative difference of the P&L of group **A** versus that of group **B**. It thus reflects the evolutionary success or failure of the two competing groups over time. An expected deviation of group **A** by a value of $\Delta \hat{m}$ affects Φ^+ to become either positive or negative. Depending on the sign of Φ^+, either the strategies of group **A** start to dominate the population ($\Phi^+ > 0$) or alternatively, if $\Phi^+ < 0$, the strategies of group **B** spread and dominate in the population.

2.4.2 Evolutionary dominance of disadvantageous inequity averse agents

Consider a population of size n that initially consists only of purely self-regarding and selfish acting agents. This homogeneous population is assumed to be in an evolutionary equilibrium state. As identified in the previous sections, self-regarding agents play the ESS $s^a = [m^a, k^s]$ with

$$k^s = \frac{n - g}{(n-1) \cdot n \cdot r}$$

and

$$m^a \approx \frac{c_{\text{fix}}}{g - 1}$$

as given by the equations (2.28) and (2.37). Replacing one agent in this population by a disadvantageous inequity averse agent leads to a heterogeneous population that consists of two homogeneous subgroups. In the following, we

analyze the co-evolutionary dynamics of this heterogenous population of agents that is composed of a group **A** with size $n-1$ of purely self-regarding agents and a group **B** with a single disadvantageous inequity averse agent and size $d=1$.

In contrast to the self-regarding agents, disadvantageous inequity averse agents play the ESS given by $s^b = [m^b, k^b]$ with

$$k^b = \frac{2}{n \cdot (r-1)}$$

and

$$m^b = \bar{m}$$

as defined in equations (2.41) and (2.30). Substituting $k_1 = k^b$, $k_2 = k^s$ and $d=1$ into equation (2.45) yields

$$\begin{aligned}\Phi^\star(1, n, p_1, k^b, k^s) = &\frac{(p_1-1)\cdot(2-n)\cdot\Delta m p_1 \cdot (n-g)\cdot\Delta m}{n\cdot r} + \\ &\frac{(g\cdot(2-3\cdot p_1 + n\cdot(2\cdot p_1 - 1)))}{n}\,.\end{aligned} \tag{2.46}$$

The logically consistent relation between the evolutionary success or failure, viewed either from the perspective of group **A** or from group **B**, reads:

$$\Phi^\star_{\mathbf{B}} := \underbrace{\Phi^\star(1, n, p_1, k^b, k^s)}_{\text{perspective of group } \mathbf{B}} \stackrel{!}{=} \underbrace{-\Phi^\star(1, n, 1-p_1, k^s, k^b)}_{\text{perspective of group } \mathbf{A}} =: -\Phi^\star_{\mathbf{A}} \tag{2.47}$$

If $\Phi^\star_{\mathbf{B}} > 0$, group **B** that initially consists of a single disadvantageous inequity averse agent, outperforms group **A** that has $n-1$ members of self-regarding agents. Consequently, the strategy $s^b = [m^b, k^b]$ spreads in the population. In contrast, if $\Phi^\star_{\mathbf{A}} > 0$, group **A** becomes predominant and strategy $s^a = [m^a, k^s]$ spreads in the population. The resulting condition for the disadvantageous inequity aversion trait to become dominant is defined by

$$\Phi^\star_{\mathbf{B}} > 0 \;\wedge\; \Phi^\star_{\mathbf{A}} < 0\,. \tag{2.48}$$

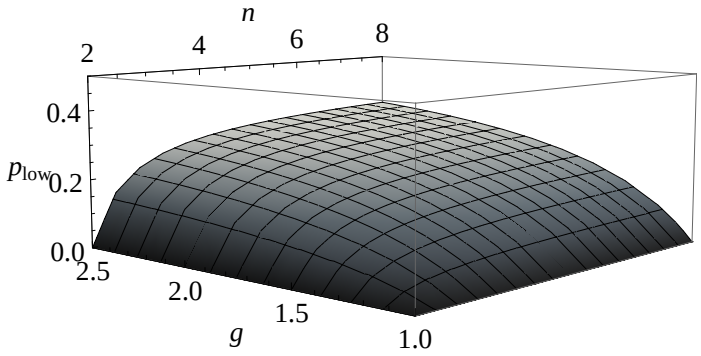


Figure 2.4: Minimum probability threshold p_1 given by expression (2.49), above which a single disadvantageous inequity averse agent can invade a population of selfish agents by deviating from the contribution of the selfish agents with $\Delta\hat{m} = p_1 \cdot (-\Delta m) + (1 - p_1) \cdot \Delta m$.

Reducing condition (2.48) over the set of reasonable parameter values with $\Delta m > 0$, $n \geq 2$, $r > 1$ and $0 < g < n$ reveals that $\Phi^{\star}_{\mathbf{B}}$ becomes positive if the probability p_1 falls into the range

$$p_{\text{low}} < p_1 \leq 1 \tag{2.49}$$

with

$$p_{\text{low}} = \frac{(n-2) \cdot (g-1)}{2 - 3 \cdot g + n \cdot (2 \cdot g - 1)} \ . \tag{2.50}$$

Figure 2.4 shows the surface defined by expression (2.49) for p_{low} as a function of n and r in the range $2 < n < 8$ and $1 < g < 2.5$. The domain above the surface corresponds to p_1 values for which a single disadvantageous inequity averse agent can invade a population of selfish agents by deviating from the contribution of the selfish agents. A scenario with a population consisting of 4 agents with 3 agents being self-regarding and one agent being disadvantageous inequity averse, playing a public goods game with a per capita return of 0.4 MUs per invested MU, i.e. $g = 1.6$, results in a $1 - p_{\text{low}} = 82\%$ chance for the single disadvantageous inequity averse agent to outperform at each period.

For all reasonable parameter values, $n > 2$ and $0 < g < n$, the lower bound p_{low} is always smaller than $\frac{1}{2}$. This means that the probability for the dis-

advantageous inequity averse agent to invade the population of selfish agents over time is always larger than one-half. The range of p_1 defined by equation (2.49) shows that, if the single disadvantageous inequity averse agent in group **B** deviates on average by a negative value, i.e. $p_1 > \frac{1}{2}$, from the contribution m_2 of the selfish agents (group **A**), she always wins since $\Phi^{\star}_{\mathbf{B}} > 0$.

Such a single agent can win even though she may be strongly out-numbered by the $n-1$ selfish agents who tend to defect, because the minimum required consumption c_{fix} per period forces the population to contribute on average at least an amount of

$$\frac{d \cdot m_1 + (n-d) \cdot m_2}{n} \approx \frac{c_{\text{fix}}}{g-1}$$

MUs in order not to go extinct.

On the other hand, if the single disadvantageous inequity averse agent contributes on average more than the group of self-regarding and selfish agents, it must hold that

$$\frac{g+n-2}{2-3 \cdot g + n(2 \cdot g - 1)} < \Delta \hat{m} \tag{2.51}$$

in order for that agent to have a larger P&L than the self-regarding agents of group **A**. Coming along with the condition $\Phi^{\star}_{\mathbf{B}} > 0$, the disadvantageous inequity averse agent in group **B** can be thought of as being more fertile than the self-regarding agents of group **A**, which results in $d(t+1)$ being larger than $d(t)$ over time. In addition, with an increasing number d of agents in group **B** and, consequently, a decreasing number $n-d$ of agents in group **A**, the lower limit for p_1 declines until it becomes zero for $d = \frac{n}{2}$. This means that, as soon as half of the total population consists of disadvantageous inequity averse agents, the self-regarding and selfish agents are doomed, as the probability for group **B** to take over the entire population becomes 1 independent of their contribution decisions.

In summary, for arbitrary initial conditions, we have established that disadvantageous inequity averse preferences and the corresponding ESS s^b have significantly more than 50% chance of spreading in the population. At large times and for finite populations, in the presence of a larger than 50% probability to grow their relative population ($1 - p_{\text{low}} > \frac{1}{2}$), the population of the disadvantageous inequity averse agents will with probability one reach half the

total population, at which point they invade with certainty the whole population due to their self-reinforcing advantage explained above. This can be summarized by the following set of inequalities:

$$\begin{aligned} & 1 - p_{\text{low}} > \frac{1}{2} \Rightarrow \Pr\left[\Phi^{\star}_{\mathbf{B}}(t) > 0\right] > \frac{1}{2} \\ & \Rightarrow \Pr\left[d(t+1) > d(t)\right] > \frac{1}{2} \Rightarrow \Pr\left[\Phi^{\star}_{\mathbf{B}}(t+1) > \Phi^{\star}_{\mathbf{B}}(t)\right] > \frac{1}{2} \\ & \Rightarrow \lim_{t\to\infty} \Pr\left[\Phi^{\star}_{\mathbf{B}}(t) > 0\right] = 1 \Rightarrow \lim_{t\to\infty} d(t) = n \ . \end{aligned} \tag{2.52}$$

In conclusion, our third main result can be summarized as follows:

Result 2.3: *On long enough time scales, disadvantageous inequity averse preferences always invade and dominate pure self-regarding and selfish preferences in an evolutionary system.*

2.5 Conclusion

Previous works on economic theories about fairness, altruistic punishment and cooperation in voluntary contribution situations have systematically underestimated the importance of evolutionary dynamics and in particular the role of natural selection for the emergence of prosocial behavior and fairness preferences. We have combined an evolutionary approach together with an expected utility model to identify and explain the mechanisms that account for the emergence of fairness preferences and altruistic punishment. In particular, we designed an expected utility model that allowed us to calculate an optimal strategy profile for the level of punishment in public goods games, depending on the fairness preferences of the agents in the population.

In particular, we considered two specific types of agents: (1) purely self-regarding and selfish acting agents and (2) agents who are disadvantageous inequity averse. We find that the evolutionary optimal strategy profile of disadvantageous inequity averse agents matches the behavior of subjects in the experiments and explains quantitatively the observed level of altruistic punishment without adjustable parameters. Our results imply that subjects show a strong predisposition for disadvantageous inequity aversion which, in

turn, seems to be the driving force behind the observed altruistic punishment behavior. Finally, we showed that disadvantageous inequity aversion is an evolutionary dominant and stable strategy when compared to the pure self-regarding behavior, in a heterogeneous population of agents. Our theory offers new predictions that are testable by running future experiments with different numbers of subjects, modified payoff levels or a varied efficiency of the punishment.

In conclusion, we believe that path-dependent evolutionary processes, together with the self-organizational aspects of individual utility maximization, provide an important explanatory basis for the emergence of cooperation, altruism and prosocial behavior in general. Future research on social preferences should take the time dimension and the evolutionary dependencies of many social system more carefully into account.

The results and findings presented in this chapter derive from an expected utility model that integrates an evolutionary perspective. By definition, the model is constructed based on certain simplifying assumptions in order to ensure its computability. In particular, evolutionary comparisons with respect to the fitness of individuals or the dominance of strategies are implemented using a simplistic two-person view. This approach is common practise in game theory and economics and is widely applied in this area of research. However, we stress that the inherent characteristics of evolutionary systems require a more sophisticated approach in order to fully understand the dynamics and underlying mechanisms. The following chapter 3 presents a numerical approach that takes the full complexity of evolutionary path dependencies, n-player interactions and mechanisms such as adaptation, selection, cross-over and mutation into account.

3. The co-evolution of fairness preferences and altruistic punishment

This chapter studies the co-evolutionary emergence of fairness preferences in the form of other-regarding behavior and its effect on the origination of altruistic punishment behavior using a numerical simulation model. Our approach closely combines empirical results from three public goods experiments with an evolutionary simulation model whose formulation borrows ideas from evolutionary biology, behavioral sciences and -economics as well as complex system science. As a principal result, we show that the evolution among interacting agents inevitably involves a built-in sense for fairness in the form of disadvantageous inequity aversion that emerges in the presence of effective selection pressure. The evolutionary dominance and stability of disadvantageous inequity aversion is demonstrated by enabling agents to co-evolve with different self- and other-regarding preferences in a competitive resource limited environment. Disadvantageous inequity aversion leads to the emergence of altruistic punishment behavior and quantitatively explains the level of punishment observed in contemporary lab experiments. Our findings corroborate, complement, and interlink the experimental and theoretical literature that has shown the importance of other-regarding behavior in various decision settings. This

chapter can be considered as the logical and consistent extension to the findings from chapter 2. Therefore, we increase the complexity of the evolutionary dynamics and the interactions by replacing the analytical approach in chapter 2 with the numerical simulation model presented in this chapter.

3.1 Introduction

Why do we show altruistic- and other-regarding behaviors? Why have we developed a sense for fairness? Is such behavior compatible with Darwin's principle of fitness maximization and/or with the economic axiom of rational decision making? Which evolutionary mechanisms dominate the evolution of our pro-sociality? With the genesis of more complex forms of life and organisms, evolution has been working on multiple scales, ranging from the level of genes and phenotypic traits to the emergence of norms, culture and social institutions. In continuation of the previous findings in chapter 2, this chapter aims at shedding further light on the puzzling behavior of pro-sociality. By allowing for more complex evolutionary dynamics using a numerical simulation model we attempt to analyze the roots of pro-sociality and the reciprocal effects at different scales of the evolutionary mechanisms. This chapter presents a transdisciplinary approach to explain the emergence of fairness preferences and altruistic punishment behavior, which is motivated by perspectives from biology, evolutionary psychology, sociology and economics.

There is evidence from a variety of studies that fairness preferences have emerged in hominids over hundreds and thousands of years, with roots in our genetic heritage as evidence from recent studies on primates and the genetic encoding of social behavior suggests (Brosnan and de Waal, 2003; Silk et al., 2005; Jensen, Hare, Call, and Tomasello, 2006; Jensen et al., 2007b; Hamlin et al., 2007; de Waal et al., 2008; Robinson et al., 2008; Takahashi, Shimomura, and Kumar, 2008; Fowler and Schreiber, 2008). The importance of our genetic heritage for the structural basis of our pro-sociality appears to be plausible: Our genes encode the essential protein structures that are required to build up our physical-, cognitive- and computational capabilities. These capabilities allow us e.g. to perceive others' behavior, to compare quantities and to interact either physically or by communication with our environment.

Furthermore, they build the fundamental basis that allows us to express, transmit and externalize our cumulative knowledge, our culture. Vice versa, our cultural evolution promotes those genes which are beneficial to the cultural evolution itself. Culture and genes thus appear to be subjected to more complex, co-evolutionary processes occurring over a spectrum of different time scales. Cultural evolution is shaped by biological conditions, while, simultaneously, genes are altered in response to the evolutionary forces induced by the cultural context. As a consequence, the perception of fairness and the reaction to unfair behavior as well as the individual's response to its social environment in general seem to be encoded both, in cultural norms and in genes (Boyd and Richerson, 1988; Laland, Smee, and Feldman, 2000; Gintis, 2003; Sinha, 2005; Jablonka and Lamb, 2007; Jasny, Kelner, and Pennisi, 2008; Efferson, Lalive, and Fehr, 2008).

On all scales, living things tend to organize and group together. Cells which consist in large parts of complex molecules such as proteins group into organisms. Proteins are transcriptions of one or multiple genes. Higher level organisms arrange themselves in groups and populations. Groups again often organize in societies. Throughout the hierarchical levels of biological and social disciplines, previously independent entities on a lower scale reassemble to an new unique and individual entity on a higher scale. As an ultimate result, the coordination and convergence of individual attitudes to common group behavior and the emergence of social norms as well as their enforcement by informal social sanctions are often observed in groups of animals and human societies (Homans, 1974; Coleman, 1998; Whiten, Horner, and de Waal, 2005; Bernhard, Fischbacher, and Fehr, 2006; Guererk, Irlenbusch, and Rockenbach, 2006). From small cliques to the social order in groups and tribes, all the way to the legal frameworks of countries, punishment is a widespread mechanism underlying the formation of social norms (Fehr, Fischbacher, and Gaechter, 2002; Fehr and Fischbacher, 2004; Henrich et al., 2006). In particular, altruistic punishment, i.e., the punishment of norm violators at one's own cost without personal benefit, is frequent in social dilemmas and is often used to explain the high level of cooperation between humans (Fehr and Gachter, 2000, 2002; Rockenbach and Milinski, 2006; Henrich et al., 2006; Herrmann, Thoni, and Gachter, 2008). From an evolutionary perspective, natural se-

lection should discriminate against altruists who incur costs to themselves in order to provide benefits to non-relatives and to strangers in one-shot interactions. Within Darwin's theory as well as in economic and game theoretic models, which rely on rational selfishness and the dominance of self-regarding preferences, such behaviors are puzzling, if not disrupting. This observation calls for the identification of the generative mechanism(s) underlying altruistic punishment.

Models of kin selection, inclusive fitness, reciprocity, network reciprocity, group-level and multi-level selection have been developed to explain the presence of pro-social behavior. Laboratory experiments and field studies suggest that egalitarian motives and other-regarding preferences, which relate a person's decision to her social environment, have a significant influence in social dilemmas, coordination and bargaining games (Fehr and Fischbacher, 2002; Fowler, Johnson, and Smirnov, 2005; Fehr, Bernhard, and Rockenbach, 2008; Tomasello and Warneken, 2008). As a result, psychological models of inequity aversion have been formulated that included descriptions of other-regarding preferences. These models are based on motivation functions that include relative income preferences, envy, inequality aversion and altruism (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness, Gary and Rabin, Matthew, 2002; Fehr and Schmidt, 2006). The quantitative comparison with empirical data often remains unsatisfactory as most models aim at explaining stylized facts rather than providing quantitative explanations of the underlying mechanisms. Although, while based on plausible assumptions, their evolutionary validation is not manifested: Can other-regarding preferences emerge, evolve and ultimately dominate pure self-regarding and selfish behavior? What are the consequences of other-regarding behavior for our social interactions? Can the presence of other-regarding preferences cause the emergence of altruistic feedback mechanisms such as costly punishment?

Experiments on public goods and social dilemma games provide convenient tools to study social preferences in well-defined scenarios under controlled conditions. As already presented in chapter 2, these experiments allow to study in details what controls the predisposition of humans to bear the costs associated with punishment of free riders, and how it may improve the welfare of the group. The observed behavior in the experiments can be interpreted as sam-

pling the statistically stationary characteristics of a cultural group of subjects which have evolved over a long time horizon. Their response to specific social dilemma situations are then revealed through the present-day experiments.

In particular, when provided with the opportunity to punish norm deviators at own costs, altruistic behavior is manifested (Fehr and Gachter, 2000, 2002; Decker, Stiehler, and Strobel, 2003; Masclet et al., 2003; Noussair and Tucker, 2005). Even in one-shot interactions in public good games in which reputation and reciprocal effects are absent, costly punishment, which at a first sight seems to be in contradiction with individual fitness maximization, natural selection and rational choice theory, is frequently observed (Fehr and Gachter, 2000, 2002; Anderson and Putterman, 2006; Fudenberg and Pathak, 2009). One should, however, keep in mind that other patterns of behaviors may have emerged in the presence of different norms, environmental conditions and genetic endowments. E.g. subjects from 15 diverse populations display various behavioral patterns when playing an ultimatum game (Henrich et al., 2006). The diversity of behavioral traits found in different human cultures may result from different evolutionary trajectories as well as from distinct relative influences of the cultural versus genetic heritages and a varying intensity of the selection pressure (Cason, Saijo, and Yamato, 2002; Henrich et al., 2006; Hil and Gurven, 2004).

The co-evolutionary dynamics and inter-dependencies of genes and cultural norms constitute our starting point to understand the properties of our prosocial behavior and our sense of fairness, as observed in lab experiments, field studies and, of course, in real life. To identify and fully understand the mechanisms underlying our prosocial behavior, we design an evolutionary simulation model that mimics the dynamics of individuals being exposed to a social dilemma situation. To verify our theoretical results, we compare them with observations previously obtained in three independently conducted lab experiments. As a most important result, we find that evolution favors a build-in predisposition for fairness concerns: In the presence of a sufficiently large selection pressure, individuals inevitably develop an aversion to unfairness. Secondly, the dislike of unfair situations - not to be confused with a preference for fairness in general - promotes altruistic behavior in the form of costly punishment that occurs even in one-shot interactions as frequently observed

in lab experiments. Thus, altruistic punishment is a consistent consequence of our conditional evolutionary predisposition to unfairness aversion.

In the following section, we will present our model, motivate, discuss and verify the obtained results and draw conclusions about the evolution of fairness preferences, altruistic punishment and moral behavior.

3.2 Method

We develop a simulation model consisting of synthetic agents that describes the long-term co-evolution of cultural norms and genes accounting for fairness preferences and altruistic punishment behavior in populations being exposed to a competitive voluntary contribution dilemma. Specifically, we compare our model with the results of three public goods game experiments conducted by Fehr/Gachter and Fudenberg/Pathak (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009). Our modeling strategy is to see the empirical observations in the experiments as a snapshot within a long-term evolutionary dynamics: on the short time scales of the experiments, the traits of the human players probed by the games can be considered fixed for each player. These traits might be encoded in the cultural context, in genes, or both.

Our model does not aim at simulating and explaining strategic short-term behavior of agents in social dilemmas, but instead mimics the culture-gene co-evolution that has occurred over tens of thousands of years. Aiming at two goals, we validate our model by comparing its results with the observed behavior in the experiments. In a first step, we quantitatively identify the underlying other-regarding preference relation that explains best the contemporary behavior. Here, we specifically look into a set of common assumptions made by researchers to account for fairness preferences and its observable consequences in the form of altruistic punishment behavior. Other-regarding preferences are expressed as inequality or inequity aversion. Initialized with different variants of these other-regarding preferences, the traits of our agents converge after long transients to statistically stable values, which are taken to describe the present-day characteristics of modern humans. In a second step, we verify that the identified preference relation which explains best the contemporary behavior is evolutionary stable and dominates the remaining

variants of self- and other-regarding preferences. We do this by allowing the other- and self-regarding preferences to co-evolve over time within a heterogeneous population. Our final goal is to reveal the ultimate mechanisms and the conditions under which agents develop spontaneously a propensity to "altruistically" punish, starting from an initial population of self-regarding and selfish-acting non-punishers.

The design of our model is inspired by three public goods game experiments with punishment conducted by Fehr/Gachter and Fudenberg/Pathak (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009). In these experiments, subjects[1] are arranged in groups of $n = 4$ persons and play a two stage game. At the beginning of each period, in stage one, subjects received an initial endowment of 20 monetary units (MUs). Thereafter, subjects could invest $m \in [0, 20]$ MUs to a common group project, which returned $g = 1.6$ MUs for each invested MU. The total return from the project was equally split and redistributed to all group members. Thus, the return per capita was $g/n = 0.4$. As long as $g/n < 1$, the game has a vivid social dilemma component, since it is rationally optimal not to cooperate, even though the group is better off if each member cooperates[2]. Thus the setup is susceptible to defection through material self-interest and we consider the subjects' investment as their level of cooperation.

In the second stage of the game, subjects were provided with the opportunity to punish other group members, after they had been informed about the individual contributions[3]. The use of punishment was associated with costs for both parties, in which each MU spent by a punisher led to $r_p = 3$ MUs taken from the punished subject (Fehr and Gachter, 2002; Fudenberg and Pathak, 2009)[4]. Experiments were played both in a partner treatment (Fehr

[1] Here undergraduate students from the Federal Institute of Technology (ETH) and the University of Zurich as well as subjects from the Boston area universities.

[2] If all agents contribute one MU (cooperate), they each obtain 1.6 MU. If only one does, the three others (free-riders) pocket 0.4 MU on top of their own uninvested MU while the single contributor is left with just 0.4 MU and thus takes a loss of 0.6 MU.

[3] In (Fudenberg and Pathak, 2009), subjects also played an unobserved treatment in which they learned the contributions of other group members not until the last period has been played. However, this variation in the design of the experiment did not lead to a significantly different level of observed punishment.

[4] In (Fehr and Gachter, 2000), the punisher paid approximately 2 MUs to take an additional 10% from the punished subject's period profit.

and Gachter, 2000), in which the group composition did not change across periods, and in a stranger treatment (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009). In the later, subjects were reassigned to new groups at each period using an anonymous random matching procedure and thus were only engaged in one-shot interactions during the entire runtime of the experiment. In total, the experiments were played for $T_1 = 10$ (Fehr and Gachter, 2000; Fudenberg and Pathak, 2009) and $T_2 = 6$ periods (Fehr and Gachter, 2002) respectively.

The data from Fehr/Gaechter and Fudenberg/Pathak as well as from several other public goods experiments (Decker et al., 2003; Masclet et al., 2003; Noussair and Tucker, 2005) show that people, if provided the opportunity, frequently punish defectors, even if this is costly to themselves and not immediately observable to others. In the case of repeated interactions, as in the partner treatment, such behavior might be explained by the "direct reciprocity" mechanism. What is more surprising is that subjects continue to punish at a cost to themselves even in one-shot interactions for which there is no feedback mechanism in action that would work e.g. by direct or indirect reciprocity. This behavior is referred to as "altruistic punishment" to emphasize the conflict with the behavior expected from purely rational agents. The question we address here is why humans behave in a way that seemingly contradicts individual fitness maximization and rational choice.

3.2.1 The computational model

We construct an evolutionary simulation model adapted from the design of the experiments in (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009) that consists of a population of agents who play a public goods game with punishment. In this model, agents are characterized by three traits. The first two traits characterize the agent's level of cooperation m and their propensity to punish k. The third trait q characterizes the agent's preferences for self- and other-regarding behavior, respectively. All traits can adapt and evolve over long periods according to generic evolutionary dynamics: adaptation, selection, crossover and mutation. In order to capture the possible evolution of the population, agents adapt and die when unfit. Newborn agents replace dead ones, with traits taken from the pool of the other surviving agents. The

adaptation and replication dynamics are described in detail in section 3.2.2 and 3.2.3, respectively.

A given simulation period t is decomposed into two sub-periods:

1. Cooperation: Each agent i chooses an amount of $m_i(t)$ MUs to contribute to the group project in period t. This value of $m_i(t)$ reflects the agent's intrinsic willingness to cooperate and thus is referred to as her level of cooperation. As in the experiments, each MU invested in the group project returns $g = 1.6$ MUs to the group. Combining all the contributions by all group members and splitting it equally leads to a per capita return given by equation (3.1).

$$r(t) = (g/n) \cdot \sum_{j=1}^{n} m_j(t) \tag{3.1}$$

This results in a first-stage profit-and-loss (P&L) of

$$s_i(t) = r(t) - m_i(t) = (g/n) \cdot \sum_{j=1}^{n} m_j(t) \; - \; m_i(t) \; , \tag{3.2}$$

for a given agent i, which is equal to the difference between the project return and its contribution in period t. The willingness to cooperate embodied in trait $m_i(t)$ evolves over time as a result of the experienced success and failures of agent i in period t. The adaptation and replication rules are described in detail in sections 3.2.2 and 3.2.3.

2. Punishment: Given the return from the group project $r(t)$ and the individual contributions of the agents, $\{m_j(t), j = 1, ..., n\}$, which are revealed to all, each agent may choose to punish other group members according to the rule defined by the equation (3.3) below. To choose the agents' decision rules on when and how much to punish, we are guided by figure 2.1 in chapter 2. Resulting from the data of three experiments, figure (2.1) shows the empirically reported average expenditure that a punisher incurs as a reaction to the negative or positive deviation of the punished opponent.

One can observe an approximate proportionality between the amount spent for punishing the lesser contributing agent by the greater contributing agent and

the pairwise difference $m_j(t) - m_i(t)$ of their contributions. The figure includes data from all three experiments (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009). In our model, this linear dependency, with threshold, is chosen to represent how an agent i decides to punish another agent j by spending an amount given by

$$p_{i\to j}(t) = \begin{cases} k_i(t) \cdot (m_i(t) - m_j(t)) & m_i(t) \geq m_j(t) \ , \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

The coefficient $k_i(t)$, which represents the propensity to punish, is the second trait that characterizes agent i at time t. It is allowed to vary from agent to agent and it evolves as a function of the successes and failures experienced by each agent, as explained in sections 3.2.2 and 3.2.3. Given that certain other-regarding preferences are active, we will show that evolution makes the punishment propensities $k_i(t)$ self-organize towards a value fitting remarkably well the empirical data, without the need for any adjustment.

As a result of being punished, the fitness of the punished agent j is reduced by the amount spent by agent i multiplied by the punishment efficiency factor r_p. As in the experiments, we fix the punishment efficiency factor to $r_p = 3$.[5]

The total P&L $\hat{s}_i(t)$ of an agent i over one period of her lifetime is thus the sum of three components: (i) her first stage P&L $s_i(t)$ from the group project (equation (3.2)), (ii) the MUs $\sum_{j\neq i} p_{i\to j}(t)$ spent to punish others and (iii) the punishments $r_p \sum_{j\neq i} p_{j\to i}(t)$ received from others, where $p_{i\to j}(t)$ and $p_{j\to i}(t)$ are given by (3.3):

$$\hat{s}_i(t) = s_i(t) - \sum_{j\neq i} p_{i\to j}(t) - r_p \sum_{j\neq i} p_{j\to i}(t) \ . \tag{3.4}$$

Equation 3.4 represents the second stage P&L of agent i in period t.

[5] In the first experiment of Fehr/Gachter (Fehr and Gachter, 2000), the punishment efficiency factor was determined based on the first stage payoff of the punished individual. However, it can be considered to be approximately equal to the factor 3 as in the remaining two experiments.

3.2.2 Adaptation Dynamics

Adaptation is an heritable phenotypic trait that affects the individual's fitness as a result of facing short- and long-term changes in the environment (Williams, 1996; Reeve and Sherman, 1993; Drickamer, Vessey, and Miekle, 1995). In the context of the analyzed public goods game, this translates into the underlying mechanism that makes subjects adjust their individual willingness to cooperate $m_i(t)$ and their propensity to punish $k_i(t)$ as a consequence of experienced environmental conditions.

It has been argued (Simon, 1982; Arthur, 1994; Holland, Holyoak, Nisbett, and Thagard, 1989; Gigerenzer and Selten, 2002) that humans (and our ancestors) are likely to use heuristics and inductive reasoning to make decisions. In particular, this means that humans tend to replace working hypotheses with new ones when the old ones cease to work. We adopt this bounded rational approach to define the adaptation mechanism that controls the dynamics of the propensity to punish and the level of cooperation.

The two traits $[m_i(t); k_i(t)]$, characterizing each agent i at a given period t, evolve with time according to standard evolutionary dynamics: adaptation, selection, crossover and mutation. While selection, crossover and mutation operate on the individual fitness level, i.e. are controlled by the birth-death process, adaptations are individually performed by each agent during its lifetime. We model this phenotypic expression using a third trait, $q_i(t)$. In contrast to $[m_i(t); k_i(t)]$, which are continuous measures, $q_i(t)$ represents a discrete indicator variable that corresponds to a specific boolean expression. The associated boolean expression translates into a specific adaptation condition that expresses a self- or other-regarding preference relation. We focus in particular on the set of inequality and inequity aversion preferences, which have been identified as important determinants in the human decision process and that of other species (Brosnan and de Waal, 2003; Brosnan, Talbot, Ahlgren, Lambeth, and Schapiro, 2010; Almas, Cappelen, Sorensen, and Tungodden, 2010; Tricomi, Rangel, Camerer, and O/'Doherty, 2010; Range et al., 2008; Fehr et al., 2008; Braeuer, Call, and Tomasello, 2006). If a particular condition becomes satisfied, an unbiased adaptation of $[m_i(t); k_i(t)]$ is triggered. This allows each agent to adapt $[m_i(t); k_i(t)]$, either solely based on the individually

experienced P&L values, or depending on the P&L and contributions of all group members.

Inequality aversion refers to the dislike of unequal profits, ignoring a potential inequality in the individually contributed efforts. In contrast, inequity aversion relates the personal profits directly to the personal efforts that has been contributed to the group project. The following six preference types represent the fundamental set of variants of inequality and inequity aversion preferences: (A) inequity averse, (B) inequality averse, (C) disadvantageous inequality averse, (D) disadvantageous inequity averse, (E) advantageous inequality averse and (F) advantageous inequity averse. "Disadvantageous" indicates that agents are only inequality/inequity averse if the inequality/inequity plays to their disadvantage, while "advantageous averse" agents do the opposite. In contrast, pure inequality or inequity averse agents dislike both situations in which they are discriminated against or are discriminating others. We as well analyze purely self-regarding and selfish-acting agents (G), i.e. agents who adapt their traits independently of the actions and the outcomes of other agents.

Figure 3.1 depicts schematically the possible variants of inequality and inequity aversion preferences introduced above. While inequity aversion (first row) is determined by a combinatorial condition relating the P&L to the performed effort (contribution) of an agent, inequality aversion (second row) is determined only by the agent's P&L value. Disliked regions of individual P&Ls (inequality aversion) or combinations of P&Ls and contributions (inequity aversion) are highlighted by boxes filled using the same pattern: E.g. inequity averse agents (first row, left column) dislike situations in which they contribute more than the average and their P&L is less than the average (combination indicated by 1) or vice versa (combination indicated by 2). The following list describes the set of analyzed phenotypic expression $\hat{q} \in Q$ in detail:

A: **Inequity averse agents**: such an agent i updates her cooperation level and her propensity to punish according to eq. (3.5) below, if...

...she has contributed less than (or equally) to her group fellows ($m_i(t) \leq \bar{m}(t)$), where the average $\bar{m}(t)$ is performed over the contributions of the other members of her group and, at the same time, has received a total P&L $\hat{s}_i(t)$ defined in (3.4) larger than (or equal) to the group average

$(\hat{s}_i(t) \geq \bar{s}(t))$, where the average $\bar{s}(t)$ is performed over the other group members...

...or she has contributed more than (or equally) to her group fellows $(m_i(t) \geq \bar{m}(t))$ and, at the same time, has received a total P&L less than (or equal) to the group average $(\hat{s}_i(t) \leq \bar{s}(t))$.

For inequity averse agents, the boolean expression is defined as $\hat{q}_A := [m_i(t) \leq \bar{m}(t) \wedge \hat{s}_i(t) > \bar{s}(t)] \vee [m_i(t) < \bar{m}(t) \wedge \hat{s}_i(t) \geq \bar{s}(t)] \vee [m_i(t) \geq \bar{m}(t) \wedge \hat{s}_i(t) < \bar{s}(t)] \vee [m_i(t) > \bar{m}(t) \wedge (\hat{s}_i(t) \leq \bar{s}(t)]$.

B: **inequality averse agents**: such an agent i updates her cooperation level and her propensity to punish if her P&L $\hat{s}_i(t)$ given by (3.4) is not within a specific tolerance range $[-l, +l]$ around the average P&L of the other members of her group, i.e. if $\hat{s}_i(t) < \bar{s}(t) - l$ or $\hat{s}_i(t) > \bar{s}(t) + l$. When this occurs, agent i updates her traits $[m_i(t); k_i(t)]$ according to equation (3.5). It is clear that inequality averse agents do not take the individually contributed efforts explicitly into account, in contrast with the inequity aversion agents (A).

For inequality averse agents, the boolean expression reads $\hat{q}_B := [\hat{s}_i(t) < \bar{s}(t) - l] \vee [\hat{s}_i(t) > \bar{s}(t) + l]$

We run multiple simulations initialized by different values for l as presented in the results section.

C: **disadvantageous inequity averse agents**: as for agents of type (A), disadvantageous inequity averse agents compare their P&L to their contributions, however they only dislike situations in which the inequity is detrimental to them. If an agent i has contributed equally or more than her fellows in the group $(m_i(t) \geq \bar{m}(t))$ and, at the same time, has received a total P&L $\hat{s}_i(t)$ defined in (3.4) smaller than or equal to the group average $(\hat{s}_i(t) \leq \bar{s}(t))$, then she updates her traits $[m_i(t); k_i(t)]$ according to eq. (3.5).

For disadvantageous inequity averse agents, the boolean expression is defined by $\hat{q}_C := [m_i(t) \geq \bar{m}(t) \wedge \hat{s}_i(t) < \bar{s}(t)] \vee [m_i(t) > \bar{m}(t) \wedge (\hat{s}_i(t) \leq \bar{s}(t)]$

D: **advantageous inequity averse agents**: these agents correspond to the antithesis of agents of type (C). If an agent i has contributed equally or less than her fellows in the group ($m_i(t) \geq \bar{m}(t)$) and, at the same time, has received a total P&L $\hat{s}_i(t)$ defined in (3.4) larger than or equal to the group average ($\hat{s}_i(t) \leq \bar{s}(t)$), then she updates her traits $[m_i(t); k_i(t)]$ according to eq. (3.5).

For advantageous inequity averse agents, the boolean expression is $\hat{q}_D := [m_i(t) \leq \bar{m}(t) \wedge \hat{s}_i(t) > \bar{s}(t)] \vee [m_i(t) < \bar{m}(t) \wedge \hat{s}_i(t) \geq \bar{s}(t)]$

E: **disadvantageous inequality averse agents**: these agents only dislike situations in which the inequality is to their disadvantage. An agent i updates her cooperation and her propensity to punish only if her P&L $\hat{s}_i(t)$ given by (3.4) is smaller than the average P&L of the other members of her group, i.e. $\hat{s}_i(t) < \bar{s}(t)$. When this occurs for an agent i, she updates her traits according to equation (3.5).

The corresponding boolean expression for disadvantageous inequality averse agents is $q_E := [\hat{s}_i(t) < \bar{s}(t) - l]$

F: **advantageous inequality averse agents**: these agents only dislike situations in which the inequality is to their advantage as opposed to setup (E). An agent i updates her cooperation and her propensity to punish only if her P&L $\hat{s}_i(t)$ given by (3.4) is larger than the average P&L of the other members of her group, i.e. $\hat{s}_i(t) > \bar{s}(t)$. When this occurs for an agent i, she updates her traits according to equation (3.5).

Advantageous inequality aversion is defined by the boolean expression $q_F := [\hat{s}_i(t) > \bar{s}(t) + l]$

G: **self-regarding agents**: such an agent updates her cooperation and propensity to punish if her P&L $\hat{s}_i(t)$ given by (3.4) turns out to be smaller than the P&L in the previous period $t - 1$.

Pure self-regarding and selfish behavior is defined by the boolean expression $\hat{q}_G := [\hat{s}_i(t) < \hat{s}_i(t-1)]$

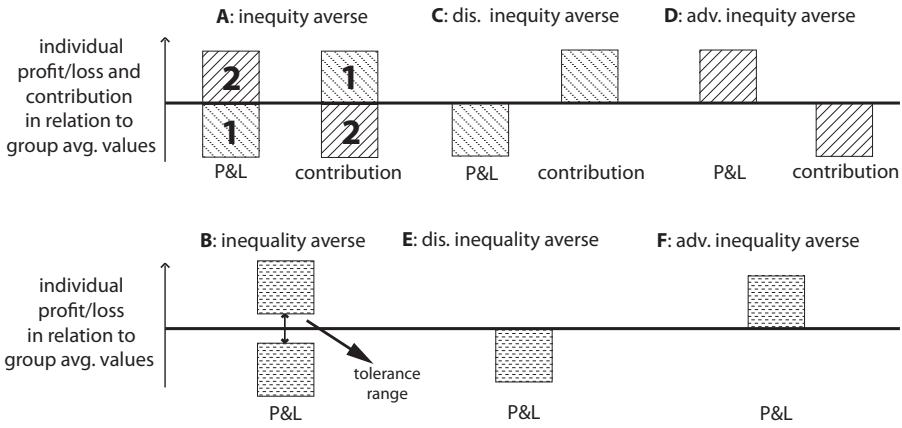


Figure 3.1: Scheme of the different possible variants of inequality and inequity aversion preferences introduced in the text.

In addition, each agent needs at least to consume an amount of $c_{\text{fix}} > 0$ per period in order to match the minimum costs of living, i.e. this value reflects the absolute lower limit required for survival. Thus agents in all dynamics additionally adapt their traits if their P&L is less than c_{fix} in avoidance of becoming extinct.

The update an agent performs if the predominant condition from the set of conditions $Q := [\hat{q}_A; \hat{q}_B; \hat{q}_C; \hat{q}_D; \hat{q}_E; \hat{q}_F; \hat{q}_G]$ applies consists in an unbiased random increment according to

$$\begin{aligned} m_i(t+1) &= m_i(t) + \epsilon_{[-0.005,0.005]} \\ &\text{and} \\ k_i(t+1) &= k_i(t) + \kappa_{[-0.005,0.005]} \ . \end{aligned} \tag{3.5}$$

The random variables ϵ and κ are uniformly distributed within the interval indicated in the subscript. Since contributions and punishment expenditures are non-negative, draws of ϵ and κ are truncated to avoid realizations that would lead to negative values of $m_i(t+1)$ and/or $k_i(t+1)$. Our results are robust to changes of the width of the interval, as long as it remains symmetric around zero.

3.2.3 Replicator Dynamics: Selection, crossover and mutation

In addition to the adaptation of the agents' traits $[m_i(t); k_i(t)]; q_i(t)$ described above, survival, viability and fertility selection occur by replacing underperforming agents. As we do not include a population dynamic, our model assumes a constant group size equal to n, with each death being followed by a corresponding birth. Selection occurs if an agent's wealth drops below zero, i.e. $w_i(t) < 0$. In this case, the agent dies and is replaced by a new one with different traits $[m_i(t+1), k_i(t+1), q_i(t+1)]$, determined by those of the surviving agents of the group. We tested our model with the following three variants of the selection mechanism:

S1: The first variant includes a survival-, viability- and fertility-selection mechanism. At each period, consumption absorbs an amount $c(t)$ of the agents' fitness. As mentioned above, each agents requires a minimum consumption of $c_{\text{fix}} > 0$ per period to satisfy the minimum expenses associated with the survival capability. Additionally, we implement a realistic driving force to select for successful traits. Traits, carried by agents that perform better than the group average over time, are selected using a consumption that is proportional to the average P&L of the group. In total, the consumption of an agent in the current period is determined by:

$$c(t) = \text{Max}[\frac{1}{n}\sum_i \hat{s}_i(t); c_{\text{fix}}] \ . \tag{3.6}$$

When an agent's fitness $w_i(t)$ drops below zero, the agent dies and is replaced.

S2: A second variant incorporates only a viability- and fertility-selection mechanism. Here, the death- and rebirth-event of an agent occurs randomly in time but the viability is proportional to the wealth. For each simulation period, the agent with the lowest wealth (fitness) in the group dies with a probability ξ and is subsequently replaced. We have varied ξ in a range $0.0001 < \xi < 0.01$ resulting in essentially the same output. To avoid negative values of wealth, which might occur as a result of

continuously realized negative P&L values, agents are endowed with an initial wealth $w_i(0) \gg 0$.

S3: In the third investigated variant, survival and viability selection is not active. Selection occurs based on a simple mechanism with non-overlapping generations, i.e. all agents have the same predefined lifespan. After one generation has reached its maximum age, the entire population of agents is replaced. The traits $[m_i(t+1), k_i(t+1), q_i(t+1)]$ of the new generation are inherited proportionally to the realized wealth of the agents in the previous generation. Agents receive an initial endowment with $w_i(0) \gg 0$ to prevent negative values of wealth (fitness).

Our results are robust to all three selection mechanisms (*S1*,*S2* and *S3*), i.e. all variants essentially create the same quantitative output. To be specific, without loss of generality, we obtained all results described in the following sections using replicator dynamic *S1*.

To simulate fertility selection, i.e. the fact that successful individuals produce more offsprings, we initialize reborn agents with traits inherited proportional to the fitness of the surviving agents. In this way, more successful traits are more strongly propagated than less successful ones. In detail, the process of crossover and mutation for the first two traits, $m_i(t+1)$ and $k_i(t+1)$, is determined as follows:

$$\begin{aligned} m_i(t+1) &= \bar{m}(t) + \epsilon_{[-0.005,0.005]} \\ &\text{and} \\ k_i(t+1) &= \bar{k}(t) + \kappa_{[-0.005,0.005]} \ . \end{aligned} \tag{3.7}$$

$\bar{m}(t)$ and $\bar{k}(t)$ correspond to the fitness weighted average values calculated over the surviving population and ϵ and κ reflect the individual mutation rates in the form of an unbiased uniformly distributed random increment over the interval indicated by the subscript. Again, draws of ϵ and κ are adjusted in a way to ensure the non-negativeness of the $m_i(t+1)$ and $k_i(t+1)$ values.

Crossover and mutation for the discrete indicator variable $q_i(t+1)$ occurs analogously. Our model implementation allows us to pairwise compare different self- and other-regarding preferences, i.e. a heterogeneous population

can co-evolve along two different adaptation rules $\hat{q}_x, \hat{q}_y \in Q$ across time. The value $q_i(t)$ determines which of the two conditions $\hat{q}_x, \hat{q}_y$ is active for agent i: if agent i's indicator value $q_i(t) = 0$, then she adapts $m_i(t)$ and $k_i(t)$ according to rule $\hat{q}_x$. In contrast, if $q_i(t) = 1$, adaptation occurs according to the second rule $\hat{q}_y$. Crossover and mutation operates on $q_i(t+1)$ as follows:

$$q_i(t+1) = \begin{cases} 1, & \text{if } \tau_{[0,1]} \leq \bar{q}(t) + \xi_{[-0.005,0.005]} \\ 0, & \text{if } \tau_{[0,1]} > \bar{q}(t) + \xi_{[-0.005,0.005]} \end{cases} \tag{3.8}$$

First, the fitness weighted average of the surviving population $\bar{q}(t)$ is calculated and mutated by a random variable ξ that is uniformly distributed in $[-0.005, 0.005]$. Second, a $[0,1]$-uniformly distributed random number τ is drawn and compared to the value $\breve{q}(t) := \bar{q}(t) + \xi_{[-0.005,0.005]}$. If τ is less than or equal to $\breve{q}(t)$, $q_i(t+1)$ becomes one and zero otherwise.

3.3 Results and Discussion

This section is structured in two parts. In the first part, we aimed at determining which superordinate regime ($q^\star \in Q$) of self- or other-regarding preferences has led our ancestors to develop traits promoting altruistic punishment behavior to a level that is observed in the experiments. To answer this question, we let the first two traits $[m_i(t); k_i(t)]$ co-evolve over time while keeping the third one, $q_i(t)$, fixed to one of the phenotypic traits defined in $Q := [q_A; q_B; q_C; q_D; q_E; q_F; q_G]$. In other words, we account only for a homogeneous population of agents that acts according to one specific self-/other-regarding behavior during each simulation run. Starting from an initial population of agents which displays no propensity to punish defectors, we will find the emergence of long-term stationary populations whose traits are interpreted to represent those probed by contemporary experiments, such as those of Fehr/Gachter or Fudenberg/Pathak.

The second part focuses on the co-evolutionary dynamics of different self- and other-regarding preferences embodied in the various conditions of the set $Q := [q_A; q_B; q_C; q_D; q_E; q_F; q_G]$. In particular, we are interested in identifying which variant $q^* \in Q$ is a dominant and robust trait in presence of a social

dilemma situation under evolutionary selection pressure. To do so, we analyze the evolutionary dynamics by letting all three traits of an agent, i.e. m, k and q co-evolve over time. Due to the design of our model, we always compare the co-evolutionary dynamics of two self- or other-regarding preferences pairwise, and we consider all possible combination in $q_x, q_y \in Q$ with $x \neq y$. Again starting from an initial population of agents with no disposition for other-regarding behavior and for altruistic punishment, we report below a remarkable consistency between (a) the evolutionary dominance of a variant of other-regarding behavior and (b) our findings from the first part of the analysis that focused on the empirical identification and validation. Additionally, we will learn that the findings presented in this chapter precisely agree with the results obtained by our analytical model presented in chapter 2.

The results presented below correspond to groups of $n = 4$ agents with a punishment efficiency factor of $r_p = 3$ and a per capita return per contributed MU of 0.4 ($g = 1.6$) as in the experiments. The minimum consumption value has been set to $c_{\text{fix}} = 0.0001$. We have run our simulation with thousands of independent groups over 10 million simulation periods.

3.3.1 The effect of other-regarding preferences on the evolution of altruistic punishment

To identify if, and if so which variant of self- or other-regarding preferences drives the propensity to punish to the level observed in the experiments, we test the single adaptation conditions defined in $Q := [q_A, q_B, q_C, q_D, q_E, q_F, q_G]$. In each given simulation, we use only homogeneous populations, that is, we group only agents of the same type and thus fix $q_i(t)$ to one specific phenotypic trait $q_x \in Q$. In this setup, the characteristics of each agent (i) thus evolve based on only two traits $[m_i(t); k_i(t)]$, her level of cooperation and her propensity to punish, that are subjected to evolutionary forces.

Each simulation has been initialized with all agents being uncooperative non-punishers, i.e., $k_i(0) = 0$ and $m_i(0) = 0$ for all i's. At the beginning of the simulation (time $t = 0$), each agent starts with $w_i(0) = 0$ MUs, which represents its fitness. After a long transient, we observe that the median value of the group's propensity to punish k_i evolves to different stationary levels or

exhibit non-stationary behaviors, depending on which adaptation condition (q_A,q_B,q_C,q_D,q_E,q_F or q_G) is active. We take the median of the individual group member values as a proxy representing the common converged behavior characterizing the population, as it is more robust to outliers than the mean value and reflects better the central tendency, i.e. the common behavior of a population of agents.

Figure 3.2 compares the evolution of the median of the propensities to punish obtained from our simulation for the six adaptation dynamics (A to F) with the median value calculated from the Fehr/Gachter's and Fudenberg/-Pathak empirical data (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009). The propensities to punish in the experiment have been inferred as follows. Knowing the contributions $m_i > m_j$ of two subjects i and j and the punishment level $p_{i\to j}$ of subject i on subject j, the propensity to punish characterizing subject i is determined by

$$k_i = -\frac{p_{i\to j}}{m_j - m_i} \ . \tag{3.9}$$

Applying this recipe to all pairs of subjects in a given group, we obtain a measure of propensities to punish per group. Sampling all groups and all periods, we calculate the median of all k_i values as shown in figure 3.2 (continuous horizontal line). Figure 3.3 additionally shows a magnification of figure 3.2 for adaptation dynamics C and D including their 20/80 quantiles.

Figures 3.2 and 3.3 reveal that purely self-regarding and selfish-acting agents adapting their traits according to dynamics (G) remain weak-punishers as shown in figure 3.2. In contrast, for agents endowed with inequality or inequity aversion preferences (adaptation conditions A to F), different stationary and non-converging states of the propensity to punish emerge spontaneously, each with different characteristics.

We now state our first main result:

Result 3.1: *For all adaptation rules (A to F), it holds that altruistic punishment has emerged endogenously as a trait in a competitive social dilemma scenario that is subject to evolutionary selection pressure.*

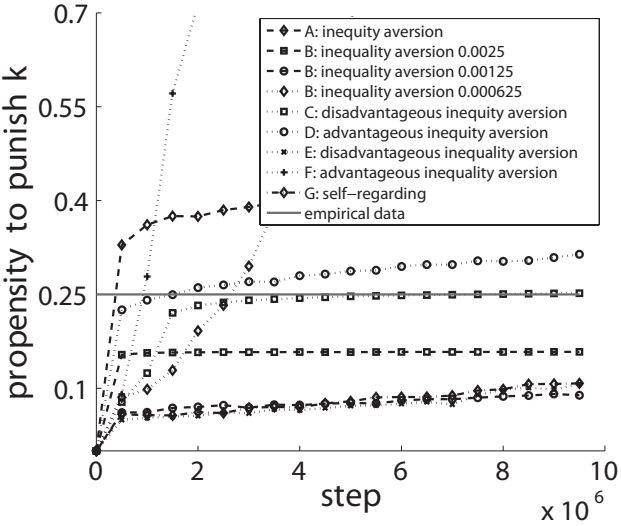


Figure 3.2: Evolution of the propensity to punish as a function of time. The values correspond to the population's median of the individual k_i values as a function of time for the seven different adaptation dynamics (A to G). The values for each adaptation dynamic result from 800 system realizations with a total of 3200 agents. The empirical median value calculated from all three experiments of Fehr/Gachter's and Fudenberg/Pathak (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009) is shown as the continuous horizontal line. For adaptation dynamic (B), the plot shows the obtained median values for all tolerance range parameters $l \in 0.0025; 0.00125; 0.000625$. The parameters of our simulation are: $n = 4, g = 1.6, r_p = 3, c_{\text{fix}} = 0.0001$.

In detail, we find that, for self-regarding and selfish-acting agents (dynamics G), the level of punishment that evolved remains too small to explain the empirical results of Fehr/Gachter and Fudenberg/Pathak. For the inequality averse population (B), we find that, for a set of reasonable values of the tolerance range parameter l, the empirical distribution can not be reproduced. Figure 3.2 shows the median value of the propensity to punish for adaptation dynamics B with the following values of the tolerance range parameter $l \in [0; 0.0025; 0.00125; 0.000625]$. While a large tolerance range causes altruistic punishment to remain weak, a narrow tolerance range results in continuously increasing and thus non-stationary levels of punishment. For

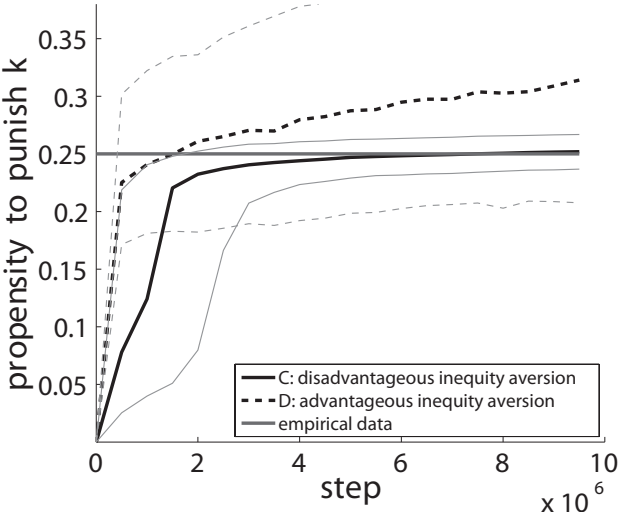


Figure 3.3: Magnification of figure 3.2 for adaptation dynamics C and D including their 20/80 quantiles (thin continuous grey line (C) and thin dotted grey line (D)). The horizontal continuous line corresponds to the median value of the empirically observed propensities to punish.

inequity- and altruistic inequity averse agents (dynamics A and D) as well as for disadvantageous inequality- and altruistic inequality averse agents (dynamics E and F), we find levels of altruistic punishments that far exceed the empirical evidence. We find that the adaptation dynamics C (disadvantageous inequity averse agents) causes the values k_i of the propensity to punish to converge towards the empirically observed norm. The quantitative comparison with the Fehr/Gachter and Fudenberg/Pathak experiments supports the hypothesis that human subjects are well-described as being disadvantageous inequity averse (dynamics C), corroborating and complementing previous evidence (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Fehr and Schmidt, 2006). A Mann-Whitney test does not reject the equality of the median values between the results obtained by the adaptation dynamics C and the empirical data observed in the experiments with a p-value of 0.943. For all other adaptation dynamics, the equality of the obtained median propensity to punish and the experimental value is clearly rejected. Also, the 95% confidence

interval for the sample median, either for each of the three experiment data sets independently, in pairs or as a whole does not allow us to reject adaptation dynamic C, whereas all other dynamics (A,B,D-G) can be rejected. The propensities k_i to punish exhibits a median around $k^\star \simeq 0.25$, which means that most punishers spend an amount approximately equal to one-fourth of the experienced differences in contributions in the given setup with 4 players. Note that the value of the median around $k^\star \simeq 0.25$ is close to the slope of the straight line fitting the empirical data shown in figure 2.1. In chapter 2 the value $k^\star \simeq 0.25$ has also been identified analytically as a Nash equilibrium strategy resulting from the maximization of the evolutionary expected utility problem with disadvantageous inequity aversion preferences. Given the simplicity of our model and of its underlying assumptions, it is striking to find such detailed quantitative agreement for one of our dynamics. This immediately raises the question of the generating underlying mechanisms that control these dynamics.

It is important to stress that the competitive evolutionary environment with its distinct selection pressure has no build-in mechanism that a priori favors the emergence of altruistic behavior such as the costly punishment of defectors. In order to understand how altruistic traits are selected in our simulation model, we analyze the evolution of the individual realized fitness- and P&L-values across time. Additionally, we inspect the micro behavior of the adaptation conditions A-G on a per step level to understand why and when agents adapt their traits $m_i(t)$ and $k_i(t)$. Figure 3.4 shows the evolution of a population of disadvantageous inequity averse agents (adaptation dynamics C). The figure reveals that the preference for disadvantageous inequity aversion together with the evolutionary dynamics, in form of survival/viability and fertility selection, is responsible for the emergence of altruistic punishment behavior in our model. The interplay of the evolutionary selection- and the individual adaptation-processes causes the propensity to punish k to evolve to a level that matches the empirical observations. Remarkably, a symmetric inequity aversion, i.e. an aversion for disadvantageous and advantageous inequity, is not needed as a condition to let altruistic punishment emerge.

We now state our second main result:

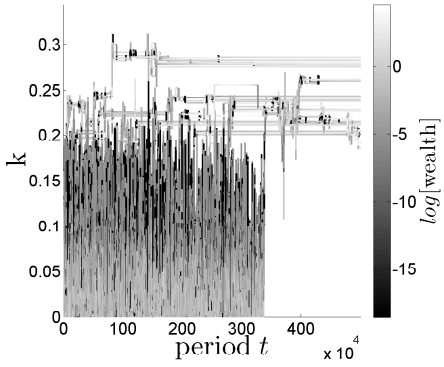
Result 3.2: *Disadvantageous inequity aversion is sufficient to explain the spontaneous emergence of altruistic punishment, with a median level of the propensity to punish that precisely match empirical data.*

Figure 3.4(a) shows the average group fitness of the agents across time on a logarithmic scale[6]. This plot reveals the existence of two evolutionary attraction points $k = 0$ and $k = 0.25$, which are identified by two discrete horizontal ranges around $k = 0.25$ and $k = 0$ for which the fitness takes the largest values (brighter shape of grey). Both evolutionary equilibria are separated by a range of values $0.125 < k < 0.2$, in which the evolution is unstable (darker grey shape). This evolutionary barrier can also be observed in figure 3.4(b), showing the higher rate of deaths/births in the range of $0.125 < k < 0.2$ indicated by a brighter shape of grey. Note that this evolutionary barrier matches the identified bifurcation hurdle k^+ defined in equation (2.27) of chapter 2 and can also be observed as a pivot value in figure 2.2 of chapter 2.

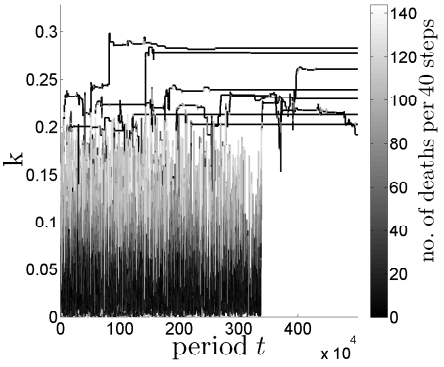
Figure 3.4(c) depicts the value of the boolean condition $\hat{q}_C$ on a group level across time, i.e. it quantifies whether all 4 agents per group are satisfied with their realized P&L and the ratio of their contributions in a way that $\hat{q}_C$ becomes *false* (the agents are "happy"). If this applies, no adaptation is performed by the agents. For values of $k < 0.125$, this is clearly not the case, causing $k_i(t)$ and $m_i(t)$ to continuously evolve. In addition, figure 3.4(d) reveals that the agents' viability condition, i.e. P&L-consumption≥ 0, is only constantly satisfied for levels of the propensity to punish $k > 0.2$ while it continuously alternates between positive and negative values below this boundary level.

As described above, fertility selection occurs by replacing dead agents with newborns whose traits are taken proportional to the wealth of the surviving group members. This results in k-values that are associated with a higher fitness to dominate and to spread in the population as a function of time in the presence of an ongoing deaths/births process. Figure 3.4(a) shows that more and more agents with $k \simeq 0.25$ start to dominate the heredity transmission

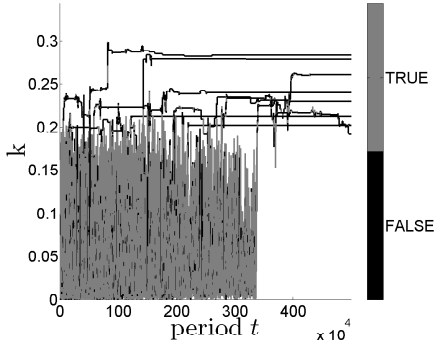
[6] We use a logarithmic scale as it better highlights the wealth dynamics across time.



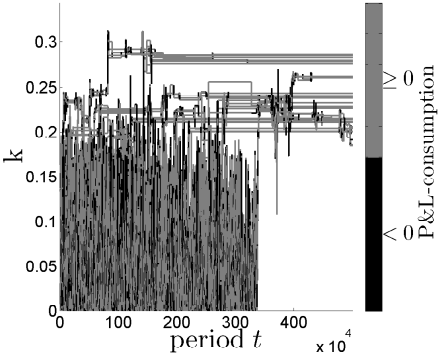
(a) log[wealth]



(b) avg. number of deaths per group in 100 steps



(c) $\hat{q}_C$ =true/false per group



(d) $\hat{s}_i(t) - c(t)$ is ≥ 0 or < 0

Figure 3.4: Evolution of the propensity to punish k (y-axis) over 5 million time steps (x-axis) (sample taken every 100 steps) resulting from 8 system realizations with a total of 32 agents in 8 groups. The shade of grey indicates (a) the evolution of the agents' fitness/wealth (upper left in log scale), (b) the number of deaths per group within 100 simulation steps (upper right), (c) if the disadvantageous inequity aversion condition $\hat{q}_C$ is true or false for a given group (lower left) and (d) the positiveness/negativeness of the difference between P&L minus the consumption in each period for each of the 32 agents (lower right).

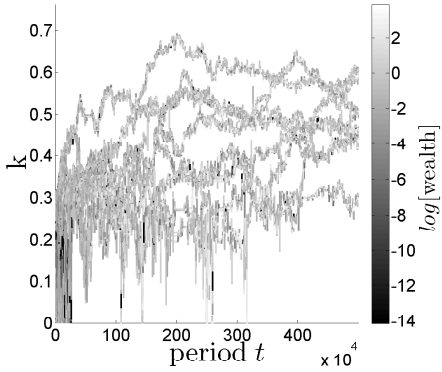
mechanisms, i.e. they spread their propensity to punish ($k \simeq 0.25$) much more than those with $k \ll 0.25$. This is because their fitness is higher and, at the same time, the deaths of agents with $k \ll 0.25$ occur more frequently. This becomes visible in figure 3.4(a) in the form of an increasing brighter shape of grey along the time line for realizations corresponding to a $k \simeq 0.25$, while those with $k \ll 0.25$ remain at a lower fitness level and disappear by-and-by. The identified level of the propensity to punish at $k \simeq 0.25$ is consistent with the findings obtained by our analytical framework in chapter 2. The optimal propensity to punish k^{ieq} that is defined in equation (2.39) matches exactly the value of $k \simeq 0.25$ for the specific game setup with $n = 4$ agents and a punishment efficiency of $r = 3$ as can be seen in the analysis of section 2.3.3.

In summary, we observe the co-evolution of three processes:

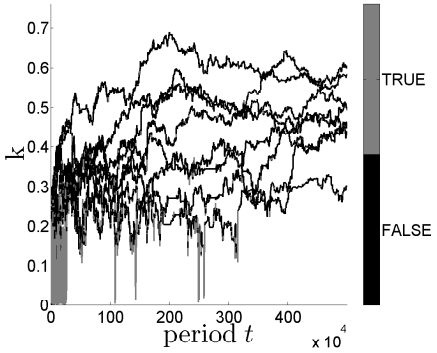
- Aversion to disadvantageous inequity makes agents adapt and explore values of their propensity to punish at levels $k > 0.125$.
- This leads them into a evolutionary unstable state associated with the range $0.125 < k < 0.2$.
- Subsequently, the evolutionary dynamics in the form of selection, cross-over and mutation, makes agents converge towards an equilibrium of their propensity to punish at a value around $k \simeq 0.25$.

This equilibrium is shaped by the two main conditions, i.e. the aversion to situation of disadvantageous inequity and evolutionary viability condition P&L-consumption≥ 0. The two conditions can be fulfilled simultaneously only for $k \approx 0.25$.

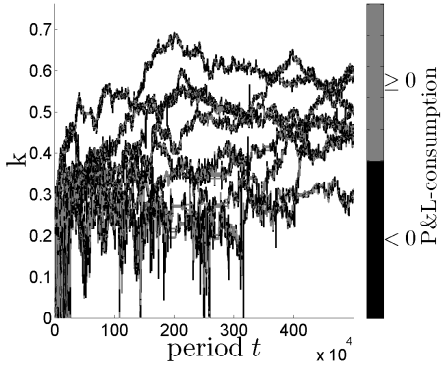
Figures 3.5 to 3.10 present an overview of the micro-dynamics of the remaining self- and other-regarding preferences (dynamics A,B,D-G). The three subplots show the evolution of the propensity to punish k (y-axis) over 5 million time steps (x-axis) (sample shown every 100 steps) resulting from 8 system realizations with a total of 32 agents in 8 groups. In subplot (a), the shade of grey indicates the evolution of the wealth. Subplot (b) depicts, if the other-regarding preferences condition $\hat{q}_{A,B,D,E,F,G}$ is true or false for a given group.



(a) log[wealth]

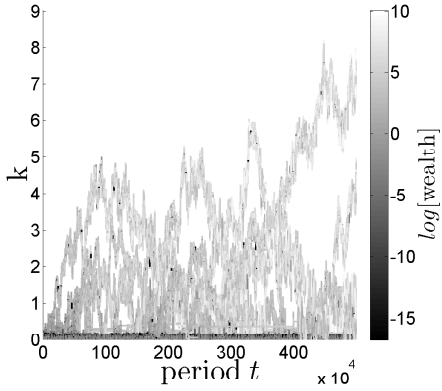


(b) $\hat{q}_A$ =true/false per group

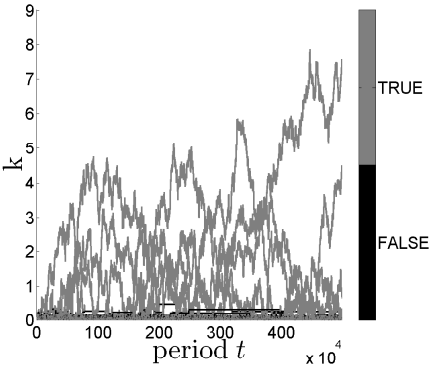


(c) $\hat{s}_i(t) - c(t)$ is ≥ 0 or < 0

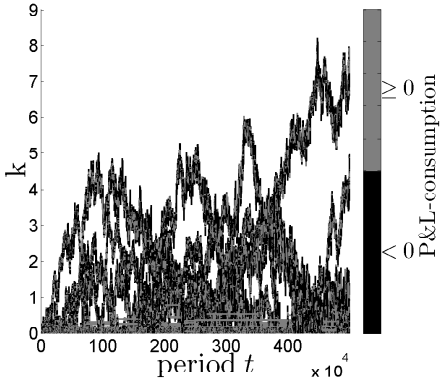
Figure 3.5: Dynamics A - inequity aversion



(a) log[wealth]

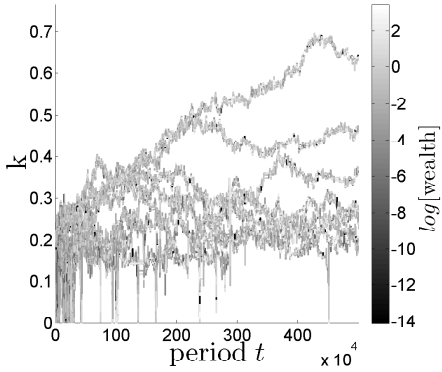


(b) $\hat{q}_B$ =true/false per group

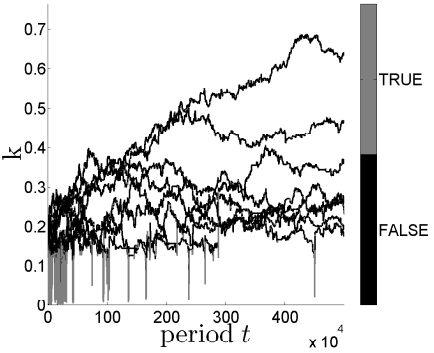


(c) $\hat{s}_i(t) - c(t)$ is ≥ 0 or < 0

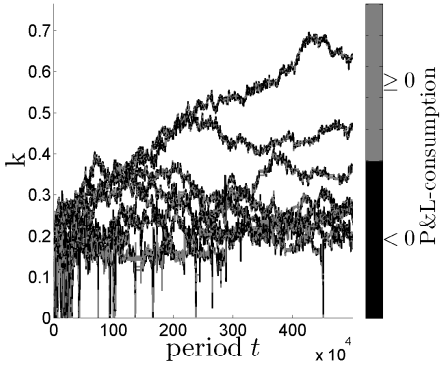
Figure 3.6: Dynamics B - inequality aversion with $l = 0.000625$



(a) log[wealth]

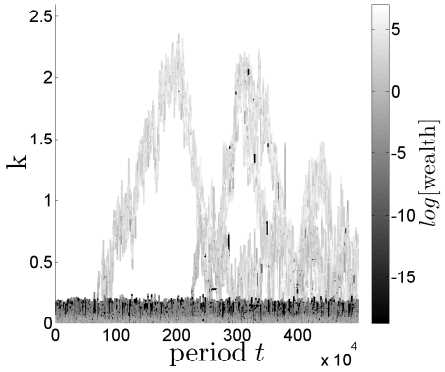


(b) $\hat{q}_D$ =true/false per group

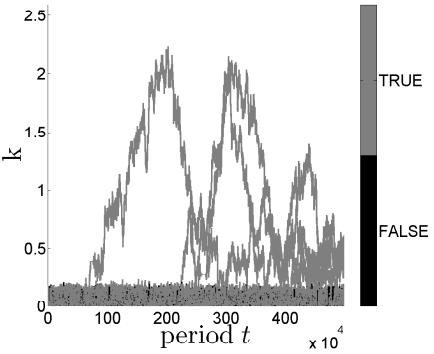


(c) $\hat{s}_i(t) - c(t)$ is ≥ 0 or < 0

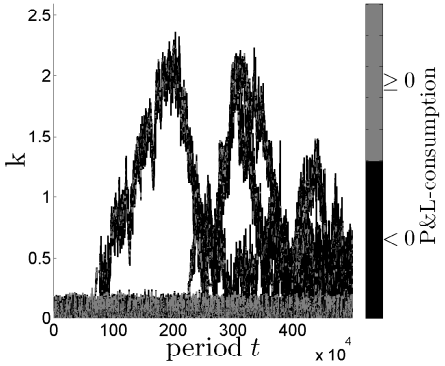
Figure 3.7: Dynamics D - advantageous inequity aversion



(a) log[wealth]

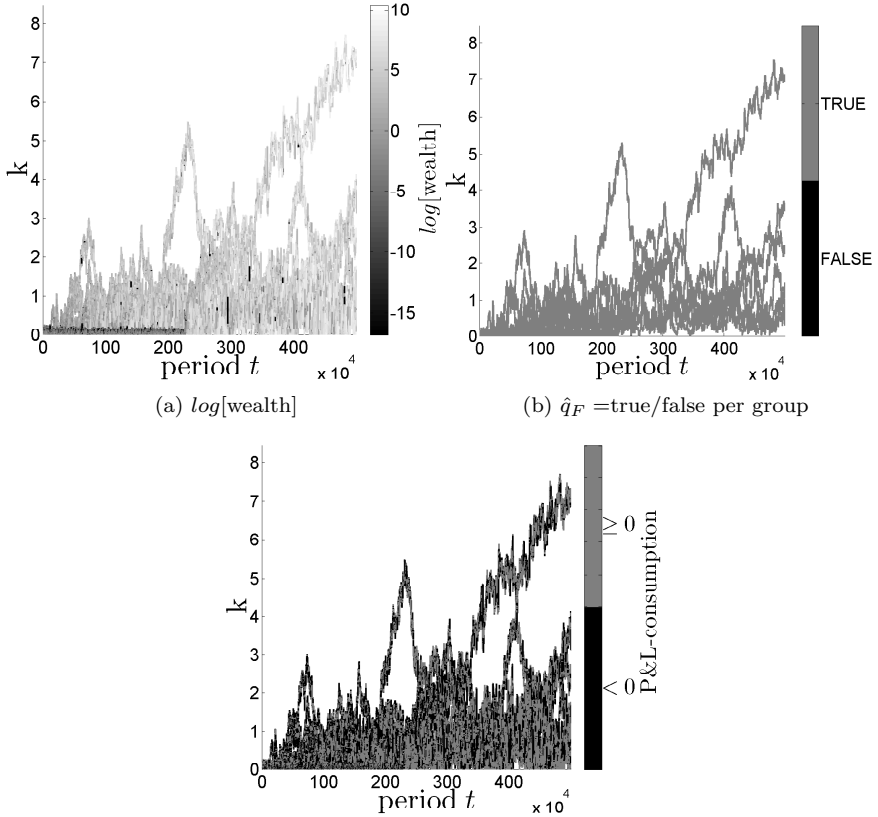


(b) $\hat{q}_E$ =true/false per group



(c) $\hat{s}_i(t) - c(t)$ is ≥ 0 or < 0

Figure 3.8: Dynamics E - disadvantageous inequality aversion

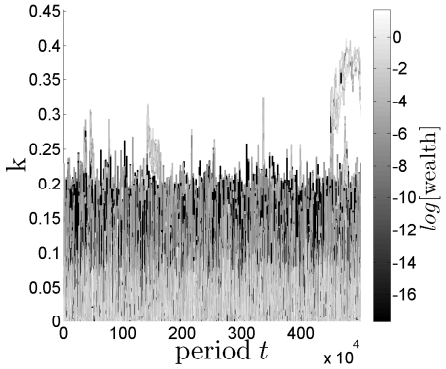


(a) log[wealth]

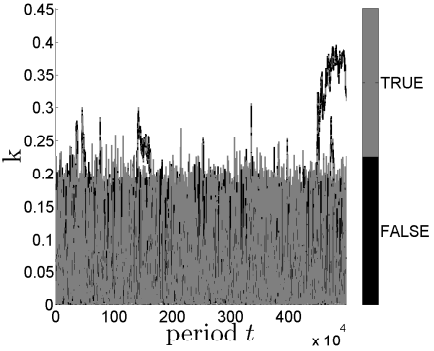
(b) $\hat{q}_F$ =true/false per group

(c) $\hat{s}_i(t) - c(t)$ is ≥ 0 or < 0

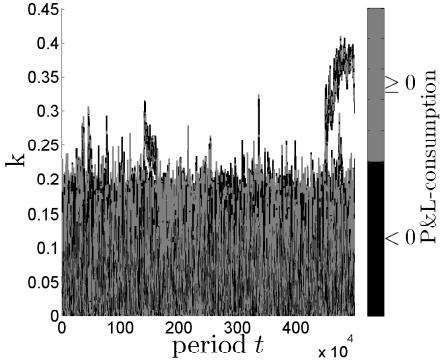
Figure 3.9: Dynamics F - advantageous inequality aversion



(a) log[wealth]



(b) $\hat{q}_G$ =true/false per group



(c) $\hat{s}_i(t) - c(t)$ is ≥ 0 or < 0

Figure 3.10: Dynamics G - self-regarding agents

Subplot (c) shows the positiveness/negativeness of the difference between P&L and consumption in each period and for each of the 32 agents.

Figure 3.2 has shown that altruistic punishment emerges not only in the presence of disadvantageous inequity aversion but also in the presence of the other variants of other-regarding preferences (dynamics A,B, D-F). However, a closer look on the micro-behavior data presented in the figures 3.5-3.10 reveals that the evolutionary dynamics A,B and D-F make agents not to converge to a stable stationary propensity to punish.

Figures 3.5 and 3.7 reveal that preferences of symmetric inequity aversion and advantageous inequity aversion (dynamics A and D) make agents to quickly explore values $k > 0.125$. In contrast to disadvantageous inequity aversion (dynamic C), the conditions $\hat{q}_A$ and $\hat{q}_D$ can not permanently be resolved to *false* for $k > 0.125$. In addition, there exists no unique equilibrium with respect to the fitness, for values of k larger than 0.2 as shown in figures 3.5(a) and 3.7(a). This causes adaptation and evolutionary selection to operate continuously. As a consequence, the populations continue to evolve without achieving a stable evolutionary state.

Altruistic punishment also originates in the three analyzed variants of inequality aversion (dynamics B, E and F). Figures 3.6, 3.8 and 3.9 suggest that viability and fertility selection operate in the opposite direction to the inequality aversion preferences, keeping agents away from achieving a potential stable state. While $\hat{q}_B$ can only become *false* for $k < 0.2$, agents with $k \gg 0.2$ outperform those with a smaller propensity to punish as indicated by brighter shades of grey for $k \gg 0.2$ in figures 3.6(a), 3.8(a) and 3.9(a). This leads to an evolutionary dynamic with no statistically stationary behavior and thus results in a heterogeneous population of agents with respect to k.

Purely self-regarding and selfish-acting agents (dynamic G) do not evolve a significant level of propensity to punish. Figure 3.10(a) reveals the existence of a single attraction point $k = 0$ indicated by brighter grey tones towards this value that lasts for the entire simulation. The purely self-regarding and selfish adaptation condition $\hat{q}_G$ does not allow agents to achieve an evolutionary stable state in the range of $0 < k < 0.2$, as can be observed in figure 3.10(b). As with the inequality aversion preferences, evolutionary selection, with its attraction

point at $k = 0$, works in the opposite direction to the adaptation condition G. This results in a population of agents that stray in an evolutionary non-stable range of $0 < k < 0.2$.

We find that the agents have an average lifetime of $\sim$ 160 periods with a median value of $\sim$ 90 periods. Therefore, a typical simulation run allows the occurrence of tens of thousands generations [7].

3.3.2 The co-evolution of self- and other-regarding preferences

The results obtained in the previous section in combination with the findings of section 2.3 in chapter 2 suggest that the punishment behavior of subjects observed in the experiments is driven by an aversion against disadvantageous inequity. Consequently, this raises the question if the identified adaptation dynamic C (disadvantageous inequity aversion) is an evolutionary stable and dominant phenotypic trait that emerges and prevails in an competitive resource-limited environment together with other variants of self- and other-regarding preferences. A first investigation of the evolutionary dominance of disadvantageous inequity aversion has been presented in the section 2.4 of chapter 2. Using the evolutionary simulation model the dominance of adaptation dynamic C can be verified by allowing agents with an aversion against disadvantageous inequitable outcomes to co-evolve along with other agents that act based on one of the remaining adaptation conditions (A,B,D-G) in our model.

In the following, we run our model with a population that consists of members who are either disadvantageous inequity averse or have the phenotypic trait of one of the other self- or other-regarding preferences. In this way, we can compare the reciprocal effects of the co-evolutionary dynamics from each variant A,B,D-G against disadvantageous inequity aversion preferences C. This results in 6 pairwise comparisons. In the beginning of each simulation run, the model is initialized with a preliminary homogeneous population of agents which are *not* disadvantageous inequity averse but moreover act according to one of the dynamics defined by A,B,D-G. The evolutionary updates of the

[7] These lifetimes correspond to a population of selfish inequity avers agents (C).

two competing adaptation traits is performed as described in section 3.2.3, i.e. $q_i(t)$ alternates between $q_i(t) = 0$ and $q_i(t) = 1$ according to the results of selection, crossover and mutation.

Running our simulation, we observe that the population of agents becomes always dominated by disadvantageous inequity aversion preferences, independent of which competing variant of self- or other-regarding adaptation dynamics has been seeded at step $t = 0$.

To further understand why we observe this behavior, for each of the six pairwise comparison, we plot in figures 3.11 to 3.16:

(a) the fraction of disadvantageous inequity averse agents compared to the whole population,

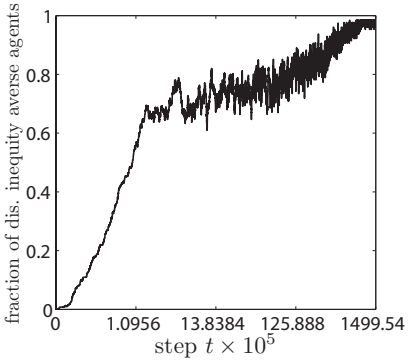
(b) the average wealth per agents in each phenotypic trait class,

(c) the relative frequencies of $Z = \hat{s_i}(t) - c(t)$, i.e. the P&L minus the consumption, for periods in which groups were heterogeneous, i.e. agents with both phenotypic traits were present in the group,

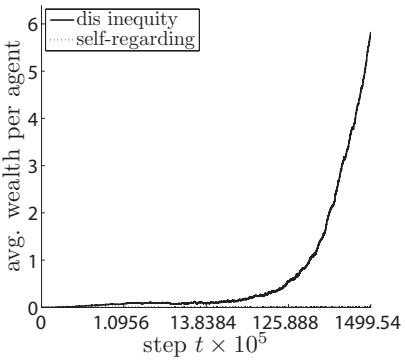
(d) the fraction of the total wealth taken by each phenotypic trait class and

(e) the average age at death for each phenotypic trait class.

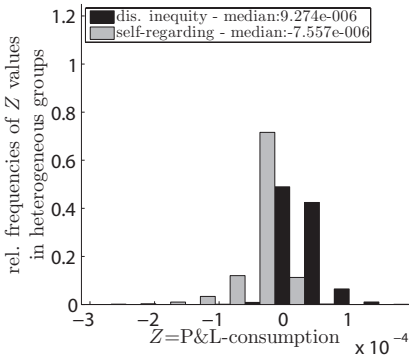
The resulting set of a total of 6 pairwise comparisons are depicted in figures 3.11-3.16, each of them showing the 5 subplots described above. Time steps (x-axis) are indicated in a non linear scale with a total of 10000 y-value samples taken over the whole simulation steps. The results correspond to 128 system realizations with a total population of 512 agents in 128 groups. The plots show nicely the impact of survival, viability and fertility selection on the population of agents. The indicated metrics in the different subplots conclusively demonstrate how disadvantageous inequity aversion always ends up dominating the population.



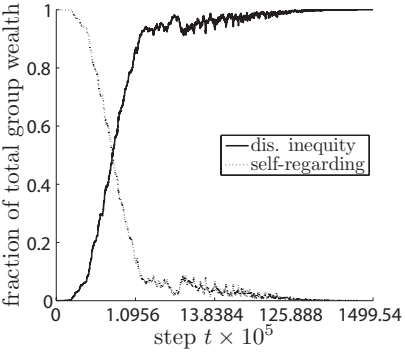
(a) fraction of disadvantageous inequity averse agents in the population



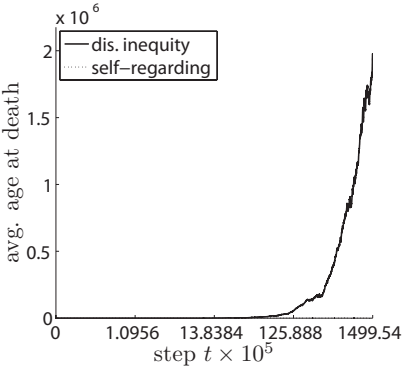
(b) average wealth per agent



(c) distribution of $\hat{s}_i(t) - c(t)$ values for steps t with heterogeneous groups

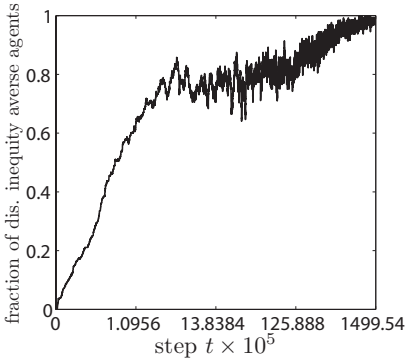


(d) fraction of the total population wealth

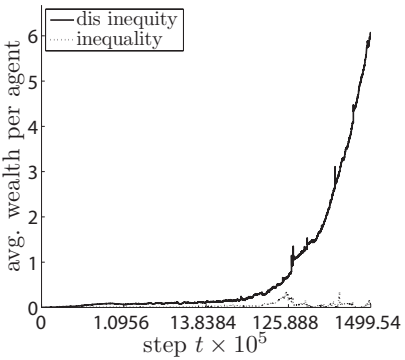


(e) average age of agents at death

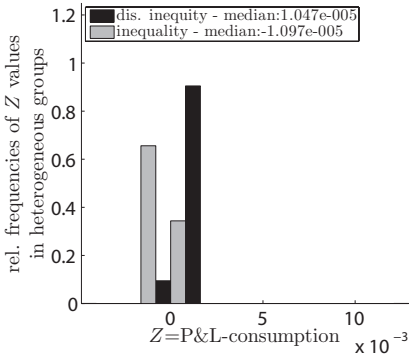
Figure 3.11: disadvantageous inequity aversion (C) vs. self-regarding (G)



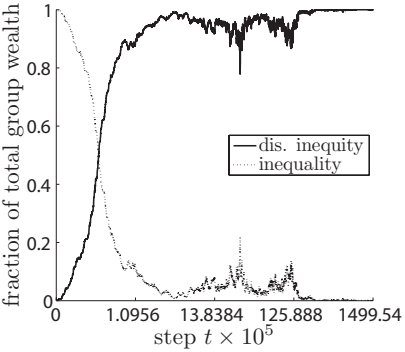
(a) fraction of disadvantageous inequity averse agents in the population



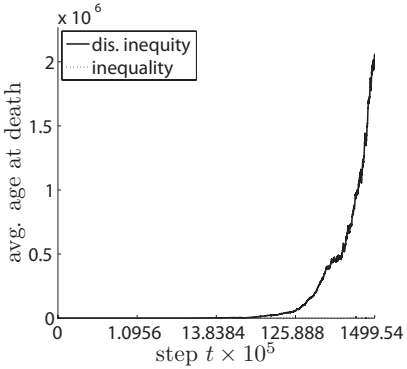
(b) average wealth per agent



(c) distribution of $\hat{s}_i(t) - c(t)$ values for steps t with heterogeneous groups

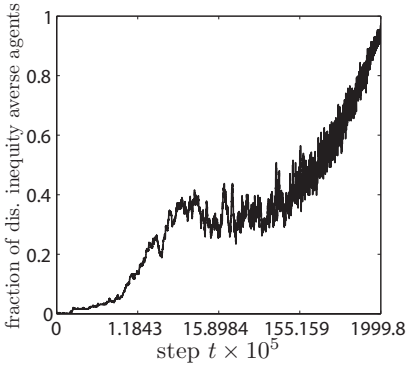


(d) fraction of the total population wealth

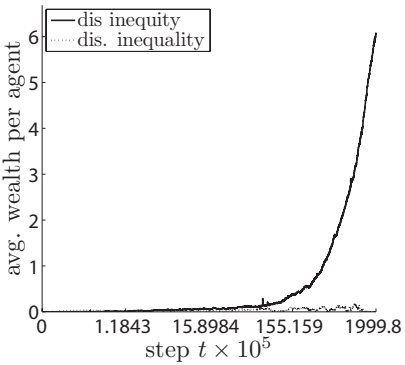


(e) average age of agents at death

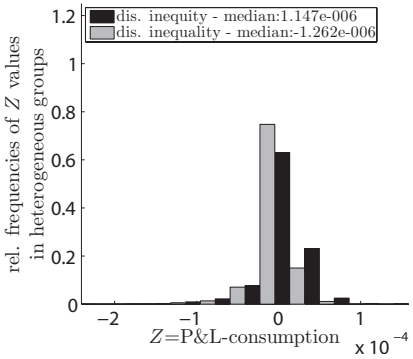
Figure 3.12: dis. inequity aversion (C) vs. inequality aversion (B)



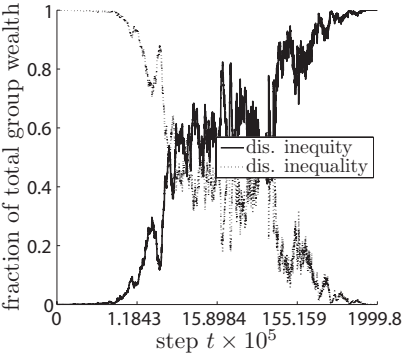
(a) fraction of disadvantageous inequity averse agents in the population



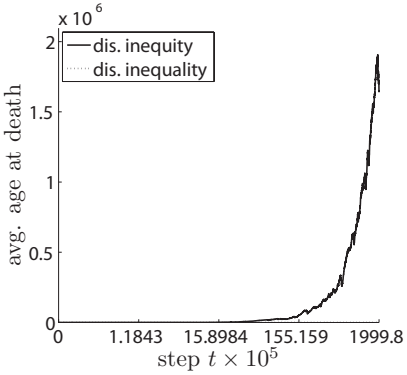
(b) average wealth per agent



(c) distribution of $\hat{s}_i(t) - c(t)$ values for steps t with heterogeneous groups

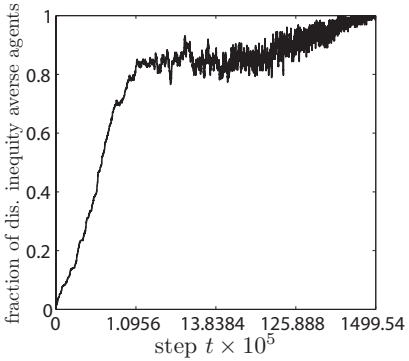


(d) fraction of the total population wealth

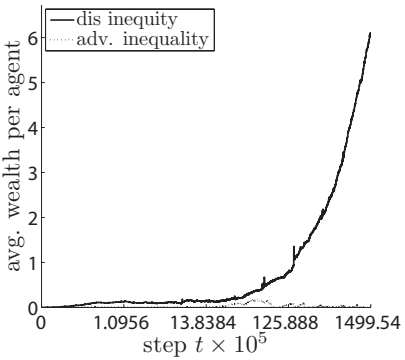


(e) average age of agents at death

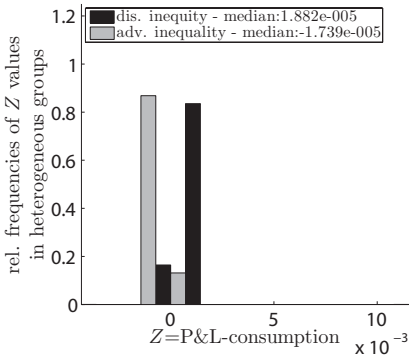
Figure 3.13: dis. inequity aversion (C) vs. dis. inequality aversion (E)



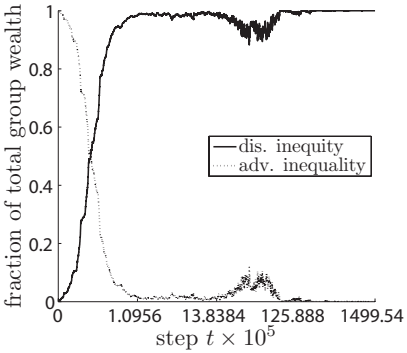
(a) fraction of disadvantageous inequity averse agents in the population



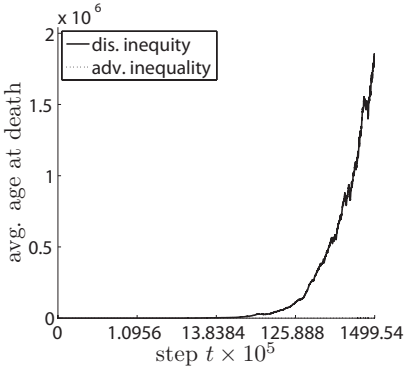
(b) average wealth per agent



(c) distribution of $\hat{s}_i(t) - c(t)$ values for steps t with heterogeneous groups

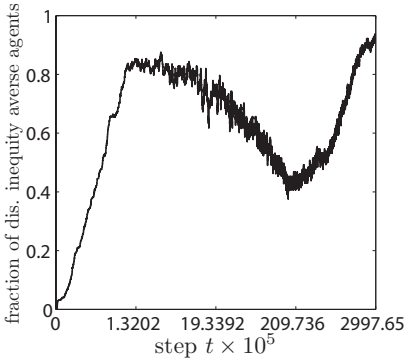


(d) fraction of the total population wealth

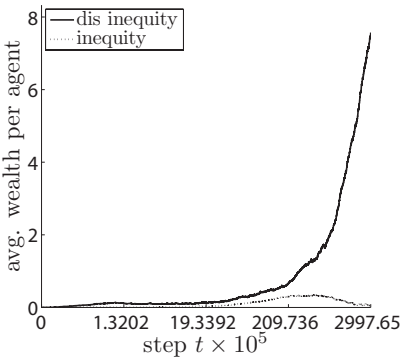


(e) average age of agents at death

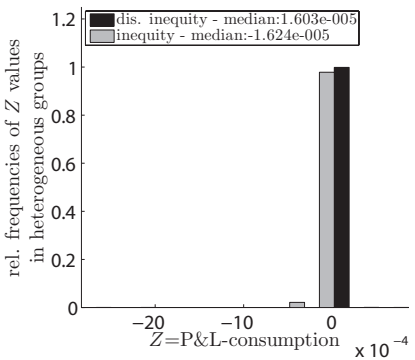
Figure 3.14: dis. inequity aversion (C) vs. adv. inequality aversion (F)



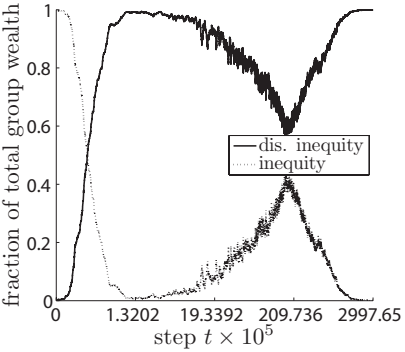
(a) fraction of disadvantageous inequity averse agents in the population



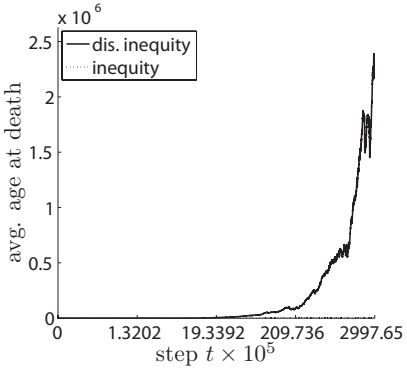
(b) average wealth per agent



(c) distribution of $\hat{s}_i(t) - c(t)$ values for steps t with heterogeneous groups

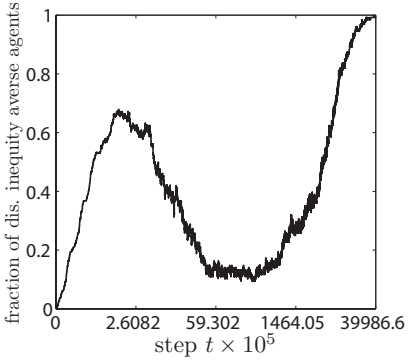


(d) fraction of the total population wealth

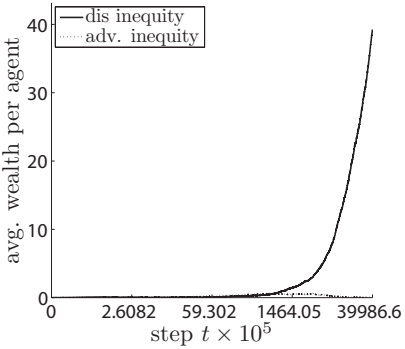


(e) average age of agents at death

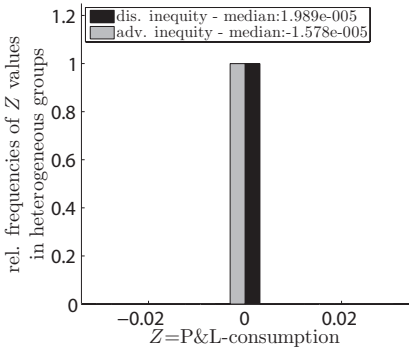
Figure 3.15: dis. inequity aversion (C) vs. inequity aversion (A)



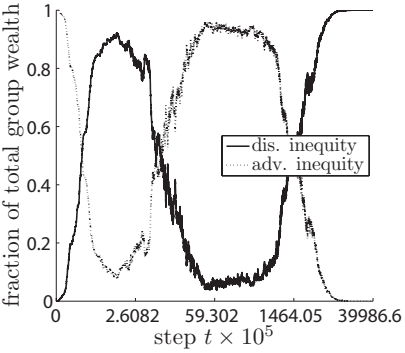
(a) fraction of disadvantageous inequity averse agents in the population



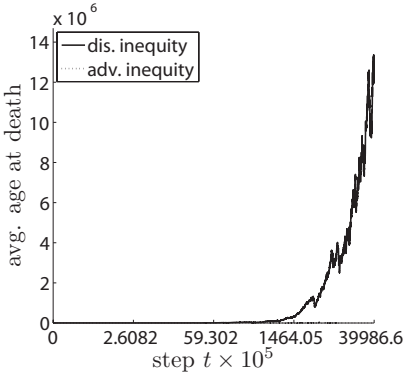
(b) average wealth per agent



(c) distribution of $\hat{s}_i(t) - c(t)$ values for steps t with heterogeneous groups



(d) fraction of the total population wealth



(e) average age of agents at death

Figure 3.16: dis. inequity aversion (C) vs. adv. inequity aversion (D)

Figure 3.11(a) shows the evolution of the number of disadvantageous inequity averse agents as a fraction of the total number of agents in the population across time. The impact of fertility selection is depicted in figure 3.11(b) with disadvantageous inequity averse agents being able to maintain on average a higher wealth value - also due to the longer lifetimes. Consequently, they are better able to promote their traits in the population. Figure 3.11(c) shows that, in periods where agents of both phenotypic traits are present, those acting based on disadvantageous inequity aversion clearly outperform self-regarding and selfish-acting agents on the short run. This is indicated by a right-shifted distribution (positive values of P&L-consumption) of the disadvantageous inequity averse agents compared to the left-shifted distribution of those being purely self-regarding and selfish-acting. Disadvantageous inequity averse agents do perform better here because they are less volatile in their adaptations as shown in figure 3.4(c) compared to the fluctuating behavior of self regarding agents shown in figure 3.10(b). In this way, agents with adaptation dynamic C suffer from less losses as a result of differences in contributions and punishments, respectively. Additionally, we provide the median value of the two distributions printed in the plot's legend. The fraction of wealth of disadvantageous inequity averse agents compared to the total wealth of the population starts to dominate as can be seen in figure 3.11. This result indicates that disadvantageous inequity averse agents typically invade and take over groups that are heterogeneous with respect to the phenotypic trait $\hat{q}$. The effect of survival selection is shown in figure 3.11(e). Groups of disadvantageous inequity averse agents are much more stable and are characterized by, on average, longer lifetimes with correspondingly a lower number of deaths. This makes them being less exposed to cross-over and mutations than compared to purely self-regarding and selfish-acting agents.

Essentially the same results and lines of argumentation hold for the remaining 5 comparisons, ie. inequity aversion (A) vs. disadvantageous inequity aversion (C), inequality aversion (B) vs. (C), advantageous inequity aversion (D) vs. (C), disadvantageous inequality aversion (E) vs. (C) and advantageous inequality aversion (E) vs. (C) as shown in figures 3.12 to 3.16.

We now state our third main result:

Result 3: *The three effects together (1-higher average wealth, 2-smaller volatility in their adaptation, 3-longer lifetimes) lead to the emergence of disadvantageous inequity aversion and its prepotency compared to the 6 self- and other-regarding preferences listed above.*

These findings together with those reported in the previous section and in chapter 2 suggest that disadvantageous inequity aversion does not only describes best the punishment behavior observed in lab experiments but moreover is also consistent and coherent with evolutionary dynamics in a competitive, resource limited environment. It seems that evolution inevitably pushes towards the development of a sense for fairness (disadvantageous inequity aversion) in population of adaptive and evolving interacting agents. This likely shapes the contemporary behavior of subjects and provides an explanation for the altruistic behavior observed in modern experiments in the form of the altruistic punishment of defectors.

3.4 Conclusion

This chapter studied the evolution of fairness preferences in the form of other-regarding behavior and its effect on the origination of altruistic punishment behavior. For this, empirical results from three public goods experiments has been combined together with an evolutionary simulation model. The model borrows ideas from evolutionary biology, behavioral sciences and -economics as well as complex system science.

Our first principal result is that, in a evolutionary-competitive resource-limited environment, altruistic punishment behavior can spontaneously emerge in a population of agents who are initially non-punishers, if other-regarding preferences are present. We have shown how this derives from an evolutionary process with adaptation, selection, crossover and mutation for different variants of inequality or inequity aversion.

Our second main result is the identification of disadvantageous inequity aversion as the most relevant underlying mechanism to explain the emergence and

the degree of altruistic punishment observed in public goods experiments. The results has been obtained by combining empirical data with an evolutionary simulation model in an innovative way. Our findings substantiate and extend the results obtained by an analytical utility framework presented in chapter 2. Our simulation model is able to reproduce quantitatively, without adjustable parameters, the experimental results concerning the level of punishment behavior. This result is of particular importance to substantiate the assumptions made by researchers in order to describe realistic behavior within the framework of rational choice: Humans exhibit other-regarding, and in particular, disadvantageous inequity aversion preferences in their decision process when facing public goods dilemmas with punishment opportunity.

As a third main result, we have demonstrated that disadvantageous inequity aversion is an evolutionary stable preference which dominates pure self-regarding and selfish behavior and also all other analyzed variants of inequity- and inequality aversion in a competitive resource-limited environment. This numerically calculated result supports, substantiates and extends the findings presented in chapter 2. We showed that standard evolutionary dynamics indeed have a built-in affinity to promote other-regarding behavior. This results from the fact that individuals so-to-speak hold each other mutually in bay to first ensure their own survival and second to preferably promote their own genetic and cultural heritage. Other-regarding behavior in the form of disadvantageous inequity aversion is often interpreted as a sense for fairness that serves to explain altruism. However, we find that disadvantageous inequity aversion and altruistic punishment, respectively, are just natural evolutionary consequences in the presence of competitive selection pressure.

The next chapter examines the effect of punishment behavior on the evolution of cooperation in public goods games. In addition, various mechanisms of group and multi-level selection are analyzed and discussed with respect to their contribution to the emergence of cooperative behavior in social dilemmas and competitive, resource-limited environments.

4. The effect of punishment on cooperation

In contrast to the punishment behavior that was discussed in the previous chapters, this chapter focuses on the interaction and the mutual effects of cooperation and punishment behavior in public goods game experiments. In the first part of this chapter, we identify and discuss behavioral patterns of cooperation and defection observed in a pool of subjects from three previously conducted lab experiments (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009). Therefore, a modification of the analytical utility framework that was introduced in chapter 2 is used to quantify the subjects' preferences either to cooperate or to defect. This provides us with a probability distribution of the subjects' intrinsic motivation to cooperate or to defect. In the second part of this chapter, we analyze and discuss the effect of (altruistic) punishment on the level of cooperation in public goods games.

Our previous findings presented in chapters 2 and 3 revealed that an aversion against disadvantageous inequitable outcomes causes altruistic punishment behavior to emerge to a level that precisely matches observations recorded in the three lab experiments. In addition, it was shown that disadvantageous inequity aversion is an evolutionary dominant and stable preference. In order to better understand the effect that the propensity to punishment has on the evolution of the cooperation level, we analyze the micro-level behavior of subjects from

three public good game with punishment (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009). This reveals that altruistic punishment promotes cooperation only among individuals who repeatedly interact with each other. In contrast, punishment among strangers, i.e. in one-shot interactions, only serves to maintain a preexisting level of cooperation. Given the widespread high levels of cooperative behavior which are present today in almost all areas of human life, punishment alone does not provide a sufficient explanatory solution to the puzzle of the evolution of cooperation. In particular, we find that punishment causes a convergence, i.e. a consolidation, of the behavior of agents and thus requires an additional countervailing mechanism that maintains heterogeneity in a population in order to keep the system evolving. Relevant mechanisms that account for the maintenance of heterogeneity in a population are subsequently discussed and analyzed in chapter 5.

4.1 Introduction

While some studies argue that punishment, and in particular altruistic punishment, accounts for the evolutionary emergence of cooperation (Boyd and Richerson, 1992; Fehr and Gachter, 2000, 2002; Masclet et al., 2003; Noussair and Tucker, 2005; Guererk et al., 2006; Nikiforakis and Normann, 2008; Herrmann et al., 2008; Gachter et al., 2008; Egas and Riedl, 2008), other studies conclude differently and argue that punishment can only sustain cooperative behavior and does not explain its evolutionary origin (Dreber, Rand, Fudenberg, and Nowak, 2008; Fudenberg and Pathak, 2009; Wu, Zhang, Zhou, He, Zheng, Cressman, and Tao, 2009; Boyd, Gintis, and Bowles, 2010; Mathew and Boyd, 2011). Various other articles combined the opportunity to punish with additional mechanisms, such as the possibility to abstain from voluntary actions or to switch between punishment and no-punishment treatments (Hauert, De Monte, Hofbauer, and Sigmund, 2002; Brandt, Hauert, and Sigmund, 2006; Hauert, Traulsen, De Silva, Nowak, and Sigmund, 2008). A further group of studies investigates the differences between peer and pool punishment scenarios (Sigmund et al., 2010; Marlowe, Berbesque, Barrett, Bolyanatz, Gurven, and Tracer, 2011; Baldassarri and Grossman, 2011) or addresses questions regarding the negative effects that punishment might cause on the evolution of the population welfare (Jensen, 2010; Holmas, Kjerstad, Luras, and Straume,

2010). Many of these studies apply methods from evolutionary game theory or provide insights using empirical investigations such as laboratory and field experiments. Apparently, there is no existing consensus in the literature about the role, function and importance of punishment for the evolution of cooperation. We add our findings from the previous chapters to this debate and put them into perspective with respect to the effect of punishment behavior on the level of cooperation in public goods games.

In the previous two chapters, we discussed the evolution of fairness preferences and the evolution of altruistic punishment behavior that is observed in lab experiments. This allowed us to reveal that fairness preferences in the form of disadvantageous inequity aversion (i) explains the altruistic punishment observed in the lab experiments and (ii) that an aversion against disadvantageous inequitable outcomes is a stable and dominant evolutionary strategy. In the first part of this chapter, we apply the analytical framework presented in chapter 2 and extend it in order to reveal the individual cooperation preferences of subjects in the three public goods games experiments. In particular, we provide a descriptive model of cooperation preferences by characterizing subjects by means of their intrinsic willingness to cooperate or to defect and subsequently classify subjects based on a continuous scale ranging from pure defection to unconditional cooperation. In the second part of this chapter, we focus on and analyze micro-level data from three public goods games (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009) experiments in order to understand the effect of punishment on the short-term dynamics of cooperative behavior. Additionally, we use the evolutionary simulation model introduced in chapter 3 to study the effect of punishment in an evolutionary competitive environment and to further identify its role, function and importance for the evolution of cooperation.

4.2 Cooperation preferences among subjects in experiments

The following section builds on the utility model introduced in the first chapter of this thesis and extends it with the objective to reveal the distribution of cooperation preferences in the subject pool. The *cooperation preference* of

a subject is defined by the ratio between the own performed effort and the performed effort of the other subjects in the reference group. In this way, subjects can be characterized by their intentional willingness to cooperate or by their purposeful defection. Previous work on the classification of cooperator types introduced specific experiments to measure and to classify subjects into a distinct set of categories, such as conditional cooperators, free-rider and altruists (Fischbacher, 2001; Houser and Kurzban, 2003; Bardsley and Moffatt, 2007; Herrmann and Thoeni, 2009; Rustagi, Engel, and Kosfeld, 2010). These experiments mainly base on choice/preference ranking tasks in the form of "*given the others provide an effort of X, which effort level Y do you choose?*" (Fischbacher, 2001). Our method complements these approaches by using the observed punishment behavior of individuals as an indirect benchmark for the level of disappointment about the opponents' behavior. The punishment reactions in three previously conducted public goods experiments (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009) are used to estimate the players' first-order beliefs about the contributions of their group fellows. Based on this estimation, we construct the distribution of cooperation and defection preferences and are able to reveal that the majority of subjects are imperfect conditional cooperators.

4.2.1 Theoretical framework of cooperation preferences

We start our analysis using the same evolutionary expected utility model that was introduced in chapter 3. Agents are arranged in groups with size n and play a public goods game with punishment. Each agent i invests an amount of m_i MU into the public good. The public good yields a per capita return of $\frac{g}{n}$ monetary units (MU) per invested MU. If $\frac{g}{n} < 1$ the public goods game has a social dilemma component that provides an incentive to defect and to exploit other group members by choosing $m_i = 0$. Each agent i is assumed to punish other group fellows j proportional to their negative deviation from the own contribution. As discussed in chapter 2, this behavioral pattern is widely observed in experiments and field studies that were conducted in the western cultural area. In general, punishment is costly, but efficient, i.e. each MU spent by the punisher reduces the fitness of the punished agents by $r > 1$ MUs. The intensity of punishment is determined by the population's intrinsic

propensity to punish k, which has evolved over hundreds of thousands of years as a result of gene-culture co-evolution. The expected P&L of an agent is defined by

$$\begin{aligned} E_i[f_i(m_i)] = & -m_i + \frac{g}{n} \cdot m_i \\ & + \frac{g}{n} \cdot (n-1) \cdot \int_0^{\infty} m_j \cdot P_i(m_j) dm_j \\ & - (n-1) \cdot k \cdot r \cdot \int_{m_i}^{\infty} (m_j - m_i) \cdot P_i(m_j) dm_j \\ & - (n-1) \cdot k \cdot \int_0^{m_i} (m_i - m_j) \cdot P_i(m_j) dm_j \ . \end{aligned} \tag{4.1}$$

Agent i has no ex ante information about the contributions m_j of her group fellows. However, all subjects tend to harmonize their contribution behavior as a result of the coordination regime that originates from the subjects' aversion against disadvantageous inequitable outcomes and the evolutionary dynamics (cf. chapter 2 and 3). With time, individual preferences and behavior converges to common norms that ultimately aggregates into culture. This joint cultural background allows each agent to form her first-order beliefs about the others' contributions (Gintis, 2009; Bernheim, 1994; Messick, 1999; Bardsley and Sausgruber, 2005; Henrich, 2004). The intuition for the expected contribution of her group fellows is embodied in the subjective probability distribution $P_i(m_j)$. Group fellows are indistinguishable from agent i's perspective. Thus, the expected utility of a representative agent j is simply given by

$$\begin{aligned} E_i[f_j(m_i)] = & -\int_0^{\infty} m_j \cdot P_i(m_j) dm_j + \frac{r_1}{n} \cdot m_i \\ & + \frac{g}{n} \cdot (n-1) \cdot \int_0^{\infty} m_j \cdot P_i(m_j) dm_j \\ & - k \cdot r \int_0^{m_i} (m_i - m_j) P_i(m_j) dm_j \\ & - k \cdot \int_{m_i}^{\infty} (m_j - m_i) P_i(m_j) dm_j \ . \end{aligned} \tag{4.2}$$

As presented in chapter 2, the expected evolutionary utility of agent i, is defined by the sum of the differences between agent i's expected P&L and the expected P&L of the $n-1$ group fellows

$$u_i(E_i[f_i(m_i)], E_i[f_j(m_i)]) = (n-1) \cdot (E_i[f_i(m_i)] - E_i[f_j(m_i)]) \tag{4.3}$$

The first order condition for an extremum of (4.3) is given by

$$\frac{\partial u_i(E_i[f_i(m_i)], E_i[f_j(m_i)])}{\partial m_i} \stackrel{!}{=} 0 \ , \tag{4.4}$$

with

$$\begin{aligned} \frac{\partial u_i(E_i[f_i(m_i)], E_i[f_j(m_i)])}{\partial m_i} =& \Big(-1 - k \cdot (1 + r_2 - n \cdot r_2) \int_{m_i}^{\infty} P_i(m_j) dm_j \\ & + k \cdot (1 - n + r_2) \cdot \int_0^{m_i} P_i(m_j) dm_j \Big) \cdot (n-1) \\ =& -1 + \frac{r_1}{n} + k \cdot (n-1) \cdot (a_i(m_i) \cdot r_2 + a_i(m_i) - 1) \ . \end{aligned} \tag{4.5}$$

As shown in chapter 2, the welfare maximizing strategy for subjects in the coordination regime is to choose a contribution that corresponds to the median values of the subjective probability distribution, so that $a_i(m_i) = 1/2$. In reality however, people's cooperation preferences are heterogeneous and complex. Transforming equation (4.4) to represent $a_i(m_i)$ as a function of the propensity to punish k, and not the reverse as done in chapter 2 yields

$$a_{m_i}(k) = \frac{1 + k \cdot (n - 1 - r)}{k \cdot (n-2) \cdot (r+1)} \ . \tag{4.6}$$

Remember that by the definition in equation (2.16) it holds that

$$a_{m_i}(k) := 1 - CDF_i(m_i) = P_i(\{m_j > m_i\}) = \int_{m_i}^{\infty} P_i(m_j) dm_j \ . \tag{4.7}$$

Equation (4.6) in combination with the definition in (4.7) describes a functional dependency between the expected fraction $a_{m_i}(k)$ of group fellows that subject i believed to contribute more than her own contribution m_i and the

propensity to punish k_i as well as the group size n and the punishment efficiency r. This relation allows to estimate the first-order beliefs of subjects about the contributions of their group fellows.
In principle, $a_{m_i}(k)$ in equation (4.6) is a function of the propensity k_i to punish that is specific of agent i, because k_i and $a(m_i)$ are the two sides of the coin on how agent i reveals her first-order beliefs concerning the contribution of others. However, because the experiments provide through expression

$$k_{i,j} = \frac{p_{i \to j}}{m_i - m_j}$$

a more fine-grained information, we can interpret the characteristic propensity to punish of agent i as a random variable that fluctuates around the culturally and genetically determined intrinsic value of k_i. In particular, k_i varies slightly from subject to subject and across periods. It has to be stressed that the dynamic inconsistency of k_i is not the consequence of a lack of rationality, i.e. that subjects are not acting according to the identified Nash equilibrium $k^\star$ as defined in equation (2.18). Moreover, they reflect an uncertainty, i.e. a lack of information, regarding the norm-conforming behavior with respect to the contributions of others (Manski, 1977; McFadden, 1974, 1981). This leads to interpret expression (4.6) as determining $a_{m_i}(k)$ as a function of all possible realizations of $k_{i,j}$. Using the empirical distribution of the observed punishment reactions and the distribution of the observed contributions together with equation (4.6) provides a way to infer the cooperation preferences of subjects, given that the subjects punished evolutionary optimally by playing a strategy profile that results from the Nash solution obtained from problem (2.13). As stated above, we define a subject's *cooperation preference* by the ratio between her level of contribution and the expected contributions of her group fellows (first-order belief).

4.2.2 Empirical cooperation preferences of subjects

First, we need to address a caveat in order to appropriately infer the cooperation preferences of the subjects from the empirical data: In contrast to the propensity to punish, the observed contributions m_i in the first period of the

game and in the corresponding full-period sample are distinctly different. This can be seen for instance by performing a two sample Kolmogorov-Smirnov test comparing these two distributions, which rejects the hypothesis that they result from the same underlying statistical distribution (p-value= 0.0057). A possible origin of this difference is that people make strategic decisions about their contributions while playing a public goods game (Fischbacher, 2001; Fischbacher and Gachter, 2010; Herrmann and Thoeni, 2009). In iterating game plays, subjects may adapt and/or learn to evolve their strategic behavior concerning the level of contribution m_i. In contrast, the corresponding Kolmogorov-Smirnov test regarding the propensity to punish k_i does not allow to reject the null hypothesis that observations form the first period of the experiment and the full-sample data set originate from the same underlying distribution. Hence, the propensity to punish k_i is not subject to time effects in the form of adaptation or learning.

In order to eliminate any bias from the observed contributions m_i that could result from such short term adaptation and/or learning, we consider in the following only data from the first period played in each treatment. In this way, the measured contributions and punishments for each subject reveal the true first-order beliefs about the others' contributions embodied in $P_i(m_j)$ and the personal cooperation preference. The observed contribution can then be interpreted as the focal action resulting from a decision process that is not affected by strategic considerations but based only on the individual cooperation preference and the internalized cultural norms (Messick, 1999; Bardsley and Sausgruber, 2005; Bernheim, 1994).

By the definition of equation (4.6), the term $a_{m_i}(k_{i,j})$ reflects the probability that the own contribution is less than the expected contributions of the others. This allows us to quantify to which extent subjects consciously choose to deviate from their first-order beliefs. As stated above, a reciprocal and norm-conforming behavior regarding m_i corresponds to the median value of subjective probability distribution about the group fellows contributions. Using the data set from the first period of each experiment and applying equation (4.6) with $n = 4$ and $r_2 = 3$ to all empirically observed values $\{k_{i,j}\}$ inferred from equation (4.2.1) yields a set of cooperation preferences $a_{m_i}(k_{i,j})$, which is broadly distributed. By definition (4.7), $0 \leq a_{m_i}(k_{i,j}) \leq 1$, which implies

via equation (4.6) that $0.125 \leq k_{i,j} < +\infty$. Thus, subjects who do not punish according to the value range that is specified by the evolutionary optimal behavior are not considered in our model. Reasons for the deviation from the evolutionary optimal behavior include different utility objectives or a different bounded rationality. Restricting our analysis to those subjects who punished negative deviators, i.e., with non-zero $k_{i,j}$'s, we find that 49 out of the 292 available observations correspond to $k_{i,j} < 0.125$. These are ignored in our analysis. Each of the remaining 243 observations $a_{m_i}(k_{i,j})$ represent the ex ante expected fractions of group fellows that a given agent i believed to contribute more than herself, given she chose a contribution of m_i. In other words, subject i intentionally chose her contribution m_i with the belief that a fraction $a_{m_i}(k_{i,j})$ (respectively $1-a_{m_i}(k_{i,j})$) of her group fellows would contribute more (respectively less) than herself.

The cooperation preferences of the different subjects participating in the three experiments can thus be classified roughly into three categories:

- **malevolent or free-rider**: For $a_{m_i}(k_{i,j}) \gg 0.5$, subject i has intentionally chosen to contribute an amount of $m_i^\star$ MUs for which she expects that on average a fraction of more than $\gg 50\%$ of the group fellows will contribute more than this value.

- **conditional cooperator**: For $a_{m_i}(k_{i,j}) \simeq 0.5$, subject i has reciprocated (conditionally cooperated) by choosing to contribute an amount of $m_i^\star$ MUs that she expects to match the norm-conforming behavior of the other participants.

- **benevolent or unconditional cooperator**: For $a_{m_i}(k_{i,j}) \ll 0.5$, subject i has intentionally chosen to contribute an amount of $m_i^\star$ MUs for which she expects that on average a fraction of less than $\ll 50\%$ of the remaining group members will contribute more than this value.

The distribution of cooperation preferences $a_{m_i}(k_{i,j})$ in the population is depicted as a normalized histogram in figure 4.1. One can observe that the histogram in figure (4.1) is skewed towards values of $a_{m_i}(k_{i,j})$ larger than 0.5, which implies that a larger fraction of subjects tends to be imperfect conditional cooperators: Subjects are more likely to intentionally contribute less

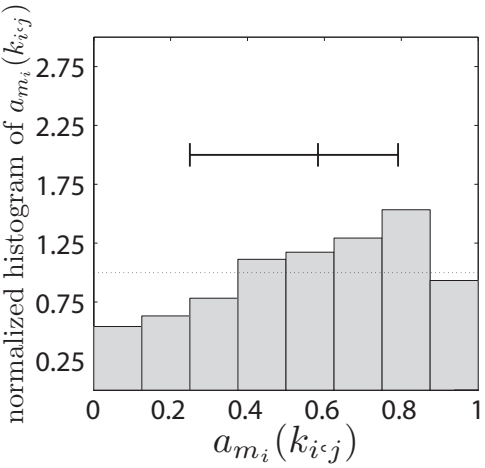


Figure 4.1: Normalized histogram of the distribution of cooperation preferences $a_{m_i}(k_{i,j})$ in the subject pool of three public goods games with punishment. The median and 20% / 80% quantiles of the distribution of the values $\{a_{m_i}(k_{i,j})\}$ are indicated by the three ticks on the horizontal bar.

than their first-order beliefs about the contribution of the others. In detail, 20% of the population "defect" by choosing to contribute less than what they believed a fraction of 80% will contribute. Another 25% of the subjects contribute according to their belief that 80% of the population will contribute less than themselves. This representation of cooperation preferences provides an alternative methodology to measure and classify the type of cooperators compared to already existing approaches presented e.g. in (Fischbacher, 2001; Herrmann and Thoeni, 2009).

We now state our first main result:

Result 4.1: *The majority of subjects can be best described as imperfect conditional cooperators. 20% of the subjects defect by contributing less then what they believe 80% will contribute. At the same time, 25% of the subjects contribute more than what they expect 80% of the population will contribute.*

We now compare the distribution of the expected contributions $E_i[m_j]$ (first-order beliefs), embodied in the set of $\{a_{m_i}(k_{i,j})\}$ values, with the effectively observed distribution of contributions m_i. This provides us with a way to validate our evolutionary utility model introduced in chapter 2 by means of the consistency of both distributions. We construct a synthetic survivor function of the distribution of m_i by means of the $\{a_{m_i}(k_{i,j})\}$ values as follows: each observed contribution m_i can be ranked by the simultaneously observed $\{k_{i,j}\}$ values using equation (4.6) in combination with the definition of $a_i(m_i)$ in equation (4.7). In other words, each value $k_{i,j}$ assigns via (4.6) a probability ($a_{m_i}(k_{i,j}) \equiv P(X \geq m_i)$) that the contributions of others will be larger than or equal to m_i. To construct the synthetical survivor function of the expected others' contributions (first-order beliefs), we smooth the scattered data set defined by the pairs of $\{[a_{m_i}(k_{i,j}), m_i]\}$ values using a two step filtering method. First, we calculate the mean $\bar{m}_x$ of those $\{m_i\}$ values taken from the subsets of pairs $\{[a_{m_i}(k_{i,j}), m_i]\}$ where $a_{m_i}(k_{i,j})$ falls within a range of $[x-0.04, x+0.04]$ for all $x \in [0, 0.02, 0.04, \cdots, 0.98, 1]$. Following that, we apply a Savitzky-Golay filter on the resulting uniformly spaced data set of pairs $\{[x, \bar{m}_x]\}$ in order to obtain a smoothed and meaningful approximation of the survivor function. With an appropriate choice of the order of the smoothing polynomial, the Savitzky-Golay filtering method has the nice property of preserving the features of the underlying distribution, such as its moments of orders up to one plus the order of the smoothing polynomial (here the order is two). This procedure finally results in the synthetically calculated survivor function $A(m_i)$ of the expected contributions $E_i[m_j]$.

Figure (4.2) shows the resulting reconstructed survivor function $A(m_i)$ as a black continuous line including the one standard error band (two dotted lines). The sharp drop observed at $m_i \geq 17$ can be interpreted as a boundary effect due to the maximum endowment equal to 20. The empirical survivor function of the first period contributions for 349 subjects ($N = 440$) who punished at least once a negative deviator during the experiment is plotted as a gray continuous line. It is striking to find such an agreement between the empirical survivor function of the contribution m_i for the 349 punishers and the synthetically calculated survivor function $A(m_i)$ of the first-order beliefs. This suggests that our evolutionary utility model and its predicted functional rela-

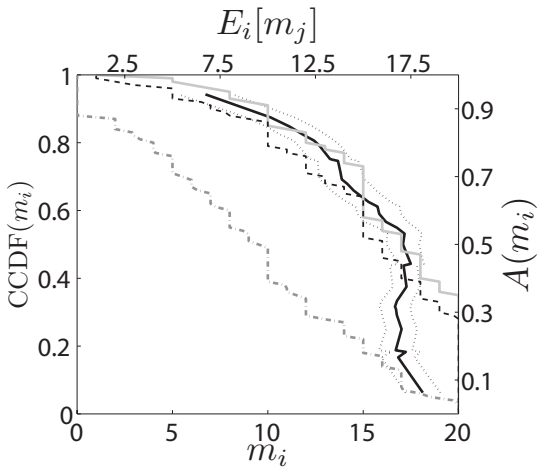


Figure 4.2: Synthetically calculated survivor function $A(m_i)$ of the expected contributions (first-order beliefs) $E_i[m_j]$ (black continuous line) including the one standard error range (dotted lines) and the corresponding empirical survivor function of the observed contributions m_i (gray continuous line) calculated from data of three public goods games with punishment. The dashed-dotted gray line shows the empirical survivor function of m_i for "anomalous" subjects which exclusively punished positive deviators. The dashed black line shows the survivor function of the contributions from subjects who did not punish at all. The quantitative agreement between the survivor function $A(m_i)$ of the expectations $E_i[m_i]$ reconstructed from the punishments and its directly observed complement m_i supports our modeling approach as applied to "normal" punishers.

tion between $a_i(m_i), a_{m_i}(k_{i,j}), k, n$ and r explains well the empirically observed data in the three experiments.

Consider in contrast the 80 subjects of the population who did not punish at all, i.e. who are second-order free riders: the survivor function of their contributions is shown as a dashed black line. This distribution departs significantly from our model's prediction, as should be expected. The deviation is even much stronger for anti-social punishers (11 subjects) who exclusively punished positive deviators during the experiment. The corresponding survivor function is shown by the dashed-dotted gray line.

Our second main result is given by:

Result 4.2: *The pool of subjects shares a common norm regarding the level of contribution and thus subjects are able to form a realistic first-order belief about the contributions of the other group fellows.*

4.2.3 Conclusion

This section provided an analysis of the distribution of preferences either to cooperate (benevolent behavior) or to defect (malevolent behavior) within the population of subjects from three public goods experiments. The *cooperation preference* of a subject is defined by the ratio between her contribution to the public good compared to her expectations about the contribution of the remaining subjects (first-order belief). To infer the empirical first-order beliefs of subjects, we used the observed intensity of punishment as a measure to quantify the punisher's disappointment about the defector's contribution given the punisher contributed a specific amount. This measure allowed us to reveal the subject's intentional ex ante preferences to cooperate (benevolent) or to defect (malevolent). We find that subjects are more likely to contribute less than what they expect the others to contribute (imperfect conditional cooperators). However, 25% of the population in the experiments chose to contribute an amount that they expected to be larger than what 80% of the population would contribute. In contrast to existing approaches that use a distinct set of behavioral patterns to classify subjects (Fischbacher, 2001; Herrmann and Thoeni, 2009; Frey and Meier, 2004), we have presented a method to quantitatively describe the distribution of *cooperation preferences*. Subsequently, our findings and the underlying evolutionary utility model are validated by means of the consistency of two distributions of contributions: (i) the effectively observed distribution of contributions and (ii) the distribution of the expected contributions (first-order beliefs) that was reconstructed from the observed punishment data. We find an excellent agreement between both distributions for the vast majority of the subjects. This finding implies that a common norm regarding the level of contribution is present among the subjects in the pool, as subjects were able to accurately estimate the norm-conforming behavior via their first-order beliefs.

4.3 The effect of punishment on the level of cooperation

This section provides a detailed analysis of the evolution of cooperative behavior among subjects who face a voluntary contribution mechanism with punishment opportunity. Existing literature already provides insights into the contribution dynamics of individuals who are exposed to the threat of punishment by group fellows (Sonnemans, Schram, and Offerman, 1999; Houser and Kurzban, 2003; Bardsley and Sausgruber, 2005; Bardsley and Moffatt, 2007). Other studies focus on the effects of different punishment efficiencies, varying group sizes or changing communication/information structures and their impact on the subjects' cooperation behavior (Decker et al., 2003). Still others analyze the effects coming along with combinations of punishment and reward opportunities (Andreoni, Harbaugh, and Vesterlund, 2003). In this section, we focus on the first-order dynamics of the contribution behavior of subjects who play a public goods game with punishment. For this purpose, we look at the adaptation dynamics of the voluntarily contributed MUs between two consecutive periods t and $t+1$ in three previously conducted lab experiments (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009) by analyzing and presenting the observed micro-level data on a per-subject level. We are able to uncover the different effects that punishment is causing in repeated interactions among partners and in one-shot interactions among strangers. While the contribution behavior in repeated interactions among partners is mainly determined by the direct reciprocity effect, punishment induces a strong tendency to conformity in one-shot interaction among strangers (Bardsley and Sausgruber, 2005).

4.3.1 Empirical foundation

In the following, we analyze the data from three public goods games with punishment, which have been conducted by Fehr/Gaechter in 2000/2002 and Fudenberg/Pathak in 2009 (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009). The public good games were played for a total of 6 (Fehr and Gachter, 2002) and 10 (Fehr and Gachter, 2000; Fudenberg and Pathak, 2009) periods. Subjects received 20 monetary units in each period which

subsequently could be contributed to the public good. The participants were arranged in groups of 4 in a constant group setup (partner treatment) and a dynamic group setup (stranger treatment). To control for direct reciprocal effects, the stranger treatment ensured that subjects were only engaged in one-shot interactions. Subjects remained anonymous in both treatments. In period t, subject i decided to contribute $m_i \in [0, 20]$ MUs to the public good that yields a return of $g = 1.6 \cdot \sum_{i=1}^{4} m_i$ back to the group. If all group members contribute an identically amount of m MUs, each of them receives $1.6 \cdot m$. In principle, the per capita gain that an individual subject received for 1 contributed MU $m_i = 1$ is $\frac{g}{n} = 0.4$ MU. As a consequence, this experiment setup is susceptible to materials self-interest and free-riding behavior. That is, a subject who decides not to contribute anything performs better than a subject who contributes. The design of public goods game is a common social dilemma that is present in many real-life situations and has been most prominently introduced as the tragedy of the commons (Hardin, 1968). After each subject i has contributed an amount of m_i MUs to the public good, they learn about the contributions of their fellows and are provided with the opportunity to punish other group fellows by spending an additional amount of their endowment. Punishment was efficient with a factor of 3, that is, one MU spent by subject i to punish subject j caused a cost of 3 MUs to subject j.

Figure 4.3 shows the average contribution per subject for the partner (diamonds) and the stranger (squares) treatment over 10 periods including a one standard deviation error bands resulting from the observations of all three experiments. The plot reveals a significant increase of cooperation for the partner treatment whereas the dynamics in the stranger treatment indicate that cooperation is only sustained but not promoted over time.

A more precise analysis of the observed behavior on a per subject level reveals the immediate effect of punishment on the cooperation dynamics: The color maps in figures 4.4 and 4.5 depict the subjects' responses $y_i(t)$ in period $t+1$ given they deviated from the group's average contribution $\bar{m}$ in period t by a value of $x_i(t) = m_i(t) - \bar{m}(t)$ MUs. For instance, subject i deviates from the average contribution $\bar{m}$ of its group in period t by a value of $x_i(t)$, then

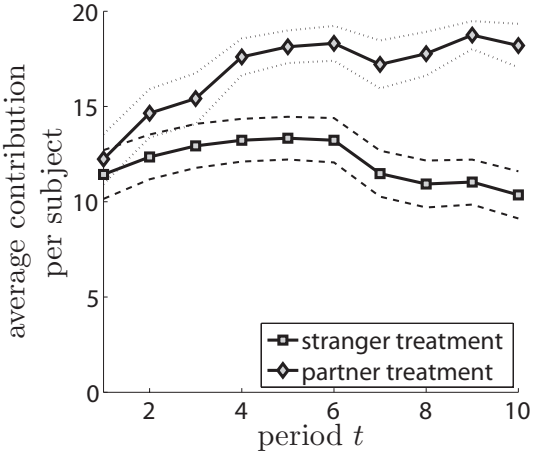


Figure 4.3: Average contribution per subject including one standard error band, as reported empirically in (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009). Data from all experiments have been pooled. The diamond markers corresponds to the results of the partner treatment, the square markers belong to the stranger treatment.

the corresponding value on the y-axis reflects the relative adaptation $y_i(t) = m_i(t+1) - m_i(t)$ of her contribution between periods t and $t+1$. Thus, the set of observed values $(x_i, y_i(x_i))$ characterizes the individual response behavior $(y_i(x))$ of subject i in period $t+1$, given she experienced a deviation of x_i from the group average $\bar{m}(t)$ in period t. The shade of gray reflects the number (in log scale) of corresponding $(x_i, y_i(x_i))$ observations in the pooled experimental result data. The plots include data from all three experiments always excluding results from the last period played, in order to control for the last-round effect. The linear regression between $y_i(x_i)$ as a function of x_i calculated separately in the two intervals $x < 0$ and $x \geq 0$ is given by function $g(x)$. To obtain $g(x)$, we apply a moving average smoothing method on the original scattered set of data $(x_i, y_i(x_i))$ that calculates the mean of all $y_i(x_i)$ values which fall in a window of $[x-1, x+1]$ for all $x \in [-20, 20]$. The function $g(x)$ corresponds to the best piecewise linear fits on the smoothed

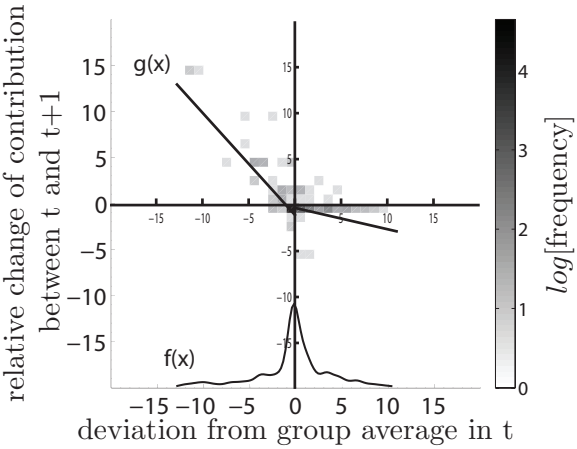


Figure 4.4: First order dynamics in the partner treatment of subject i's change of contribution $y_i = m_i(t+1) - m_i(t)$ (y-axis) against her deviation from the group average contribution $m_i(t) - \bar{m}(t)$ (x-axis) sampled across periods. The function $g(x)$ shows the best piecewise linear fit on the smoothed (moving average) data set $(x_i, y_i(x_i))$. Two fits are performed separately in the intervals $[-12.75, 0]$ and in $[0, 11]$. The lower curve $f(x)$ represents the estimation of the probability density function of the deviations x.

data for the two distinct ranges of negative and positive deviations x. In the partner treatment the term $g_p(x)$ is defined by

$$g_p(x) = \begin{cases} -1.1 \cdot x - 1.1 & \text{for } x \leq 0 \\ 0.23 \cdot x - 0.30 & \text{for } x > 0 \ . \end{cases} \tag{4.8}$$

The best linear fit $g_s(x)$ for the stranger treatment is given by:

$$g_s(x) = \begin{cases} -0.3 \cdot x + 0.4 & \text{for } x \leq 0 \\ -0.35 \cdot x & \text{for } x > 0 \end{cases} \tag{4.9}$$

The lower curves $f(x)$ in figures 4.4 and 4.5 represent the estimate of the probability density function of all observed deviations x_i, obtained by a standard kernel smoothing method.

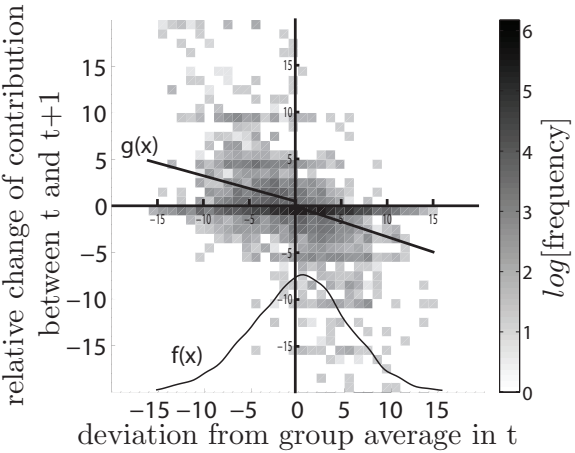


Figure 4.5: First order dynamics in the stranger treatment of subject i's change of contribution $y_i = m_i(t+1) - m_i(t)$ (y-axis) against her deviation from the group average contribution $m_i(t) - \bar{m}(t)$ (x-axis) sampled across periods. The function $g(x)$ shows the best piecewise linear fit on the smoothed (moving average) data set $(x_i, y_i(x_i))$. Two fits are performed separately in the intervals $[-15, 0]$ and $[0, 11]$. The lower curve $f(x)$ represents the estimation of the probability density function of the deviations x.

The difference between the dynamics in the partner and the stranger treatment is striking. While the distribution $f(x)$ of the deviations from the average contribution is approximately symmetrical around 0 in the partner treatment, the response y_i is highly asymmetric: for positive deviators ($x_i(t) > 0$), the distribution of the adaptation behaviors $y_i(x_i)$ between t and $t+1$ is approximately symmetric around 0. In contrast, for negative deviators ($x_i(t) < 0$) most of the y_i's are found in the second quadrant ($y_i(x_i) > 0$). The form of equation (4.8) illustrates the different decision characteristics of individuals who either defected ($x \leq 0$) or cooperated ($x > 0$): for $x < 0$, subjects react with a strong increase in contributions at the next period ($t+1$) and tend to fully compensate their negative deviations in period t. As seen from the fact that $g(x) > -x$ for negative values of x, they even show a tendency to over-contribute. On the contrary, subjects who contribute more than the group average ($x > 0$) do not significantly decrease their contributions in the next period resulting in flat

slope of function $g(x)$ for $x > 0$. This asymmetry introduces a drift towards larger overall cooperation and suggests that cooperation between partners is catalyzed by direct reciprocity. The direct reciprocity effect in combination with punishment is able to explain the emergence of cooperation in the partner treatment. However, among strangers the situation is different.

In the stranger treatment, the distribution of deviations from the mean contribution is also approximately symmetrical around 0, but the response $y_i = m_i(t+1) - m_i(t)$ is much more symmetric than for the partner treatment: The slope of $g(x)$ for negative deviations ($x < 0$) is nearly the same as for positive deviations. This means that defectors ($x < 0$) tend to cooperate more at the next period, as in the partner treatment, while, in contrast, cooperators ($x > 0$) tend to adapt their behavior and start to cooperate less. In the stranger treatment, positive deviations compensate negative deviations with successive increments and decrements causing a convergence of free-riders and cooperators across time. This allows to conclude, that the feedback provided by punishment induces a contribution dynamic that ultimately results in a homogeneous level of contributions which is ex ante determined by the average level of cooperation of the group $\bar{m}$ in the first period. The underlying dynamics work as follows: subjects who contributed less than the average increase their contributions in the next period until they contribute as much as the remaining subjects. This is characterized by $0 < g(x) < -x$ for $x < 0$. On the other hand, those who contributed more slightly decrease their contributions in the next period, which is characterized by $-x < g(x) < 0$ for values $x \geq 0$. These two differences ((i) smaller change towards cooperation for defectors and (ii) negative change towards less cooperation for cooperators) explain the stabilization of the level of cooperation in the stranger treatment as apparent from figure 4.3. For strangers, punishment thus seems to serve as a coordination mechanism, but does not lead to the emergence and reinforcement of cooperative behavior. This is in line with our finding presented in chapter 2 and 3.

Additionally to the previous analysis, the characteristic trend of the contributions in a populations of partners and strangers can offer valuable clues to the overall dynamics. The expected deviation of the population's contribution in $t+1$ can be written as the cumulative sum of the first-order deviation dy-

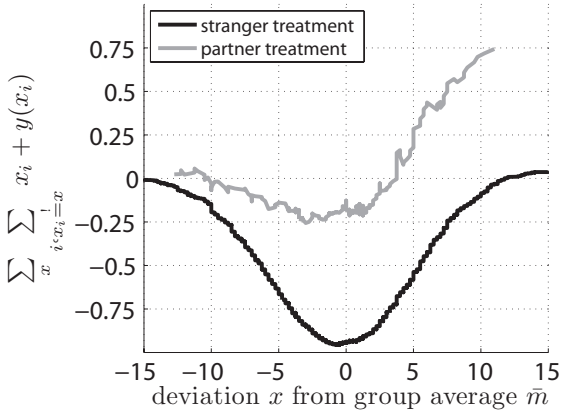


Figure 4.6: Cumulative first-order deviation dynamics as defined by equation (4.10). The curves can be interpreted as the expected relative change of the average group contribution given that only the subjects who deviated by $x_i(t) <= x$ adapt their contributions while those of subjects who deviated by $x_i(t) > x$ stay constant. The gray line corresponds to the partner treatment, the black line to the stranger treatment. The values have been normalized.

namics $x_i + y(x_i)$ under the condition that the deviation from the population average $\bar{m}$ in t is less than x. This yields the following expression:

$$\begin{aligned} E[x(t+1)|x(t) \leq x)] &:= \int_{-20}^{x} E[x(t+1)|x(t) = x)] f(x) dx \\ &:= \sum_{x} \sum_{i, x_i \stackrel{!}{=} x} x_i + y(x_i) \end{aligned} \tag{4.10}$$

Equation 4.10 represents the expected relative change of the average group contribution in the population conditional on the fact that only subjects who deviated by $x_i(t) <= x$ adapt their contribution while the contributions of subjects who deviated by $x_i(t) > x$ stay constant. The corresponding shapes of equation (4.10) for the partner and the stranger populations are shown in figure 4.6. Indeed, figure 4.6 reveals that the level of cooperation in a population of strangers on average levels out over time, which results in a homogeneous population.

We conclude with our third main result:

Result 4.3: *Altruistic punishment among stranger acts as a coordination mechanism that allows to sustain a certain level of cooperation, but cannot explain its evolutionary emergence. In contrast, providing a punishment opportunity among partners can explain an increase of cooperation, however, only due to the presence of the direct reciprocity effect.*

Even though there is a wide-spread belief that punishment, and in particular altruistic punishment, is a key candidate to explain the evolutionary emergence of cooperation (Boyd and Richerson, 1992; Fehr and Gachter, 2000, 2002; Masclet et al., 2003; Noussair and Tucker, 2005; Guererk et al., 2006; Nikiforakis and Normann, 2008; Herrmann et al., 2008; Gachter et al., 2008; Egas and Riedl, 2008), our analysis of the micro level data from the three experiments (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009) provides a contrary evidence. We have shown indeed that punishment can promote cooperation among partners by sustaining the direct reciprocal mechanism. However, it only serves to maintain a preexisting level of cooperation in one-shot stranger interactions. Again, this is in line with our findings presented in chapter 2 and 3. Given the widespread degree of cooperative behavior that is present today in almost all areas of human life, this observation requires to identify the detailed mechanisms accounting for the evolutionary origin, emergence and sustainment of cooperative behavior.

4.3.2 Simulation results

The presented results in this section originate from the same evolutionary simulation model that has been introduced in chapter 3. The agents' fairness preferences are fixed to disadvantageous inequity aversion (cf. dynamics C in section 3.2.2) which have been identified to be the predominant measure of fairness in an aggregated pool of subjects from three different experiments (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009).

The agents play a public goods game with punishment in groups of size n. The P&L of agent i in period t equals

$$\hat{s}_i(t) = \frac{g}{n} \cdot \sum_{j=1}^{n} m_j(t) - m_i(t) - \sum_{j \neq i} p_{i \to j}(t) - r \sum_{j \neq i} p_{j \to i}(t)$$

as previously defined in equation (3.4). The simulation sequence is exactly equivalent to the simulations runs in chapter two. The population of agents is arranged in groups of $n = 4$ and the group project yields a per capita return of 0.4 for each invested MU, i.e. $g = 1.6$. The punishment efficiency factor is fixed to $r = 3$, i.e. for each MU spent by the punisher, the fitness of the punished agent is reduced by 3 MU. Sampling the simulation model over a set of predefined parameter configurations, allows us to explore and to analyze the sensitivities of a population of individual agents to specific exogenously fixed conditions. In the following, the sensitivity of the level of cooperation $m_i(t)$ is analyzed with respect to the level of punishment in the population. For this purpose, the dynamics of $m_i(t)$ for fixed values of k, ranging from zero ($k = 0$) up to excessive punishment behavior with $k = 1$ are examined.

In the absence of punishment, i.e. all k's are imposed equal to 0, we find that cooperation that was maintained previously in the presence of punishment decays after a few thousand periods as shown in figure 4.7. In contrast, if punishment is restored at $k = 0.25$, cooperation remains stable at the previously initialized level. Figure 4.8 shows the average level of cooperation in a group of 4 agents after a transient period of 20,000 simulation periods for 1000 system realizations as a function of the propensity to punish k. The level of cooperation for all agents was initialized by a value drawn from a uniformly distributed random variable in $[0.99, 1.01]$. This figure reveals that the level of cooperation undergoes a phase transition at the critical value $k_c \simeq 0.125$, at which it becomes non-zero and grows rapidly to a saturation value. For propensities to punish larger than 0.25, the level of cooperation remains constant at its saturation value. The value $k^\star \simeq 0.25$ seems to be the minimum propensity to punish that enforces to sustain a maximum level of cooperation. This suggests that evolution may have selected an "optimal" propensity to altruistically punish defectors in order to sustain cooperation. To corroborate

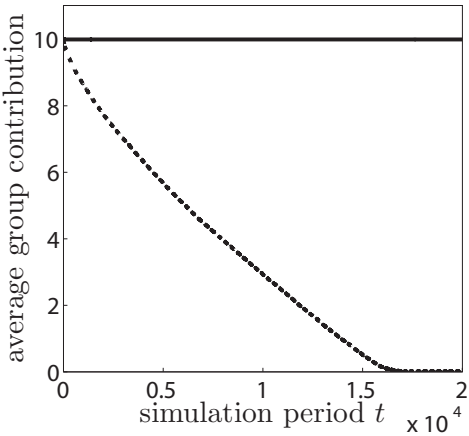


Figure 4.7: Average group contribution for a group of 4 agents with punishment ($k = 0.25$ - continuous line) and without ($k = 0$ - dashed line) for dynamic C (disadvantageous inequity aversion) over 10,000 simulation periods and 16 system realizations. The initial contribution $m_i(0)$ for all agents i of a group is randomly drawn form a uniform distribution in $[0.99, 1.01]$.

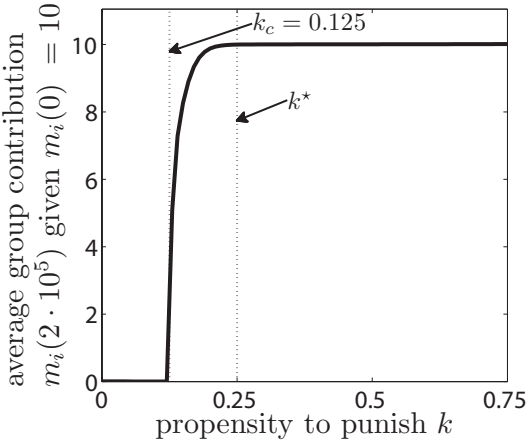


Figure 4.8: Average group contribution for a group of 4 agents as a function of k for dynamic C (disadvantageous inequity aversion) after an equilibrium time of 20,000 simulation periods and for 1000 system realizations. k is fixed to the corresponding value on the x-axis and the initial contribution $m_i(0)$ for all agents i of a group is randomly drawn form a uniform distribution in $[0.99, 1.01]$.

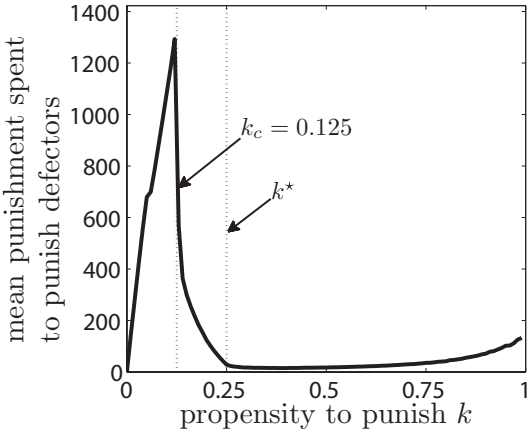


Figure 4.9: Average punishment spent to punish defectors for a group of 4 agents as a function of k after an equilibrium time of 20,000,000 simulation periods and for 100 system realizations. k is fixed to the corresponding value on the x-axis and the initial contribution $m_i(0)$ in period 0 for all agents i of a group is randomly drawn form a uniform distribution in $[4.99, 5.01]$. A value of $k^\star \simeq 0.25$ corresponds to an optimal value of the propensity to punish associated to a minimum of the global punishment expenditure.

this hypothesis, we now consider the intrinsic propensity to punish k as a measure of deterrence. Figure 4.9 plots the average amount of MUs spent to punish a defector during 5,000,000 simulation periods for 3200 system realizations as a function of the propensity to punish k. As in the setup of figure 4.8, the level of cooperation $m_i(t)$ for all agents is initialized at period $t = 0$ by a random variable uniformly distributed in $[0.99, 1.01]$. The results show clearly that for values of k above the critical value of $k_c \simeq 0.125$, which corresponds to a higher level of deterrence, effectively less exertion of costly punishment is caused in order to maintain a certain level of cooperation and norm conformity, respectively. This responsive behavior was manifested in many empirical observations (Kleiman, 2009; Kennedy, 2008; Jensen, 2010; Holmas et al., 2010). The value $k^\star \simeq 0.25$ corresponds to the minimum overall punishment cost with a stable maximum cooperation level. This substantiates that evolution may have selected an "optimal" propensity to punish to sustain cooperation and prevent defection in contexts in which people exhibit disadvantageous in-

equity aversion. Comparable results were obtained using a different simulation model, as reported in (Kleiman and Kilmer, 2009).

We now state or last main result of this chapter:

Result 4.4: *Evolution has selected an optimal level of altruistic punish, which is only able to sustain a pre-existing level of cooperation in a population of agents. Thus, punishment alone cannot explain the emergence of cooperation. The evolutionary selected level of punishment is optimal with respect to the group welfare.*

4.3.3 Conclusion

Neither the micro-data analysis of the empirical observations from three public goods experiments (Fehr and Gachter, 2000, 2002; Fudenberg and Pathak, 2009), nor the results obtained from our evolutionary simulation model suggest that (altruistic) punishment can be considered as the key mechanism in explaining the evolutionary emergence of cooperation. Moreover, the results presented in this section reveal that (altruistic) punishment provides a key stabilization mechanism for sustaining cooperation among strangers. The evolutionary rooted predisposition to punish based on disadvantageous inequity aversion does not appear to help in promoting the level of cooperation. This inevitably brings up the question for the causative mechanisms that underly the evolutionary emergence of cooperative behavior. In particular, the preexisting high level of the first period contributions in the experiments at approximately 12 MUs remains unexplained. To take on this remaining puzzle, a more detailed analysis of the interaction between conformity promoting mechanisms, such as (altruistic) punishment, and heterogeneity inducing processes is performed in the next chapter.

5. The emergence of cooperation

This chapter aims at explaining the emergence of cooperation in voluntary contribution mechanisms under evolutionary dynamics. In particular, we focus on the high levels of contributions that can be observed in public goods game experiments with punishment. To shed light on this puzzling behavior, we discuss and analyze the long-term interaction between the heterogeneity of agents and the feedback provided by punishment. The findings presented in chapter 4 revealed that the predisposition to punish unfair behavior at own costs prevents free-riding behavior in social dilemmas and, in particular, acts as a coordination mechanism in one-shot interactions among strangers: if the population displays a propensity to punishment that is sufficiently strong, agents converge to a homogeneous and stable level of cooperation that is sustained across time (Bardsley and Sausgruber, 2005; McNamara and Leimar, 2010). After a population has achieved conformity with respect to the contributions, punishment cannot serve any longer as a catalyst for cooperation but moreover remains as a passive mechanism of deterrence. This is caused by the fact that people tend to punish mainly free-riders, i.e. negative deviators, with a intensity proportional to their negative variation[1]. In other words: as soon as a population is sufficiently homogeneous with respect to the contributions, punishment becomes inactive and ultimately disappears if all contributions are equal. Thus, altruistic punishment cannot explain the

[1] see e.g. figure 2.1 in chapter 2 and results presented (Fehr and Gachter, 2002)

evolutionary origin of the high levels of cooperation that are observed today. This calls for the identification of the underlying evolutionary mechanisms that in combination with punishment account for the emergence and the sustainment of cooperation. For this reason, we extend the evolutionary simulation model introduced in chapter 3 by implementing different multi-level selection mechanisms that maintain heterogeneity both between and within groups of a population. Specifically, we look into different variants of inter and intrademic group selection.

5.1 Introduction

The emergence of cooperation among groups of organisms ranging from bacterial strains to small human tribes, regional communities all the way up to (inter-)cultural areas and states, is still considered as one of the 25 most compelling puzzles science is facing today (Pennisi, 2005). The principle of the survival of the fittest, i.e. natural selection and genetic drift, discriminates individuals who incur costly behavior that is beneficial to other members of their species. Results from lab experiments, field studies and observations in everyday life suggest that humans exhibit a high level of cooperative behavior even in one-shot interactions in which no reciprocal effects are present. This pro-social behavior strictly contradicts rational choice and is in conflict with selfish maximization and the paradigm of inclusive fitness.

One mechanism that is considered to contribute to the solution of this puzzling behavior is the punishment of free-riders and norm-violators. This hypothesis has been verified with the help of public goods game experiments, which was already discussed in the previous chapters. The empirical observations and our computational results suggest that humans show a predisposition to altruistically punish free-riders at their own costs even if they are only engaged in one shot interactions. This seems to replace one unresolved conflict by another as selfish maximization and standard evolutionary theory, at a first sight, seem to rule out the emergence of altruistic punishment behavior. In the previous two chapters we have demonstrated that altruistic punishment behavior originates and emerges as a result of fairness preferences in the form of disadvantageous inequity aversion. More importantly, an aversion against

disadvantageous inequitable outcomes is an evolutionary stable and dominant behavior that almost surely invades a population that initially only consists of purely self-regarding and selfish acting agents over time.

In section 5.2 we analyze the dynamics of agents who, in the presence of punishment, face a constant heterogeneity in their contributions and show how this affects the structure and the contribution dynamics of the population. In the section (5.3), the simulation model presented in chapter 3 is extended and used to quantitatively explore, test and verify different mechanisms that sustain a level of heterogeneity in the trait pool of the population. This allows us to reveal the conditions that are required for the evolutionary emergence of cooperative behavior. In particular, we look into different variants of multilevel selection and show how the interaction of between-group heterogeneity, within-group heterogeneity and punishment accounts for the emergence of cooperation in a competitive and resource-limited environment that is susceptible to material self-interest.

5.2 Heterogeneity, punishment and the evolution of cooperation

In chapters 2 and 3 punishment was identified as a coordination mechanisms that tends to homogenize a population of agents with respect to their contributions. For homogeneously contributing populations, punishment becomes inactive and the cooperation behavior of agents levels off. However, what happens if complementary evolutionary processes induce a steady level of heterogeneity into the groups of agents? This section provides a detailed analysis of the population dynamics of agents who continuously face a certain level of heterogeneity in their groups while playing a public goods game with punishment. The presented analysis focuses solely on the effect of the interaction between heterogeneity and punishment and does not consider potential mechanisms that cause and maintain heterogeneity in a population. An exploration of alternative mechanisms accounting for heterogeneity in a population of agents is presented in the last section of this chapter.

First, we consider a population of a finite size n that consists of two distinct sets of agents: agents in set **A** contribute an amount of $m + \frac{\Delta m}{2}$, while agents in set

B contribute $m - \frac{\Delta m}{2}$. The population is assumed to have itself coordinated around a joint level of cooperation by contributing on average m MUs to the public good. This means that on average $\frac{n}{2}$ agents contribute $m + \frac{\Delta m}{2}$ while the remaining other half of the population contributes $m - \frac{\Delta m}{2}$. In contrast to the analytical model presented in the last section of chapter 2, agents cannot arbitrarily adapt their contribution behavior. However, the proportional fraction of agents in the two sets **A** and **B** can vary given that the total number of agents in the population stays constant. Similarly to the previous sections and chapters, agents who contribute more punish those who contribute less. As a consequence, agents in set **A** punish agents in set **B** proportional to their deviation Δm and in accordance with their intrinsic propensity to punish represented by the factor k. This linear punishment behavior was observed in many empirical studies with subjects from western civilizations as shown in section 2.2.3. It is assumed that the fertility of agents is determined by their fitness, i.e. agents reproduce proportional to their realized P&L. Over time the population may thus be dominated either by agents of from set **A** or set **B** depending on their realized P&L levels.

In the following we denote the fraction of members in set **A** (**B**) compared to the total number of agents in the population with $\tilde{\mathbf{A}}$ ($\tilde{\mathbf{B}}$). The profit that an agent in set **A** gains from the public goods game is defined by:

$$\begin{aligned} P\&L_{\mathbf{A}} := n \cdot \tilde{\mathbf{A}} \cdot \frac{g \cdot \frac{\Delta m}{2}}{n} - n \cdot \tilde{\mathbf{B}} \cdot \frac{g \cdot \frac{\Delta m}{2}}{n} + \\ g \cdot m - (m + \frac{\Delta m}{2}) - n \cdot \tilde{\mathbf{B}} \cdot k \cdot \Delta m \end{aligned} \tag{5.1}$$

The first term on the right hand side of equation 5.1 represents the additional P&L gained by members of set **A**, resulting from the fact that a fraction of $\tilde{\mathbf{A}}$ contributes an amount of $\frac{\Delta m}{2}$ MU more than the average group contribution to the public good. Accordingly, the second term corresponds to the deficit realized by members of set **A** due to the fact that a fraction of $\tilde{\mathbf{B}}$ contributes $\frac{\Delta m}{2}$ MU less than the average. The third term represents the average P&L received in principle by each member of the population; this value is adjusted to fit the exact P&L of agents in set **A** by means of the first two terms described above. The fourth term represents the costs of contributing to the public good.

The last term essentially reflects the costs of punishing the fraction $\tilde{\mathbf{B}}$ of lower contributing agents. The corresponding P&L of agents in set $\mathbf{B}$ is given by:

$$P\&L_{\mathbf{B}} := n \cdot \tilde{\mathbf{A}} \cdot \frac{g \cdot \frac{\Delta m}{2}}{n} - n \cdot \mathbf{B} \cdot \frac{g \cdot \frac{\Delta m}{2}}{n} + g \cdot m - (m - \frac{\Delta m}{2}) - n \cdot \tilde{\mathbf{A}} \cdot k \cdot r \cdot \Delta m \tag{5.2}$$

The different terms of equation 5.2 are accordingly defined to equation 5.1, except that the last term reflects the costs of being punished and thus it is additionally multiplied by the punishment efficiency factor r.

The population dynamics for agents in set $\mathbf{A}$ can be written using the recurrence equation

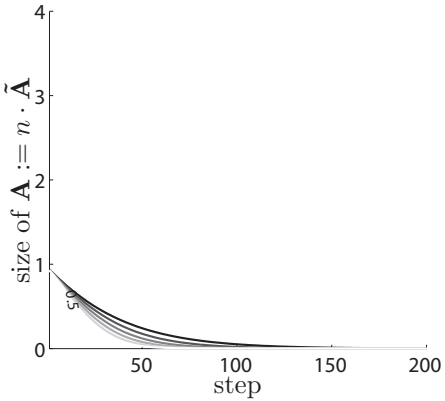
$$\tilde{\mathbf{A}}' = \frac{\tilde{\mathbf{A}} \cdot P\&L_{\mathbf{A}}}{\tilde{\mathbf{A}} \cdot P\&L_{\mathbf{A}} + \tilde{\mathbf{B}} \cdot P\&L_{\mathbf{B}}} \ , \tag{5.3}$$

where $\tilde{\mathbf{A}}'$ represents the fraction of agents in set $\mathbf{A}$ in the next generation. The absolute size of set $\mathbf{A}$ and $\mathbf{B}$ is defined by

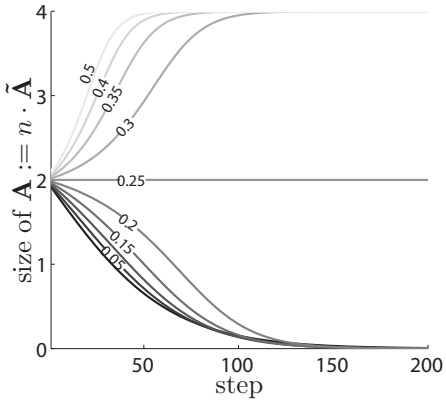
$$\begin{aligned} &\text{size of } \mathbf{A} = n \cdot \tilde{\mathbf{A}} \\ &\text{size of } \mathbf{B} = n \cdot \tilde{\mathbf{B}} = n \cdot (1 - \tilde{\mathbf{A}}) \ . \end{aligned} \tag{5.4}$$

Figure 5.1 depicts the dynamics of the size of set $\mathbf{A}$ in a population with a total size of $n = 4$ agents over 200 generations. The numbers on the curves indicate different initializations of the propensity to punish k. The subfigures (a), (b) and (c) correspond to different initial sizes (1,2,3) of $\mathbf{A}$ in the first generation at period $t = 0$. Figure 5.1(b) shows that an initially "balanced" population of size $n = 4$ consisting of 2 higher contributing agents in set $\mathbf{A}$ and 2 less contributing agents in set $\mathbf{B}$ converges over time to one of the two cooperation fixed-point regimes, i.e. either $m + \frac{\Delta m}{2}$ or $m - \frac{\Delta m}{2}$. The dynamics depend on the predefined propensity to punish: for $k > 0.25$ the population ends up with $\mathbf{A} \cdot n = 4$ whereas for $k < 0.25$ agent from set $\mathbf{B}$ end up dominating over time.

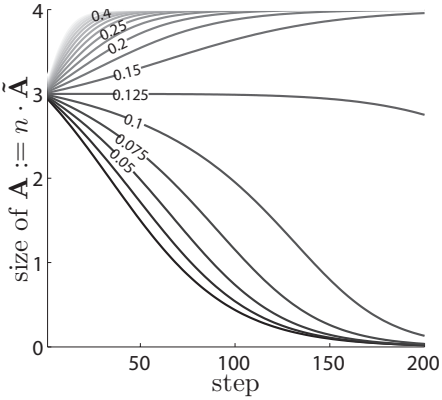
In order to illustrate the interaction between the initial population structure and the propensity to punish, figure 5.2 depicts a map of the population struc-



(a) size of $\mathbf{A} = 1$ at $t = 0$



(b) size of $\mathbf{A} = 2$ at $t = 0$

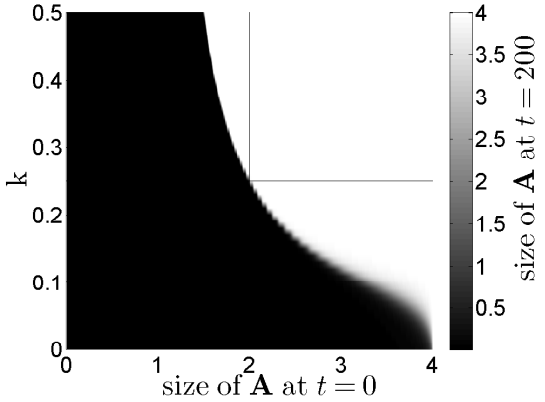


qquad

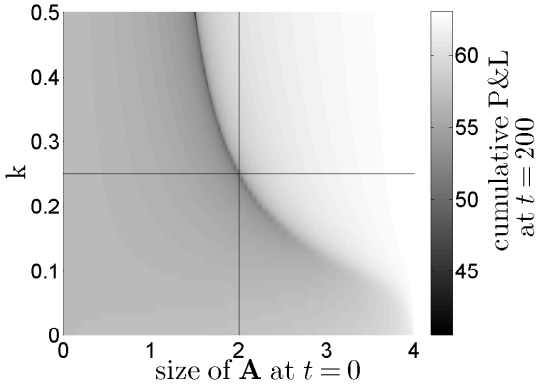
(c) size of $\mathbf{A} = 3$ at $t = 0$

Figure 5.1: Population of $n = 4$ agents playing a public goods game with punishment with $g = 1.6$, $r = 3$ and $\Delta m = 0.05$. Three populations have been initialized with either one (a), two (b) or three (c) members in set $\mathbf{A}$. The numbers on the contour lines represent the corresponding propensity to punish k.

ture after 200 generations. The map shows the effect of the initial population structure with respect to the fraction of agents in set **A** and **B** (x-axis) and the level of the propensity to punish (y-axis) on the evolution of $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{B}}$. The scale of gray indicates the number of $m + \frac{\Delta m}{2}$ contributing agents after 200 generations in a population with a total size of $n = 4$.



(a) Number (color code) of agents in set **A** after 200 generations



(b) Cumulative P&L (color code) of agents in set **A** over 200 generations

Figure 5.2: Development of the population structure/P&L over 200 generations as a function of the initial population structure (x-axis) and the propensity to punish k (y-axis).

We now add an additional layer into the population structure in order to allow for heterogeneity with respect to the propensity to punish k. Therefore, set **A** is divided into two additional sets denoted by **P** and **Q**. Agents in the set **P** display a propensity to punish of k_P and agents in **Q** punish with an intensity determined by k_Q. $\tilde{\mathbf{P}}$ and $\tilde{\mathbf{Q}}$ denote the corresponding fractions measured relative to the size of parent-set **A**. The resulting population structure is depicted schematically in figure 5.3. The individual P&L structure of agents

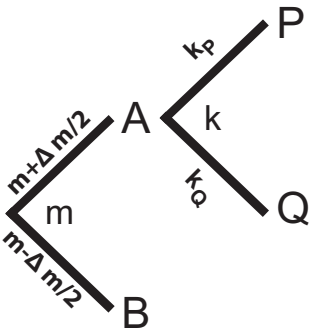


Figure 5.3: Two layer heterogeneity structure: sets **A** and **B** differ with respect to the contribution level m. Set **A** is once more divided into a set **P** of agents with a propensity to punish k_P and a set **Q** with k_Q.

in the resulting sets **P**, **Q** and **B** is given by:

$$P\&L_{\mathbf{P}} := n \cdot \tilde{\mathbf{A}} \cdot \frac{g \cdot \frac{\Delta m}{2}}{n} - n \cdot \tilde{\mathbf{B}} \cdot \frac{g \cdot \frac{\Delta m}{2}}{n} + g \cdot m - (m + \frac{\Delta m}{2}) - n \cdot \tilde{\mathbf{B}} \cdot k_P \cdot \Delta m \tag{5.5}$$

$$P\&L_{\mathbf{Q}} := n \cdot \tilde{\mathbf{A}} \cdot \frac{g \cdot \frac{\Delta m}{2}}{n} - n \cdot \tilde{\mathbf{B}} \cdot \frac{g \cdot \frac{\Delta m}{2}}{n} + g \cdot m - (m + \frac{\Delta m}{2}) - n \cdot \tilde{\mathbf{B}} \cdot k_Q \cdot \Delta m \tag{5.6}$$

$$P\&L_{\mathbf{B}} := n \cdot \tilde{\mathbf{A}} \cdot \frac{g \cdot \frac{\Delta m}{2}}{n} - n \cdot \mathbf{B} \cdot \frac{g \cdot \frac{\Delta m}{2}}{n} + g \cdot m - (m - \frac{\Delta m}{2}) - n \cdot \tilde{\mathbf{A}} \cdot \left(\tilde{\mathbf{P}} \cdot k_P \cdot r \cdot \Delta m + \tilde{\mathbf{Q}} \cdot k_Q \cdot r \cdot \Delta m\right) \tag{5.7}$$

The different terms in equations 5.5-5.6 are analogously defined as in equation 5.1. The structure of equation 5.7 corresponds to equation 5.2 except for the last term which reflects the costs of being punished proportional to the size of set **P** with the propensity to punish k_P and the size of set **Q** with the propensity to punish k_Q.
The dynamics of the agents in set **P** are defined by the recurrence equation

$$\tilde{\mathbf{P}}' = \frac{\tilde{\mathbf{A}} \cdot \tilde{\mathbf{P}} \cdot P\&L_{\mathbf{P}}}{\tilde{\mathbf{A}} \cdot \tilde{\mathbf{P}} \cdot P\&L_{\mathbf{A}} + \tilde{\mathbf{A}} \cdot \tilde{\mathbf{Q}} \cdot P\&L_{\mathbf{Q}}} \tag{5.8}$$

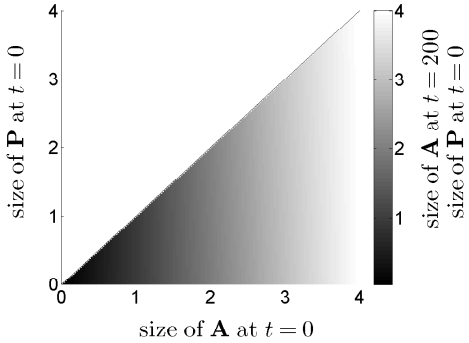
in which $\tilde{\mathbf{P}}'$ represents the fraction of agents of in set **P** in the next generation. The absolute number of agents in the sets **P** and **Q** can be obtained by

$$\begin{aligned} &\text{size of } \mathbf{P} = n \cdot \tilde{\mathbf{A}} \cdot \tilde{\mathbf{P}} \\ &\text{size of } \mathbf{Q} = n \cdot \tilde{\mathbf{A}} \cdot \tilde{\mathbf{Q}} = n \cdot \tilde{\mathbf{A}} \cdot (\mathbf{1} - \tilde{\mathbf{P}}) \ . \end{aligned} \tag{5.9}$$

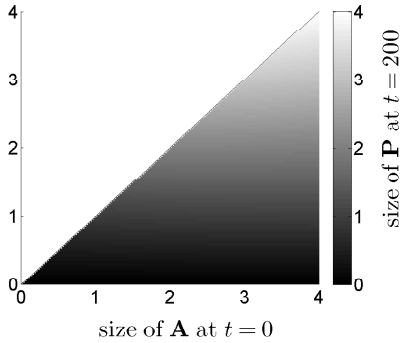
Figures 5.4 - 5.6 show the structure of the 2-layered (**A**/**P**) population after 200 generations for different fixed values of the propensities to punish k_P and k_Q and different intensities of heterogeneity, Δm, in the cooperation level as a function of the initial size of **A** (x-axis) and **P** (y-axis). The size of **P** has to be less or equal to the size of its superset **A**. Hence, only values below the indicated diagonal can be considered to be meaningful.

Figures 5.4-5.6 demonstrate that a more heterogenous population develops a potential to evolve to higher levels of cooperation with time rather than a population with a lower variability in their contributions and in their propensity to punish. The following three scenarios illustrate the effect of heterogeneity either in the contributions, in the propensity to punish or in a combination of both:

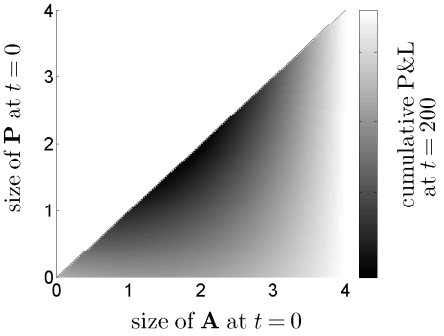
- **Scenario 1 - heterogeneity in punishment**: The population in figure 5.4 is quasi-homogeneous with respect to the contributions ($\Delta m \approx 0$), but heterogeneous in their propensity to punish, i.e. $k_p \gg k_q$ and $\frac{k_P + k_Q}{2} = \frac{1}{4}$. In this scenario the initial population structure is pre-



(a) Size of **A** (color code) after 200 generations

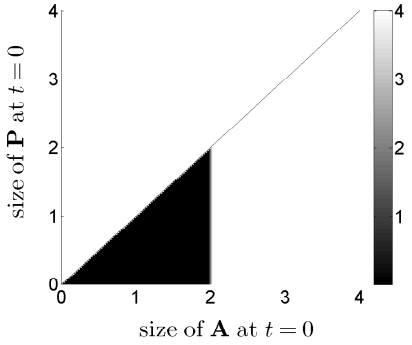


(b) Size of **P** (color code) after 200 generations

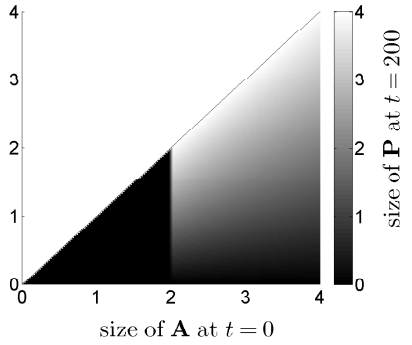


(c) Cumulative P&L (color code) of agents in set **A** over 200 generations

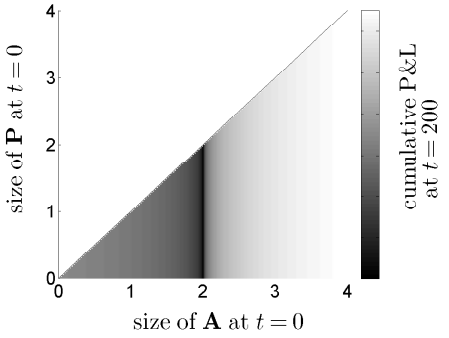
Figure 5.4: Evolution of the number of agents in sets **A** and **P** and the cumulative P&L after 200 generations. The population has been initialized with a propensity to punish of $k_P = 0.4$ and $k_Q = 0.1$ and a heterogeneity in the contributions of $\Delta m = 0.0005$. The values on the x (size of $\mathbf{A} = \mathbf{P} \cup \mathbf{Q}$) and y-axis (size of **P**) determine the initial population structure in the first generation. The color code indicates the population structure with respect to size of **A** (a), **P** (b) and the cumulative P&L (c) of the population after 200 generations.



(a) Size of **A** (color code) after 200 generations

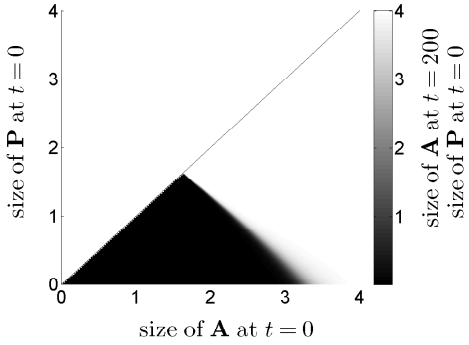


(b) Size of **P** (color code) after 200 generations

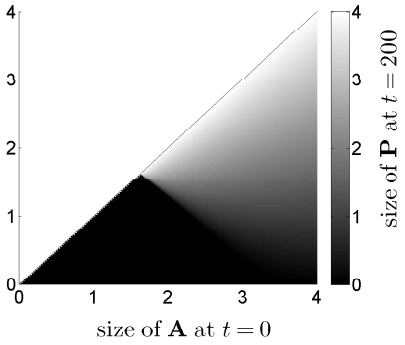


(c) Cumulative P&L (color code) of agents in set **A** over 200 generations

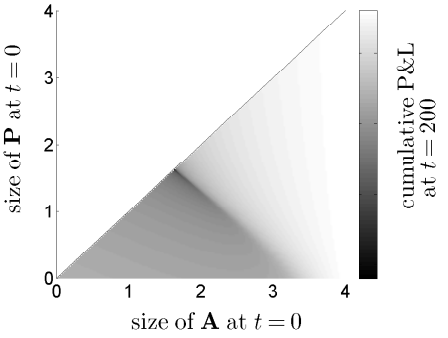
Figure 5.5: Evolution of the number of agents in sets **A** and **P** and the cumulative P&L after 200 generations. The population has been initialized with a propensity to punish of $k_P = k_Q = 0.25$ and a heterogeneity in the contributions of $\Delta m = 0.05$. The values on the x (size of $\mathbf{A} = \mathbf{P} \cup \mathbf{Q}$) and y-axis (size of **P**) determine the initial population structure in the first generation. The color code indicates the population structure with respect to size of **A** (a), **P** (b) and the cumulative P&L (c) of the population after 200 generations.



(a) Size of **A** (color code) after 200 generations



(b) Size of **P** (color code) after 200 generations



(c) Cumulative P&L (color code) of agents in set **A** over 200 generations

Figure 5.6: Evolution of the number of agents in sets **A** and **P** and the cumulative P&L after 200 generations. The population has been initialized with a propensity to punish of $k_P = 0.4$ and $k_Q = 0.1$ and a heterogeneity in the contributions of $\Delta m = 0.05$. The values on the x (size of $\mathbf{A} = \mathbf{P} \cup \mathbf{Q}$) and y-axis (size of **P**) determine the initial population structure in the first generation. The color code indicates the population structure with respect to size of **A** (a), **P** (b) and the cumulative P&L (c) of the population after 200 generations.

served and stable across time, which can be observed in subfigures (a) and (b) by the horizontal (a) and the vertical (b) color gradient.

- **Scenario 2 - heterogeneity in cooperation**: The structure depicted in figure 5.5 corresponds to a population that shares a common and uniform propensity to punish, i.e. $k_P = k_Q$, and, on the other hand is characterized by heterogeneity with respect to the level of cooperation, i.e. $\Delta m > 0$. This results in a clear fragmentation of the population structure after 200 generations that comes about with a tipping-point-like characteristic at $\tilde{\mathbf{A}}(t = 0) = 0.5$ (c.f. subfigure 5.5 (a)). Introducing heterogeneity in the contributions thus induces a strong divergent dynamic into the evolution of the population structure. In this respect, the initial fraction $\tilde{\mathbf{A}}$ of agents who deviate positively by contributing Δm more than the remaining agents in set $\mathbf{B}$ determines the evolutionary outcome.

- **Scenario 3 - heterogeneity in punishment and cooperation**: The evolution of the population structure in figure 5.6 is subject to both heterogeneity in the contributions and in the propensity to punish. The co-evolution of low and high contributing agents along with an additional variability in the punishment behavior ($\mathbf{P}$ vs. $\mathbf{Q}$) reduces the initial minimum fraction $\tilde{\mathbf{A}}$ of agents required to have set $\mathbf{A}$ dominating the population. However, this is only possible if the fraction of strong punishers ($\tilde{\mathbf{P}}$) is sufficiently large at $t = 0$. This results in the triangle shape of the population structure after 200 generations.

We now state our first main result of this chapter:

Result 5.1: *Heterogeneity both in the level of cooperation and in the propensity to punish is an indispensable precondition for the evolutionary emergence of cooperative behavior in environments which are characterized by a social dilemma component such as the analyzed public goods game.*

In the following section, we introduce and discuss different variants of heterogeneity-preserving mechanisms and explore them numerically using the evolutionary simulation model presented in chapter 3.

5.3 Heterogeneity preserving mechanisms in evolutionary dynamics

Under evolutionary dynamics the heterogeneity of the traits within a population can only be maintained by one or both of the following two mechanisms: genetic drift and mutations. In contrast, adaptation, selection and the cross-over, e.g. by sexual reproduction, reduce the heterogeneity of traits in a population across time. As argued and discussed in the previous chapters, evolutionary dynamics operate not only on the biological level but also on higher levels, e.g. in the form of the co-evolution of genes, culture and social norms. As a consequence, we use the term "trait" as an equivalent for both biological concepts such as the "genotype" or "genes" but also for concepts that apply to different levels and on different scales in the hierarchy of evolutionary processes such as an evolving cultural heritage. Starting from this perspective, we discuss and analyze different types of heterogeneity-preserving mechanisms that have been and still are subject to a lively debate in the literature (Wade, 1978; Wilson, 1983; Boyd and Richerson, 1990; West, Griffin, and Gardner, 2007; Ichinose and Arita, 2008; van den Bergh and Gowdy, 2009; Leigh, 2010). All these processes belong to the class of group or multi-level selection processes. Evolutionary dynamics with a "multi-layered" hierarchy of selection processes do not necessarily induce an upward or downward causation between the different layers. Moreover, this leads to more complex and entangled interactions with non-linear dependencies and dynamics. Multi-level selection operates across and between all biological and organizational units ranging from the individual genotype all the way up to groups, communities and cultural structures in populations(Bergstrom, 2002; Henrich, 2004; Bowles, 2003). If the evolutionary dynamics on multiple scales, e.g. on the individual and on the group level, are taken into account, an important step towards a better understanding of complex and nested organizational structures, co-evolutionary processes and emergent properties is made. Better insights into these multi-scale phenomena are gained, which ultimately contributes to unravel the puzzle of cooperation. A comprehensive and well-structured overview about multi-level selection and its application to social science and in particular to economics is presented in (van den Bergh and Gowdy, 2009).

As shown in the previous sections of this chapter, punishment is only effective and promotes cooperation if the two opposing mechanisms - drift/mutation vs. selection/cross-over - are balanced, i.e. if a certain level of heterogeneity is present in the traits of a population. Multi-level selection with its various intertwined evolution processes can add this required heterogeneity into the traits of a population by means of multifaceted evolutionary dynamics. In this section, we look specifically into two types of multi-level selection processes: inter and intrademic multi-level selection. Interdemic multi-level occurs among partially isolated sub-populations. Each sub-population is subject to a locally determined selection pressure whose strength varies individually per sub-population. The selection pressure among local sub-populations contributes to the heterogeneity of the traits between the sub-populations, i.e. the selection operates on the basis of "demographic" heterogeneity. Typical mechanisms of interdemic selection are dispersal processes, colonization and migration as well as founder effects (Wade, 1978, 1982; Ichinose and Arita, 2008; Rogers and Ehrlich, 2008). In contrast, intrademic selection occurs among sub-populations that are only isolated during some specific period(s) of their lifetime. Sub-populations are genetically subdivided but experience a uniform selection pressure that is determined on the population level. Heterogeneity in the traits of the population arises due to the periodicity in the heredity transmission of traits that occurs on a population-wide scale. The frequency of the heredity transmission constitutes a trade-off between trait heterogeneity and the population heritability. Kinship structure and social interactions are exemplary mechanisms that are characteristic for intrademic multi-level selection (Smith, 1964; West, Pen, and Griffin, 2002; Alger, 2010; Waibel, Floreano, and Keller, 2011; Mahajan, Martinez, Gutierrez, Diesendruck, Banaji, and Santos, 2011).

Both characteristic forms of multi-level selection are explored and verified by means of the evolutionary simulation model that was first introduced in chapter 3. In the competitive resource-limited world of the simulation model, agents adapt and evolve their cooperation level and their propensity to punish in response to standard evolutionary dynamics while playing a public goods game with punishment. If not otherwise specified, the agents' fairness preferences in the following are fixed to an aversion against disadvantageous in-

equitable outcomes (dynamics C - see chapter 3 subsection 3.2.2), which was identified in the previous chapters to be the predominant other-regarding strategy among subjects in three lab experiments. Analogously to the simulation runs conducted in chapter 3 the model is initialized much in the same way as described in section 3.2.1: Each monetary unit (MU) that is contributed to the public good returns a gain of $g = 1.6$ MUs back to the group. After the agents contributed $m_i \geq 0$ MUs, they learn about the contributions of the other agents and punish negative deviators according to their individual propensity to punish k_i proportional to their deviation as defined in section 3.2.1. The punishment efficiency is always fixed to a value of $r = 3$, i.e. for each MU spent to punish an agent, the punished agent loses 3 MUs. In the definition of our model, a sub-population is defined as a group of agents who jointly contribute to one public good. The entire population thereby is composed of multiple groups. Groups are indexed by j and agents are indexed by (i, j). For example the index $(4, 2)$ corresponds to the second agent in group 4. In total, the population consists of b groups, each of them with n members. In the analytical model presented in chapter 2 as well as the numerical simulation model in chapter 3, all agents are part of the same isolated group as each population essentially consists only of one group. As a consequence, all agents inherit their traits from the same local pool and interact through the same public good. This results in fitness being defined in relative terms among the agents. In other words: increasing the absolute fitness of agents in an isolated group does not affect the evolutionary dynamics (Wilson, 2004). However, if a population is composed of multiple spatial or temporary partly isolated groups this gives rise to fitness differences between groups which are caused by the heterogeneity in the traits of the co-evolving groups. An experiment on the relative fitness between groups and group competition is e.g. presented in (Burton-Chellew, Ross-Gillespie, and West, 2010).

In section 5.3.1 we investigate the effect of intrademic multi-level selection in the presence of (altruistic) punishment on the emergence and the evolution of cooperative behavior. In section 5.3.2 we analyze the mutual interaction of interdemic group selection and punishment and their impact on the evolutionary emergence of cooperation. In particular, we look into the effect of migration patterns between partially isolated co-evolving groups.

5.3.1 Variants of intrademic multilevel-selection

This section provides an analysis of the effect of intrademic multi-level selection in a population of agents who face a social dilemma in the form of a public goods game with punishment. Intrademic multi-level selection occurs in a population of temporary isolated sub-populations, i.e. groups. Being characteristic for intrademic models, the selection pressure is defined based on a global selective environment. This means that each individual in each group faces the same intensity of selection pressure which is determined on the population-wide level, i.e. across groups. As in the evolutionary simulation model presented in chapter 3, selection pressure comes in the form of a fixed consumption which is homogeneous across all groups and agents and is defined by the population's average total P&L:

$$c(t) = \text{Max}[\frac{1}{m \cdot n} \sum_{j} \sum_{i} \hat{s}_i(t); c_{\text{fix}}] \tag{5.10}$$

$\hat{s}_i(t)$ is defined by equation (3.4) in chapter 3.

The co-evolution of temporary isolated groups which are subject to a population-wide selection pressure contributes to the between group heterogeneity in the trait pool. As opposed to this, the frequency and the structure of the heredity of traits controls the transition from between group heterogeneity to within group heterogeneity. Ultimately, the interaction of punishment together with the heterogeneity in a local group promotes the emergence of cooperative behavior as discussed in detail in section 5.2: heterogeneous groups coordinate and evolve their cooperation depending on their propensity to punish k. For example, if the average propensity to punish k is less than $k < 0.25$ (cf. section 2.3.3 and 3.3) in a group of $n = 4$ agents who vary in their initial contributions m_i, the group converges towards the lowest initially present individual contribution $min(m_i)$. In other words, the traits of the least contributing agent most likely spread and start to dominate in the population. The corresponding dynamics have been illustrated in figures 5.1 and 5.2 of section 5.2.

In the following, we analyze two different heredity-mechanisms that account for the transmission of traits among successive generations within and between groups. In general, a population that consists of two or more groups can

experience two different scenarios that cause an alternation of generations with a passing on of traits to the offsprings:

- **E1 individual extinction:** This scenario reflects the case in which 1 up to $n-1$ agents of a group with size n go extinct within one period, i.e. at least 1 agent survives and has the potential chance to father new offsprings by passing on her traits to the new generation[2].
- **E2 group extinction:** This scenarios reflects the case in which all n agents of a group go extinct within one period, i.e. there are no survivors that could sustain the traits that were characteristic for the group and pass them on to a successive generation.

As our model bases on constant group and population sizes, i.e. population dynamics are not explicitly modeled, both scenarios require a replacing of the deceased agents by a succeeding generation that inherits a set of new traits. At this point the rate of heredity comes into play: For the individual extinction scenario (E1) the traits of the reborn agents can either be sampled only from the surviving part of the specific local group (low rate of heredity) or from the trait pool of global population (high rate of heredity), i.e. across groups. Similarly, in case of an group extinction (E2), the traits of the reborn agents can either be sampled from the global populations (high rate of heredity) or simply be reset to the starting conditions at the beginning of the simulation (low rate of heredity). In the following we vary the rate of heredity to better understand the effect of different heredity transmission mechanisms on the evolution of cooperation. The parts printed in cursive characters reflect the variable dimension at a time. We essentially investigate the following four variants of heredity transmission:

- **F1: *local-group* heredity** In case of individual extinction (E1) traits of reborn agents are sampled with probability h either from the trait pool of the global population, i.e. across groups, or with probability $1-h$ they are sampled from the local pool of the surviving group fellows. If

[2] We assume a minimum number of only 1 survivor as a sufficient condition for the survival of a group, even though sexual reproduction and cross-over requires the mating of at least two surviving agents. Our results are robust to variations of this assumption.

the entire group goes extinct (E2), all agents of that group are reset to the initial state at the beginning of the simulation, i.e. $m_i(t) = 0$ and $k_i(t) = 0$. If $h = 0$ the model is comparable to a propagule pool model (Wade, 1978; van den Bergh and Gowdy, 2009).

- **F2: *local-group* and population heredity** In case of individual extinction (E1) this variant is equivalent to F1, however it differs for the case of group extinction (E2): Here, traits of reborn agents are always sampled from the trait pool of the global population, i.e. sampled across all groups. For $h = 0$ this scenario is similar to the class of "migrant pool" models (Wade, 1978; Ichinose and Arita, 2008; van den Bergh and Gowdy, 2009). One well known model of the migrant pool class is the haystack model introduced by Maynard Smith (Smith, 1964). The concept of "migrant-pool" does not mean that agents migrate between groups. Moreover surviving groups send out emigrants who form a "migrant-pool". The members of the migrant pool subsequently recolonize the locations of groups that became extinct.

- **F3: local-group and *population* heredity** In case of individual extinction (E1), dead agents are always replaced by agents whose traits are sampled only from the pool of the local group. For the case that all agents of one group die at once (group extinction - E2), the traits of the reborn group are sampled with probability h from the trait pool of the global population, i.e. are sampled across groups. With probability $1 - h$ groups stay isolated and the traits of the successive generation is reset to the initial state $m_i(t) = 0)$ and $k_i(t) = 0$.

- **F4: *local group and population* heredity** In this variant the rate of heredity is varied for both scenarios E1 and E2 simultaneously. This means that the traits of reborn agents are sampled with probability h from the global population, i.e. across groups, independently of the occurrence of either scenario E1 or E2. With probability $1-h$ the traits of reborn agents are sampled from the pool of survivors in the local group - in case of an individual extinction (E1) - or reset to the initial state ($k_i(t) = 0$,$m_i(t) = 0$) - in case of group extinction (E2). If $h = 1$ the heredity mechanism is equivalent to a kinship-like structure (Lehmann,

Keller, West, and Roze, 2007; Leigh, 2010; Wade, 1978; van den Bergh and Gowdy, 2009).

As mentioned above, we modify the evolutionary simulation model introduced in chapter 3 to analyze the effect of changing rates and structures of trait transmission for both scenarios E1 and E2. Therefore the model is set up with a population of agents consisting of $b = 8$ groups each of them with $n = 4$ members. Each group plays a public goods game with punishment separately. The per capita gain in each group corresponds to $\frac{g}{n} = 0.4$ MUs. Clearly this corresponds to a social dilemma situation in which defecting, i.e. non-contributing with $m_i = 0$, pays out. The population is initialized with all $k_i(0) = 0$ and $m_i(0) = 0$ at the beginning of the simulation. As stated above the adaptation dynamics has been fixed to disadvantageous inequity aversion (cf. dynamics C in section 3.2.2) for all agents in all groups. The replicator dynamics, i.e. selection, cross-over and mutation occur analogously to the rules defined in 3.2.3, however we now vary the rate and the structure of heredity transmission both for the case of group extinction (E2) and for the individual extinction (E1) according to the 4 variants F1-F4 presented above. Figure 5.7 shows the resulting average level of cooperation m (left plot) and the propensity to punish k (right plot) in the groups after 1 million simulation periods as a function of the rate of heredity for the 4 variants F1-F4. To obtain the curves we run 256 system realization with a fixed probability value from the set of 100 values in $[0, 0.01, 0.02, ..., 1]$. For isolated groups (propagule like model - F1 with $h = 0$) the population-wide defined selection pressure (cf. equation 5.10) remains too strong to allow for "speciation" and the emergence of heterogeneity among groups in the population. Instead, one group starts to dominate and contains all of the others. Once a group reached this point, it does not require to further improve its level of cooperation and thus m remains relatively weak for all $h \in [0, 1]$ in this variant. Variants F2 and F3 complement each other and show the effect of the different heredity rates and structures on the evolution of cooperation and punishment. In case that $h = 0$ in F2 and $h = 1$ in F3 both variants are equivalent and correspond to the migration pool model. For all other values of $0 < h < 1$ the plots show nicely the transition between the two marginal scenarios, i.e. the migrant-pool and the isolated group scenario (propagule pool). The sensitivity of the level of cooperation

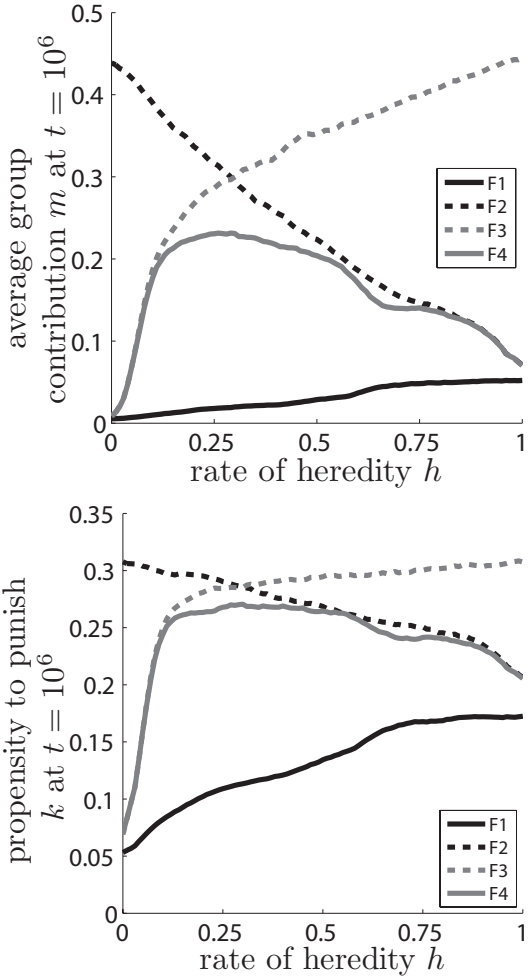


Figure 5.7: Average group contribution m (upper plot) and propensity to punish (lower plot) after 1 million simulation periods as a function of the rate of heredity transmission for all 4 variants F1-F4 described above. The curves are computed based on 256 system realizations for each probability h between $[0, 1]$ fixed for increments of 0.01.

with respect to the rate of heredity rests on the fact that the transmission of traits between groups transforms the evolved between-group heterogeneity into within-group heterogeneity. Ultimately, the evolved propensity to punish in combination with the induced within-group heterogeneity gives rise to the emergence of cooperation via the mechanism presented in section 5.2.

Finally, F4 reflects a mixture of the variants F2 and F3 and reveals the existence of an optimal rate and structure of heredity in an evolutionary population that mixes approximately in 30% of the cases on a population-wide scale and in 70% of the cases on a local-group-scale. These results contribute and provide further food for thoughts to the existing empirical work and discussions on co-residential patterns, kinship-structure and the level of cooperation in hunter-gatherer societies and among humans (Hill, Walker, Bozicevic, Eder, Headland, Hewlett, Hurtado, Marlowe, Wiessner, and Wood, 2011).

We conclude with the second main result of this chapter:

Result 5.2: *The evolution of cooperation in intrademic multi-level selection models is controlled by the rate and structure of the active heredity mechanisms. This mechanisms must provide (i) a sufficient "breeding-ground" for the emergence of inter-group heterogeneity and (ii) must transform inter-group heterogeneity into within-group heterogeneity in an appropriated ratio so that punishment stays active and promotes the emergence of cooperation.*

In the next section we turn to mechanisms of interdemic multi-level selection and explore them using our evolutionary simulation model.

5.3.2 Variants of interdemic multi-level selection

This section analyzes mechanisms of interdemic multi-level selection and their effect on the evolution of cooperative behavior in a social dilemma environment. Interdemic multi-level selection is characterized by a population that consists of isolated groups which are exposed to individually varying intensities of selection pressures. We modify the evolutionary simulation model presented in chapter 3 to include variants of interdemic multi-level selection. In doing so, we specifically look into the effect that the migration of agents between multiple isolated groups causes on the emergence of cooperative behavior (Wade,

1982). Here, "migration" corresponds to a switching of agents between groups in the population, but could also stand for any kind of (social) interaction among otherwise distinct and isolated groups, such as the transmission and adaptation of specific behaviors, cultural norms or - on a higher scale - even the implementation of laws. Migration and group-exchange patterns have also been manifested among ancient hunter-gatherer societies (Hill et al., 2011). Similar to the rate of heredity and its structure (local vs. global), migration transforms between-group heterogeneity into within group heterogeneity. In combination with the opportunity to punish, this gives rise to the emergence of cooperative behavior as shown in section 5.2. In general, there exist two basic variants of migration patterns: targeted migration, in the form of an assorted switching of agents as e.g. presented in (Enquist, 1993; Helbing and Yu, 2009; Aktipis, 2011), or purely random switching. Targeted migration occurs e.g. if humans actively choose the social environment they want to live in, e.g. if they group together along their ethnicities. In contrast, the "island-model" introduced by Wright (Wright, 1943) represents a classical random switching model. In the analysis of our model we focus on the random switching variant of models, as targeted migration models base on ex ante assumptions that inherently induce a trend towards the desired outcome. E.g. if cooperators prefer to be accompanied by other cooperators this already assumes (i) the existence of cooperators and (ii) induces a population structure that in most cases implicitly favors cooperation.

For that reason, we implement a simple "random group-migration" mechanism in our evolutionary simulation model as follows: With probability e two randomly chosen agents, each of them from a different group, are exchanged between the two groups. As our model assumes a constant group size and does not consider population dynamics, we always exchange two agents mutually. We run this modified version of our model with 8 groups and 4 members per group over 1 million time periods with fixed probabilities $e \in [0, 0.02]$ for each increment of 0.0005. Again the per capita return for each of the 8 simultaneously played public goods games is $\frac{g}{n} = 0.4$. At the beginning of the simulation, the population is initialized with only non-punishing non-cooperators, i.e. $k_i(0) = 0$ and $m_i(0) = 0$. The resulting average level of cooperation in

the population after 1 million simulation steps, i.e. $m(t = 1 \cdot 10^6)$, and the propensity to punish, $k(t = 1 \cdot 10^6)$, is shown in figure 5.8.

Figure 5.8 reveals that for a low probability of migration, i.e a low frequency of group switching, a sufficiently large propensity to punish, i.e. $k > k^+$, evolves and cooperation emerges in all groups of the population by the following mechanism: Over a period $\Delta t \gg 0$ the isolated group **A** evolves to a slightly higher level of cooperation when compared to a second isolated group **B**, i.e. between-group heterogeneity emerges such that $m_{\mathbf{A}} > m_{\mathbf{B}}$. This is possible due to the interdemic character of the model, i.e. by the isolation of the groups and the locally determined selection pressures. If now an agent $(\mathbf{B}, i)$ from group **B** is exchanged with an agent $(\mathbf{A}, j)$, agent $(\mathbf{B}, i)$ is forced by the punishment of the agents in group **A** to increase her cooperation until the group becomes homogeneous again (c.f. section 5.2) or agent $(\mathbf{B}, i)$ dies and is replaced by a newborn whose traits are sampled locally from the pool of group **A**. In reverse, agent $(\mathbf{A}, j)$ can cause an increase of the cooperation in group **B** before either the group is homogeneous (c.f. section 5.2) or agent $(\mathbf{A}, j)$ dies and is replaced by a new agent whose traits are sample from the local pool of group **B**. However, if the frequency of switching between groups becomes too large, i.e. the probability $0 \ll e \leq 1$, the population cannot develop a sufficiently large between-group heterogeneity, as the traits are constantly synchronized between groups. In other words: high rates of migration prevent the emergence of heterogeneity between groups. Another way to look at this mechanism is the Parrondo or brownian-ratchet effect (Harmer and Abbott, 2002). The Parrondo effect describes situations in which losing strategies or deleterious effects can combine to win. Here, the random behavior is rooted in the exchange between groups and the asymmetry is brought in by the punishment rule. The interaction between different rates of migration and the effectiveness of multi-level selection has been addressed and discussed in the literature e.g. in (Wade, 1982; Ichinose and Arita, 2008).

We now extend and test the interdemic model with random group migration by combining it with the 4 different heredity mechanisms (F1-F4) introduced in section 5.3.1. We analyze this combination of a interdemic-intrademic model to better understand the required conditions that render the evolutionary emergence of cooperative behavior possible. Therefore, we run the model with

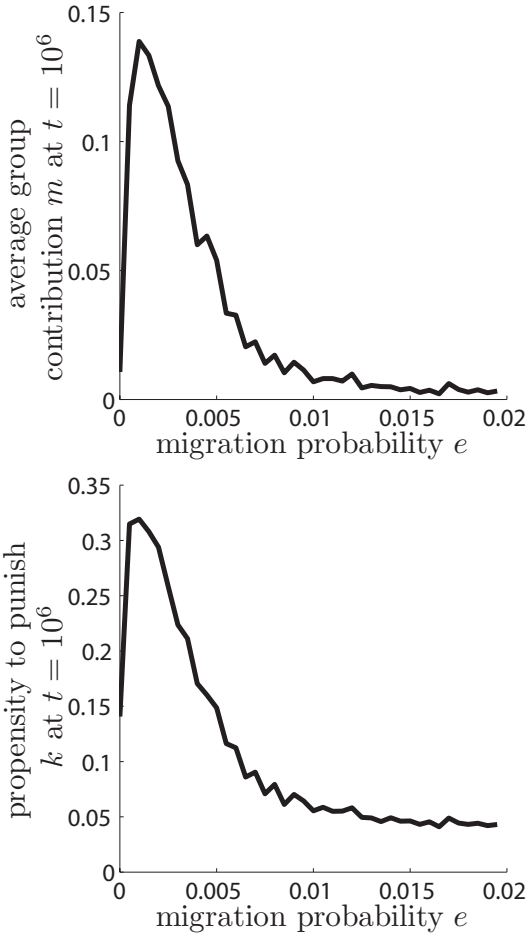


Figure 5.8: Average group contribution $m(t)$ (upper plot) and propensity to punish $k(t)$ (lower plot) after a transient period of 1 million simulation periods as a function of the migration probability e. The values have been calculated based on 256 system realizations for 40 uniformly distributed sample values in the range $[0 \geq e \geq 0.02]$.

different migration probabilities initialized in the range $e \in [0, 0.02]$ and, in parallel, set a specific rate of heredity from $h \in [0, 1]$ for each of the 4 different variants of heredity transmission F1,F2,F3 and F4. In this way, we obtain the 4 figures 5.9-5.12 that show the smoothed[3] average group contribution (left plot) and the propensity to punish (right plot) after 1 million simulation periods as a function of the migration rate e and the rate of heredity transmission h.

The figures 5.9-5.12 indicate and confirm that a population consisting out of $b = 8$ groups, each with $n = 4$ members, evolves to higher levels of cooperation and develops an optimal propensity to punish (c.f. chapters 2 and 3) only for very low migration rates. Furthermore, a transmission of traits across groups in case of individual extinction (E1) (c.f. section 5.3.1) must be absent. This can be seen by the definite peaks around $h = 0$ and $0 < e \sim 0$ in the figures 5.9, 5.10 and 5.12. In contrast, figure 5.11 reveals that in case of group extinction (E2), the rate and the structure of the heredity transmission on a population-wide scale (F3) still allows for the emergence of cooperation. This result is plausible as in interdemic multi-level selection models the selection pressure is determined individually for each group and thus complete groups go less frequently extinct compared to the inherent competition among groups in intrademic models with its globally-defined selection pressure. These findings are in line with the previously demonstrated argument that migration and the heredity transmission of traits (i) need to allow for the emergence of inter-group heterogeneity and (ii) simultaneously must transform inter-group heterogeneity into within-group heterogeneity. Only if this condition is satisfied the feedback provided by punishment can become effective and cause the emergence of cooperative behavior even in social dilemmas (c.f. section 5.2).

[3] We apply the gaussian kernel method presented in (Garcia, 2010) to smooth the surface plot.

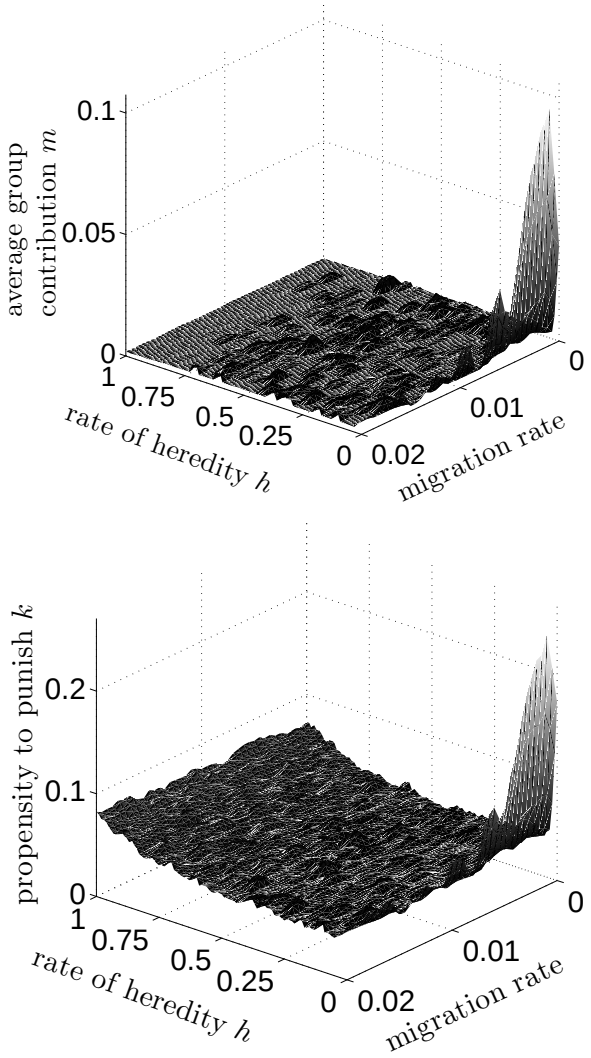


Figure 5.9: Smoothed average group cooperation (upper plot) and propensity to punish (lower plot) after a transient period of 1 million simulation periods as a function of the group exchange probability $e \in [0, 0.02]$ and the rate of heredity $h \in [0, 1]$ for the transmission mechanism F1 described in section 5.3.1.

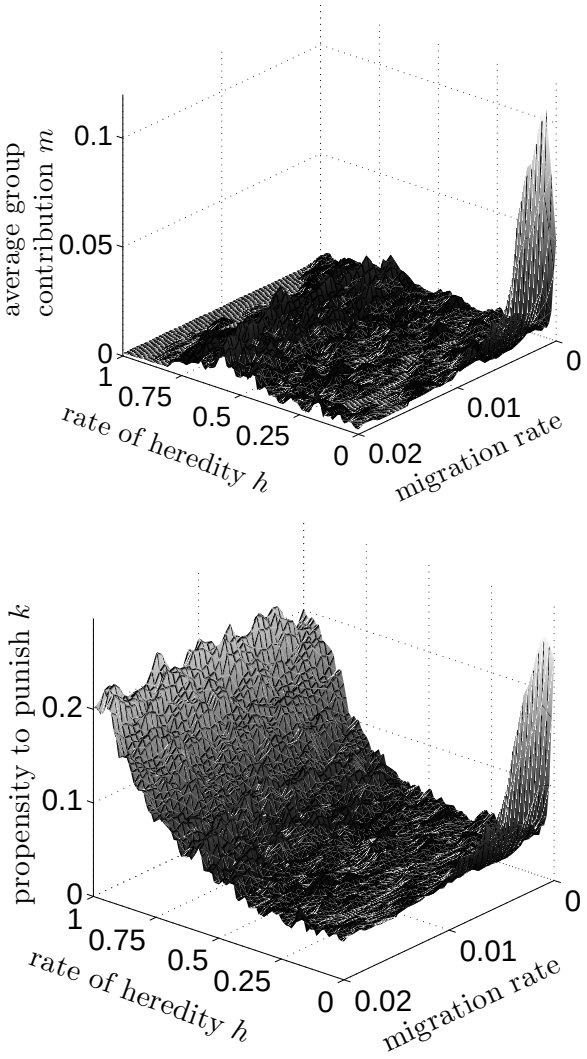


Figure 5.10: Average group cooperation (upper plot) and propensity to punish (lower plot) after a transient period of 1 million simulation periods as a function of the group exchange probability $e \in [0, 0.02]$ and the rate of heredity $h \in [0, 1]$ for the transmission mechanism F2 described in section 5.3.1.

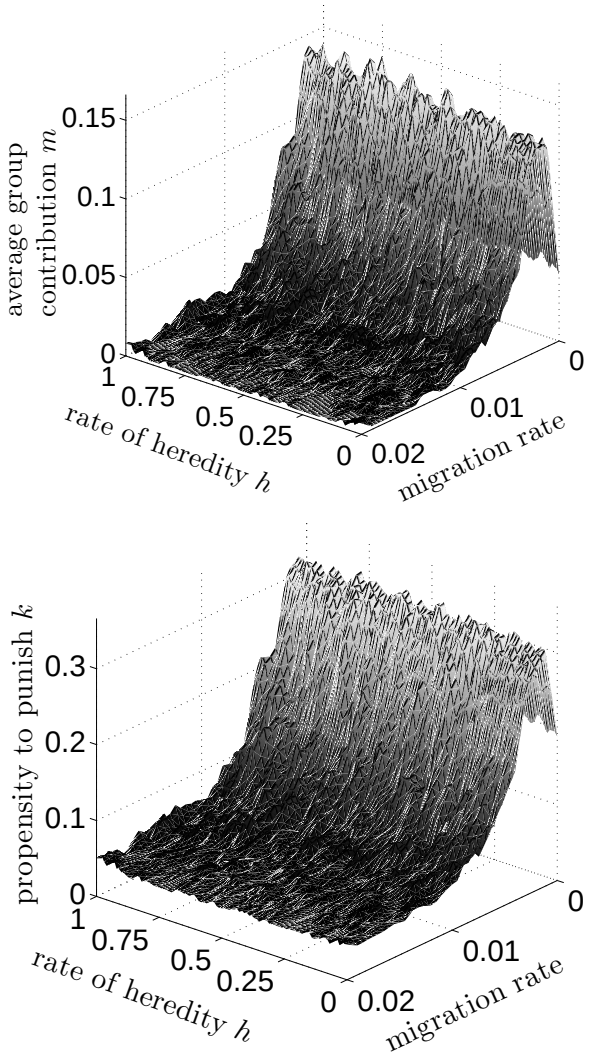


Figure 5.11: Average group cooperation (upper plot) and propensity to punish (lower plot) after a transient period of 1 million simulation periods as a function of the group exchange probability $e \in [0, 0.02]$ and the rate of heredity $h \in [0, 1]$ for the transmission mechanism F3 described in section 5.3.1.

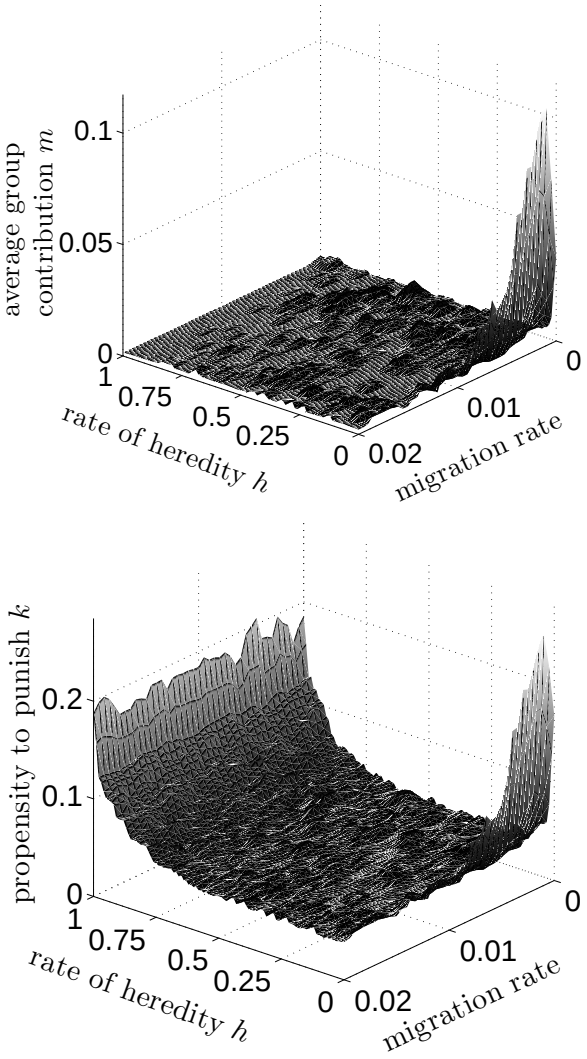


Figure 5.12: Average group cooperation (upper plot) and propensity to punish (lower plot) after a transient period of 1 million simulation periods as a function of the group exchange probability $e \in [0, 0.02]$ and the rate of heredity $h \in [0, 1]$ for the transmission mechanism F4 described in section 5.3.1.

We thus conclude with the third main result of this chapter:

Result 5.3: *Migration in interdemic group models promotes the emergence of cooperative behavior by transforming inter-group heterogeneity into within-group heterogeneity which in combination with optimal punishment finally results in an increase of cooperative behavior. The domain of qualifying rates of migration is narrow, i.e. the mechanism must provide enough time to allow co-evolving groups to become sufficiently distinct.*

5.4 Conclusion

This chapter studied the emergence of cooperation among interacting agents who face a public goods problem with punishment opportunity. In the first part of the chapter, we analyzed and discussed the interplay between heterogeneity in the agents' traits and the feedback provided by punishment and how it affects the long-term evolution of cooperative behavior in the population. This revealed that both heterogeneity in the agents' contributions m_i and in the propensity to punish k_i is beneficial for the emergence of cooperation. Thus, both mechanisms can be considered to play a key role when trying to understand the high levels of cooperation observed in social dilemma situations such as the analyzed public goods game. In the second part of this chapter, we analyzed different mechanisms that are known for generating heterogeneity in the trait pool of an agent population. Specifically, we looked into different variants of multi-level selection in the form of inter and intrademic group selection. Our findings showed that both variants are able to promote sufficient levels of between-group heterogeneity and to transform it at an adequate rate into within-group heterogeneity. The heterogeneity at different scales in the population in combination with the effect of punishment provides a conclusive explanation for the puzzle of cooperation.

6. Conclusion and Outlook

This thesis investigated the co-evolution of fairness preferences, altruistic punishment and cooperation in a population of agents who interact within the framework of a public goods problem. The ultimate goal of the thesis was to explain the high level of cooperation that can be observed among humans, even though this is in contradiction with many behavioral theories such as the rational actor model and the principle of the "survival of the fittest". We approached this puzzle of cooperation from a transdisciplinary perspective bringing together ideas and methods from evolutionary biology, evolutionary psychology, sociology, behavioral economics and complex system science. This resulted in the development of two quantitative models in which agents play a public goods game with punishment: The first model represented an analytical framework that extended the expected utility approach by adding evolutionary dynamics to the behavior of the agents. To ensure the mathematical solvability of the n-person interactions, we had to make simplifying assumptions about the heterogeneity in the population structure. The second model mitigated the assumptions made in the first model by analyzing the n-person interaction of the agents using a numerical simulation approach. The results of both models were compared and verified using data from three previously conducted laboratory experiments on a public goods game with punishment. By means of our models and the empirical data we were able to identify and to verify two important patterns of prosocial behavior that promote the emer-

gence of cooperation: First, natural selection and evolutionary dynamics cause the emergence of fairness preferences in a population of agents in the form of an aversion against disadvantageous inequitable outcomes. Second, an aversion against outcomes that are unfavorable for the own fitness is sufficient to explain the phenomenon of altruistic punishment: Agents who adapt their behavior to avoid situations in which the behavior of others plays to their disadvantage are more likely to survive. This evolutionary drift gives rise to the emergence of (altruistic) punishment behavior by means of the following mechanism: punishing unfair behaving group fellows transforms the social dilemma of the public goods problem into a coordination problem and thus allows prosocial agents to reduce and compensate the fitness advantage that is usually gained by free-riders. In the third part of this thesis, we analyzed the empirically observed cooperation behavior and the associated behavioral dynamics of subjects from three public goods experiments with punishment. We presented an alternative methodology to classify subjects along their intrinsic preferences either to cooperate or to defect based on the observed punishment reactions and complemented the findings presented in previous studies (Fischbacher, 2001; Houser and Kurzban, 2003; Bardsley and Moffatt, 2007; Herrmann and Thoeni, 2009; Rustagi et al., 2010). Furthermore, we verified that the opportunity to punish indeed induces a coordination dynamic that over time leads to a homogeneous behavior within groups. In the last part of this thesis we analyzed the effect of punishment in a public goods problem setting on the evolutionary dynamics of a population of agents who display a constant heterogeneity in their cooperation behavior. This confirmed that agents coordinate their behavior with time along the level of cooperation either of the least or the highest contributing agent in the group, depending on the intensity of the present propensity to punishment. Subsequently, we extended and modified our numerical simulation model to include mechanisms that maintain heterogeneity in the population of agents. In particular, we looked into the co-evolutionary dynamics of multiple groups of agents that were subject to different forms of multi-level selection. This ultimately allowed us to reveal how cooperation can be promoted in a competitive and resource-limited environment that is subject to material self-interest: the interplay of a sufficiently strong propensity to punish and the constant presence of within-group

heterogeneity has been identified to be the driving force behind the emergence of cooperative behavior. Even though punishment tends to harmonize the behavior of agents in a group, within-group heterogeneity can be sustained by the interplay of the following two mechanism: the co-evolution of multiple groups gives rise to the emergence of between group heterogeneity. The different forms of multi-level selection transform the between-group heterogeneity into within-group heterogeneity. Figure 6.1 illustrates the identified processes schematically. In the following we comment on our results and modeling assumption and highlight potentially interesting areas for future research.

We motivated that many decision settings can be represented as a public goods problem. Thus, we focused on a public goods game with punishment opportunity in both models of this thesis. However, other social dilemma settings and variants of the public goods problem, e.g. by providing the option to par-

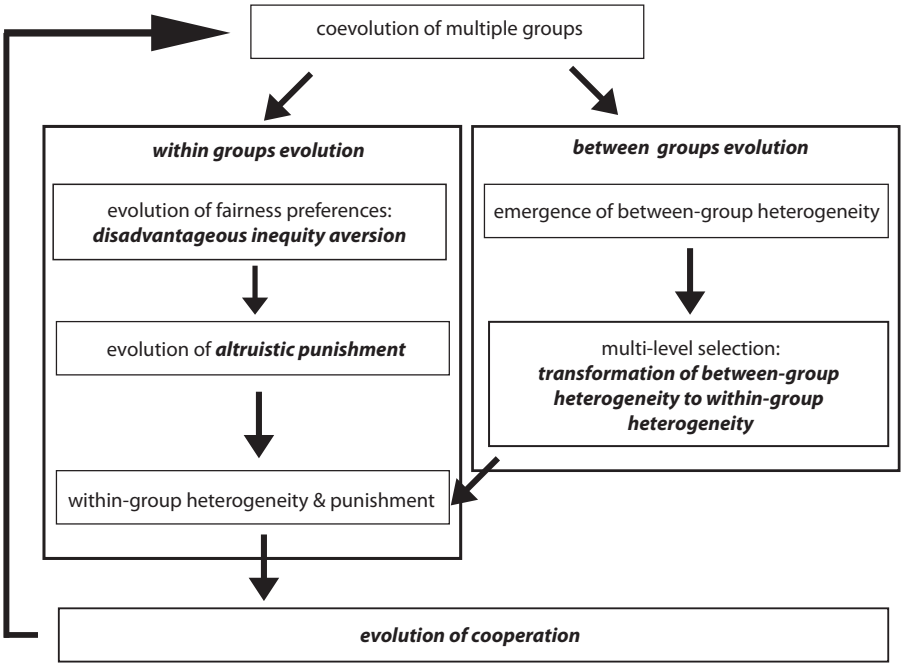


Figure 6.1: Schematic view of the chain of reasoning within our approach to explain the evolution of cooperation.

ticipate only voluntarily, have led to interesting results (Hauert et al., 2002; Brandt et al., 2006; Hauert et al., 2008). An analysis of these variants within our framework could lead to additional insights into the puzzle of pro-sociality. Furthermore, our two models base on the assumption that individuals on average only punish group fellows who contributed less than the own contribution. Even though the empirical data in figure 2.1 strongly support this hypothesis, they also show that a fraction of subjects from the three experiments displayed spiteful punishment behavior. In other words, they punished other individuals although they contributed more than themselves. This behavioral pattern is not considered by our models and could be added in a future extensions. Another extension to our model could be the co-evolution of more than two different other-regarding adaptation dynamics (A-F) (c.f. section 3.2.2) at the same time. This would allow to analyze the evolution of a more realistic population structures that consist of agents with various heterogeneous preferences of inequity and inequality aversion. Furthermore, the results presented in the last chapter only base on a few known mechanisms that generate heterogeneity within the population. Beside these multi-level selection mechanisms, other heterogeneity preserving and generating processes should be analyzed. For instance, introducing the possibility to allow for strategic short-term behavior among the agents could lead to different interesting evolutionary dynamics within the population.

In conclusion, we believe that the combination of empirical research and simulation models can provide deeper insights into the evolutionary roots of human behavior. E.g. more realistic setups in which agents play several games simultaneously so as to mimic more realistic situations may provide further insights into the nature of our prosocial behavior. With regard to the importance of understanding social peer-interactions and fostering prosocial behavior, our approach provides a flexible and powerful methodology to answer many remaining research questions. For instance, analyzing the influence of other feedback mechanisms beside the peer punishment structure implemented in our models or looking into different group structures and varying selection pressures, may lead to new insights and tools that could help to stabilize societal systems in an ever faster changing world. The recent developments of social and political change in North Africa and the Middle East, but also the

riots in the United Kingdom revealed the power of the people's perception of "fairness" and "unfairness". This highlights the importance of being able to design and implement specific social incentive mechanisms and to better understand the central role of fairness and feedback mechanisms, such as altruistic punishment, for the cooperation in social groups, communities and societies.

Reference

Aktipis, C. A., Apr. 2011. Is cooperation viable in mobile organisms? simple walk away rule favors the evolution of cooperation in groups. Evolution and Human Behavior.

Alger, I., Sep. 2010. Kinship, incentives, and evolution. American Economic Review 100 (4), 1725–58.

Allport, G. W., 1985. The historical background of social psychology. Handbook of Social Psychology. Random House, Ch. 1, pp. 1–1159.

Almas, I., Cappelen, A. W., Sorensen, E. O., Tungodden, B., May 2010. Fairness and the development of inequality acceptance. Science 328 (5982), 1176–1178.

Anderson, C. M., Putterman, L., January 2006. Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. Games and Economic Behavior 54 (1), 1–24.

Anderson, L. R., Mellor, J. M., Milyo, J., June 2008. Inequality and public good provision: An experimental analysis. Journal of Socio-Economics 37 (3), 1010–1028.

Andreoni, J., Harbaugh, W., Vesterlund, L., 2003. The carrot or the stick: Rewards, punishments, and cooperation. The American Economic Review 93 (3), 893–902.

Andreoni, J., Miller, J., 2002. Giving according to garp: An experimental test of the consistency of preferences for altruism. Econometrica 70 (2), 737–753.

Arrow, K. J., September 1970. Social Choice and Individual Values, Second edition (Cowles Foundation Monographs Series), 2nd Edition. Yale University Press.

Arthur, W. B., 1994. Inductive Reasoning and Bounded Rationality. The American Economic Review 84 (2), 406–411.

Axelrod, R., October 1985. The Evolution Of Cooperation. Basic Books.

Axelrod, R., September 1997. The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration: Agent-based Models of Competition and Collaboration (Princeton Studies in Complexity). Princeton University Press.

Axelrod, R., Hamilton, W. D., 1981. The Evolution of Cooperation. Science 211 (4489), 1390–1396.

Baldassarri, D., Grossman, G., Jul. 2011. Centralized sanctioning and legitimate authority promote cooperation in humans. Proceedings of the National Academy of Sciences 108 (27), 11023–11027.

Bardsley, N., Moffatt, P., Mar. 2007. The experimetrics of public goods: Inferring motivations from contributions. Theory and Decision 62 (2), 161–193–193.

Bardsley, N., Sausgruber, R., October 2005. Conformity and reciprocity in public good provision. Journal of Economic Psychology 26 (5), 664–681.

Bell, A. V., Richerson, P. J., McElreath, R., October 2009. Culture rather than genes provides greater scope for the evolution of large-scale human prosociality. Proceedings of the National Academy of Sciences 106 (42), 17671–17674.

Berger, U., September 2010. Learning to cooperate via indirect reciprocity. Games and Economic Behavior.

Bergstrom, T. C., 2002. Evolution of social behavior: Individual and group selection. The Journal of Economic Perspectives 16 (2), 67–88.

Bernhard, H., Fischbacher, U., Fehr, E., 2006. Parochial altruism in humans. Nature 442 (7105), 912–915.

Bernheim, B. D., 1994. A Theory of Conformity. The Journal of Political Economy 102 (5), 841–877.

Black, D., 1948. On the Rationale of Group Decision-making. The Journal of Political Economy 56 (1), 23–34.

Bochet, O., Page, T., Putterman, L., May 2006. Communication and punishment in voluntary contribution experiments. Journal of Economic Behavior & Organization 60 (1), 11–26.

Bolton, G. E., Ockenfels, A., 2000. ERC: A Theory of Equity, Reciprocity, and Competition. The American Economic Review 90 (1), 166–193.

Bowles, S., January 1998. The Moral Economy of Communities Structured Populations and the Evolution of Pro-Social Norms. Evolution and Human Behavior 19 (1), 3–25.

Bowles, S., Jul. 2003. The co-evolution of individual behaviors and social institutions. Journal of Theoretical Biology 223 (2), 135–147.

Bowles, S., Gintis, H., February 2004. The evolution of strong reciprocity: cooperation in heterogeneous populations. Theoretical Population Biology 65 (1), 17–28.

Boyd, R., Gintis, H., Bowles, S., Apr. 2010. Coordinated punishment of defectors sustains cooperation and can proliferate when rare. Science (New York, N.Y.) 328 (5978), 617–620.

Boyd, R., Gintis, H., Bowles, S., Richerson, P. J., March 2003. The evolution of altruistic punishment. PNAS 100 (6), 3531–3535.

Boyd, R., Richerson, P., Aug. 1990. Group selection among alternative evolutionarily stable strategies. Journal of Theoretical Biology 145 (3), 331–342.

Boyd, R., Richerson, P., May 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. Ethology and Sociobiology 13 (3), 171–195.

Boyd, R., Richerson, P. J., June 1988. Culture and the Evolutionary Process. University Of Chicago Press.

Braeuer, J., Call, J., Tomasello, M., December 2006. Are apes really inequity averse? Proceedings. Biological sciences / The Royal Society 273 (1605), 3123–3128.

Brandt, H., Hauert, C., Sigmund, K., January 2006. Punishing and abstaining for public goods. Proc Natl Acad Sci U S A 103 (2), 495–497.

Brandts, J., Fernanda Rivas, M., December 2009. On punishment and well-being. Journal of Economic Behavior & Organization 72 (3), 823–834.

Brosnan, S. F., de Waal, F. B., 2003. Monkeys reject unequal pay. Nature 425 (6955), 297–299.

Brosnan, S. F., Talbot, C., Ahlgren, M., Lambeth, S. P., Schapiro, S. J., March 2010. Mechanisms underlying responses to inequitable outcomes in chimpanzees, pan troglodytes. Animal Behaviour.

Burkart, J., van Schaik, C., Jan. 2010. Cognitive consequences of cooperative breeding in primates? Animal Cognition 13 (1), 1–19.

Burkart, J. M., Fehr, E., Efferson, C., van Schaik, C. P., December 2007. Other-regarding preferences in a non-human primate: Common marmosets provision food altruistically. Proceedings of the National Academy of Sciences 104 (50), 19762–19766.

Burkart, J. M., Hrdy, S. B., Van Schaik, C. P., 2009. Cooperative breeding and human cognitive evolution. Evol. Anthropol. 18 (5), 175–186.

Burton-Chellew, M. N., Ross-Gillespie, A., West, S. A., Mar. 2010. Cooperation in humans: competition between groups and proximate emotions. Evolution and Human Behavior 31 (2), 104–108.

Camerer, C. F., February 2003. Behavioral Game Theory: Experiments in Strategic Interaction (Roundtable Series in Behaviorial Economics). Princeton University Press.

Capra, F., Feb. 2004. Wendezeit. Droemer Knaur.

Carpenter, J. P., April 2007. The demand for punishment. Journal of Economic Behavior & Organization 62 (4), 522–542.

Carpenter, J. P., Matthews, P. H., Ong'ong'a, O., October 2004. Why punish? social reciprocity and the enforcement of prosocial norms. Journal of Evolutionary Economics 14 (4), 407–429.

Cason, T. N., Khan, F. U., April 1999. A laboratory study of voluntary public goods provision with imperfect monitoring and communication. Journal of Development Economics 58 (2), 533–552.

Cason, T. N., Saijo, T., Yamato, T., October 2002. Voluntary participation and spite in public good provision experiments: An international comparison. Experimental Economics 5 (2), 133–153.

Cesarini, D., Dawes, C. T., Fowler, J. H., Johannesson, M., Lichtenstein, P., Wallace, B., March 2008. Heritability of cooperative behavior in the trust game. Proceedings of the National Academy of Sciences 105 (10), 3721–3726.

Champagne, F., Francis, D., Mar, A., Meaney, M., August 2003. Variations in maternal care in the rat as a mediating influence for the effects of environment on development. Physiology & Behavior 79 (3), 359–371.

Charness, Gary, Rabin, Matthew, August 2002. Understanding Social Preferences with Simple Tests. Quarterly Journal of Economics 117 (3), 817–869.

Clayton, D., November 2000. The genomic action potential. Neurobiology of Learning and Memory 74 (3), 185–216.

Cohen, M., Axelrod, R., Riolo, R., January 2004. Evolution and Altruism. Journal of Economic Behavior and Organization 53 (1), 49–51.

Coleman, J., Aug. 1998. Foundations of Social Theory. Belknap Press of Harvard University Press.

Colman, A. M., April 2006. The puzzle of cooperation. Nature 440 (7085), 744–745.

Cox, J., Friedman, D., Gjerstad, S., Apr. 2007. A tractable model of reciprocity and fairness. Games and Economic Behavior 59 (1), 17–45.

Cox, J. C., Friedman, D., Sadiraj, V., 2008. Revealed altruism. Econometrica 76 (1), 31–69.

Cressman, R., Hofbauer, J., Feb. 2005. Measure dynamics on a one-dimensional continuous trait space: theoretical foundations for adaptive dynamics. Theoretical Population Biology 67 (1), 47–59.

Cummings, M. E., Larkins-Ford, J., Reilly, C. R. L., Wong, R. Y., Ramsey, M., Hofmann, H. A., Feb. 2008. Sexual and social stimuli elicit rapid and contrasting genomic responses. Proceedings of the Royal Society B: Biological Sciences 275 (1633), 393–402.

Dannenberg, A., Riechmann, T., Sturm, B., Vogt, C., 2007. Inequity Aversion and Individual Behavior in Public Good Games: An Experimental Investigation. SSRN eLibrary.

Darcet, D., Sornette, D., December 2008. Quantitative determination of the level of cooperation in the presence of punishment in three public good experiments. Journal of Economic Interaction and Coordination 3 (2), 137–163.

de Hooge, I. E., Zeelenberg, M., Breugelmans, S. M., Jul. 2007. Moral sentiments and cooperation: Differential influences of shame and guilt. Cognition & Emotion 21 (5), 1025–1042.

de Quervain, D. J., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., Fehr, E., August 2004. The neural basis of altruistic punishment. Science 305 (5688), 1254–1258.

de Waal, F. B., Leimgruber, K., Greenberg, A. R., 2008. Giving is self-rewarding for monkeys. Proceedings of the National Academy of Sciences 105 (36), 13685–13689.

Decker, T., Stiehler, A., Strobel, M., December 2003. A comparison of punishment rules in repeated public good games: An experimental study. Journal of Conflict Resolution 47 (6), 751–772.

Denant-Boemont, L., Masclet, D., Noussair, C., October 2007. Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. Economic Theory 33 (1), 145–167.

Dickson, B. J., Nov. 2008. Wired for sex: The neurobiology of drosophila mating decisions. Science 322 (5903), 904–909.

Donaldson, Z. R., Young, L. J., November 2008. Oxytocin, vasopressin, and the neurogenetics of sociality. Science 322 (5903), 900–904.

Dreber, A., Rand, D. G., Fudenberg, D., Nowak, M. A., March 2008. Winners don't punish. Nature 452 (7185), 348–351.

Drickamer, L. C., Vessey, S. H., Miekle, D., Dec. 1995. Animal Behavior: Mechanisms, Ecology, and Evolution, 4th Edition. William C. Brown.

Dunbar, R. I. M., 1998. The social brain hypothesis. Evol. Anthropol. 6 (5), 178–190.

Efferson, C., Lalive, R., Fehr, E., September 2008. The coevolution of cultural groups and ingroup favoritism. Science 321 (5897), 1844–1849.

Efron, B., Tibshirani, R. J., May 1994. An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability), 1st Edition. Chapman and Hall/CRC.

Egas, M., Riedl, A., January 2008. The economics of altruistic punishment and the maintenance of cooperation. Proceedings of the Royal Society B: Biological Sciences 275 (1637), 871–878.

Elster, J., Apr. 2007. Explaining Social Behavior: More Nuts and Bolts for the Social Sciences, 1st Edition. Cambridge University Press.

Englmaier, F., Wambach, A., July 2010. Optimal incentive contracts under inequity aversion. Games and Economic Behavior 69 (2), 312–328.

Enquist, M., Apr. 1993. The evolution of cooperation in mobile organisms. Animal Behaviour 45 (4), 747–757.

Falk, A., Fischbacher, U., February 2006. A theory of reciprocity. Games and Economic Behavior 54 (2), 293–315.

Fehr, E., Bernhard, H., Rockenbach, B., August 2008. Egalitarianism in young children. Nature 454 (7208), 1079–1083.

Fehr, E., Camerer, C. F., October 2007. Social neuroeconomics: the neural circuitry of social preferences. Trends in Cognitive Sciences 11 (10), 419–427.

Fehr, E., Fischbacher, U., 2002. Why social preferences matter - the impact of non-selfish motives on competition, cooperation and incentives. The Economic Journal 112 (478), C1–C33.

Fehr, E., Fischbacher, U., October 2003. The nature of human altruism. Nature 425 (6960), 785–791.

Fehr, E., Fischbacher, U., April 2004. Social norms and human cooperation. Trends Cogn Sci 8 (4), 185–190.

Fehr, E., Fischbacher, U., Gaechter, S., March 2002. Strong reciprocity, human cooperation, and the enforcement of social norms. Human Nature 13 (1), 1–25.

Fehr, E., Gachter, S., 2000. Cooperation and Punishment in Public Goods Experiments. The American Economic Review 90 (4), 980–994.

Fehr, E., Gachter, S., January 2002. Altruistic punishment in humans. Nature 415 (6868), 137–140.

Fehr, E., Gachter, S., January 2005. Human behaviour: Egalitarian motive and altruistic punishment (reply). Nature 433 (7021).

Fehr, E., Schmidt, K. M., 1999. A Theory Of Fairness, Competition, And Cooperation. The Quarterly Journal of Economics 114 (3), 817–868.

Fehr, E., Schmidt, K. M., September 2006. The Economics of Fairness, Reciprocity and Altruism - Experimental Evidence and New Theories, 1st Edition. North Holland, Ch. 8, pp. 616–653.

Fischbacher, U., June 2001. Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters 71 (3), 397–404.

Fischbacher, U., Gachter, S., 2010. Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments. American Economic Review 100 (1), 541–56.

Fletcher, G. E., 2008. Attending to the outcome of others: disadvantageous inequity aversion in male capuchin monkeys (Cebus apella). Am. J. Primatol. 70 (9), 901–905.

Fowler, J. H., May 2005. Altruistic punishment and the origin of cooperation. Proc Natl Acad Sci U S A 102 (19), 7047–7049.

Fowler, J. H., Johnson, T., Smirnov, O., January 2005. Human behaviour: Egalitarian motive and altruistic punishment. Nature 433 (7021), E1+.

Fowler, J. H., Schreiber, D., November 2008. Biology, politics, and the emerging science of human nature. Science 322 (5903), 912–914.

Frey, B. S., Meier, S., 2004. Social Comparisons and Pro-Social Behavior: Testing Conditional Cooperationïn a Field Experiment. The American Economic Review 94 (5), 1717–1722.

Fudenberg, D., Pathak, P. A., October 2009. Unobserved punishment supports cooperation. Journal of Public Economics.

Gachter, S., Renner, E., Sefton, M., December 2008. The Long-Run Benefits of Punishment. Science 322 (5907), 1510+.

Gaechter, S., Herrmann, B., Thoeni, C., September 2010. Culture and cooperation. Philosophical Transactions of the Royal Society B: Biological Sciences 365 (1553), 2651–2661.

Garcia, D., Apr. 2010. Robust smoothing of gridded data in one and higher dimensions with missing values. Computational Statistics & Data Analysis 54 (4), 1167–1178.

Gardner, A., West, S. A., 2004. Cooperation and punishment, especially in humans. The American Naturalist 164 (6), 753–764.

Gigerenzer, Selten (Eds.), August 2002. Bounded Rationality: The Adaptive Toolbox. The MIT Press.

Gigerenzer, G., 2010. Moral satisficing: Rethinking moral behavior as bounded rationality. Topics in Cognitive Science 2 (3), 528–554.

Gintis, H., September 2000. Strong reciprocity and human sociality. Journal of Theoretical Biology 206 (2), 169–179.

Gintis, H., 2001. The hitchhiker's guide to altruism: Gene-culture coevolution, and the internalization of norms. Tech. Rep. 01-10-058, Santa Fe Institute.

Gintis, H., February 2003. The Hitchhiker's Guide to Altruism: Gene-culture Coevolution, and the Internalization of Norms. Journal of Theoretical Biology 220 (4), 407–418.

Gintis, H., March 2009. The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences. Princeton University Press.

Gintis, H., Bowles, S., Boyd, R., Fehr, E., May 2003. Explaining altruistic behavior in humans. Evolution and Human Behavior 24 (3), 153–172.

Glimcher, P. W., Rustichini, A., Oct. 2004. Neuroeconomics: The consilience of brain and decision. Science 306 (5695), 447–452.

Goffman, E., Jun. 1959. The Presentation of Self in Everyday Life, 1st Edition. Anchor.

Goodson, J., Evans, A., Wang, Y., August 2006. Neuropeptide binding reflects convergent and divergent evolution in species-typical group sizes. Hormones and Behavior 50 (2), 223–236.

Grimm, V., Mengel, F., May 2009. Cooperation in viscous populations - experimental evidence. Games and Economic Behavior 66 (1), 202–220.

Grosenick, L., Clement, T. S., Fernald, R. D., January 2007. Fish can infer social rank by observation alone. Nature 445 (7126), 429–432.

Grozinger, C. M., Sharabash, N. M., Whitfield, C. W., Robinson, G. E., Nov. 2003. Pheromone-mediated gene expression in the honey bee brain. Proceedings of the National Academy of Sciences of the United States of America 100 (Suppl 2), 14519–14525.

Guererk, O., Irlenbusch, B., Rockenbach, B., April 2006. The competitive advantage of sanctioning institutions. Science 312 (5770), 108–111.

Hamlin, J. K., Wynn, K., Bloom, P., Nov. 2007. Social evaluation by preverbal infants. Nature 450 (7169), 557–559.

Hardin, G., Dec. 1968. The tragedy of the commons. Science 162 (3859), 1243–1248.

Harmer, G., Abbott, D., 2002. A review of parrondo's paradox. Fluctuation and Noise Letters 2 (2), 71–107.

Hauert, C., De Monte, S., Hofbauer, J., Sigmund, K., May 2002. Volunteering as red queen mechanism for cooperation in public goods games. Science 296 (5570), 1129–1132.

Hauert, C., Traulsen, A., Brandt, H., Nowak, M. A., Sigmund, K., Jun. 2007. Via freedom to coercion: The emergence of costly punishment. Science 316 (5833), 1905–1907.

Hauert, C., Traulsen, A., De Silva, H., Nowak, M. A., Sigmund, K., Apr. 2008. Public goods with punishment and abstaining in finite and infinite populations. Biological Theory 3 (2), 114–122.

Helbing, D., Yu, W., Mar. 2009. The outbreak of cooperation among success-driven individuals under noisy conditions. Proceedings of the National Academy of Sciences of the United States of America 106 (10), 3680–3685.

Henrich, J., January 2004. Cultural group selection, coevolutionary processes and large-scale cooperation. Journal of Economic Behavior & Organization 53 (1), 3–35.

Henrich, J., April 2006. Social science: Enhanced: Cooperation, punishment, and the evolution of human institutions. Science 312 (5770), 60–61.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., 2001. In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. The American Economic Review 91 (2), 73–78.

Henrich, J., Ensminger, J., McElreath, R., Barr, A., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C.,

Marlowe, F., Tracer, D., Ziker, J., March 2010. Markets, Religion, Community Size, and the Evolution of Fairness and Punishment. Science 327 (5972), 1480–1484.

Henrich, J., Mcelreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D., Ziker, J., June 2006. Costly Punishment Across Human Societies. Science 312 (5781), 1767–1770.

Herrmann, B., Thoeni, C., March 2009. Measuring conditional cooperation: a replication study in Russia. Experimental Economics 12 (1), 87–92.

Herrmann, B., Thoni, C., Gachter, S., March 2008. Antisocial punishment across societies. Science 319 (5868), 1362–1367.

Hetzer, M., Sornette, D., 2010. The effect of other-regarding preferences on the evolution of altruistic punishment.

Hetzer, M., Sornette, D., 2011. An theory of evolution, fairness and altruistic punishment.

Hil, K., Gurven, M., March 2004. Economic Experiments to Examine Fairness and Cooperation among the Ache Indians of Paraguay. Vol. 1. Oxford Scholarship Online Monographs.

Hill, K. R., Walker, R. S., Bozicevic, M., Eder, J., Headland, T., Hewlett, B., Hurtado, A. M., Marlowe, F., Wiessner, P., Wood, B., Mar. 2011. Co-residence patterns in hunter-gatherer societies show unique human social structure. Science 331 (6022), 1286–1289.

Hofbauer, J., Oechssler, J., Riedel, F., Mar. 2009. Brown- von neumann-nash dynamics: The continuous strategy case. Games and Economic Behavior 65 (2), 406–429.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., Thagard, P. R., March 1989. Induction: Processes of Inference, Learning, and Discovery. The MIT Press.

Holmas, T. H., Kjerstad, E., Luras, H., Straume, O. R., Apr. 2010. Does monetary punishment crowd out pro-social motivation? a natural experiment on hospital length of stay. Journal of Economic Behavior & Organization.

Homans, G. C., May 1974. Social Behavior: Its Elementary Forms. Houghton Mifflin Harcourt P.

Hopfensitz, A., Reuben, E., 2009. The importance of emotions for the effectiveness of social punishment. The Economic Journal 119 (540), 1534–1559.

Houser, D., Kurzban, R., Jul. 2003. Conditional cooperation and group dynamics: Experimental evidence from a sequential public goods game. Tech. Rep. 0307001, EconWPA.

Ichinose, G., Arita, T., Jan. 2008. The role of migration and founder effect for the evolution of cooperation in a multilevel selection context. Ecological Modelling 210 (3), 221–230.

Imhof, L. A., Fudenberg, D., Nowak, M. A., August 2005. Evolutionary cycles of cooperation and defection. Proceedings of the National Academy of Sciences of the United States of America 102 (31), 10797–10800.

Jablonka, E., Lamb, M. J., 2007. Precis of evolution in four dimensions. Behavioral and Brain Sciences 30 (04), 353–365.

Jarvis, E. D., Scharff, C., Grossman, M. R., Ramos, J. A., Nottebohm, F., October 1998. For whom the bird sings: Context-dependent gene expression. Neuron 21 (4), 775–788.

Jasny, B. R., Kelner, K. L., Pennisi, E., November 2008. From genes to social behavior. Science 322 (5903), 891+.

Jensen, K., September 2010. Punishment and spite, the dark side of cooperation. Philosophical Transactions of the Royal Society B: Biological Sciences 365 (1553), 2635–2650.

Jensen, K., Call, J., Tomasello, M., Oct. 2007a. Chimpanzees are rational maximizers in an ultimatum game. Science 318 (5847), 107–109.

Jensen, K., Call, J., Tomasello, M., August 2007b. Chimpanzees are vengeful but not spiteful. Proceedings of the National Academy of Sciences 104 (32), 13046–13050.

Jensen, K., Hare, B., Call, J., Tomasello, M., April 2006. What's in it for me? self-regard precludes altruism and spite in chimpanzees. Proceedings of the Royal Society B: Biological Sciences 273 (1589), 1013–1021.

Kahneman, D., Knetsch, J. L., Thaler, R., 1986. Fairness as a Constraint on Profit Seeking: Entitlements in the Market. The American Economic Review 76 (4), 728–741.

Kennedy, D., Norman, C., Jul. 2005. What don't we know? Science 309 (5731), 75.

Kennedy, D. M., September 2008. Deterrence and Crime Prevention: Reconsidering the Prospect of Sanction (Routledge Studies in Crime and Economics). Routledge.

Keynes, J. M., July 2006. The General Theory of Employment, Interest and Money. Atlantic Publishers & Distributors (P) Ltd.

Kleiman, M., Kilmer, B., August 2009. The dynamics of deterrence. Proceedings of the National Academy of Sciences 106 (34), 14230–14235.

Kleiman, M. A. R., September 2009. When Brute Force Fails: How to Have Less Crime and Less Punishment. Princeton University Press.

Kolm, S., Ythier, J. M. (Eds.), September 2006. Handbook of the Economics of Giving, Altruism and Reciprocity, Volume 1: Foundations (Handbooks in Economics), 1st Edition. North Holland.

Kosfeld, M., Okada, A., Riedl, A., 2009. Institution formation in public goods games. American Economic Review 99 (4), 1335–55.

Kurzban, R., Houser, D., 2001. Individual differences in cooperation in a circular public goods game. European Journal of Personality 15 (S1), S37–S52.

Kurzban, R., Houser, D., February 2005. Experiments investigating cooperative types in humans: a complement to evolutionary theory and simulations. Proc Natl Acad Sci U S A 102 (5), 1803–1807.

Laland, K., Smee, J. O., Feldman, M., 2000. Niche construction, biological evolution and cultural change. Behavioral and Brain Sciences 23, 131–146.

Lehmann, L., Keller, L., West, S., Roze, D., Apr. 2007. Group selection and kin selection: Two concepts but one process. Proceedings of the National Academy of Sciences 104 (16), 6736–6739.

Leigh, E. G., Jan. 2010. The group selection controversy. Journal of evolutionary biology 23 (1), 6–19.

Levati, M. V., Sutter, M., van der Heijden, E., October 2007. Leading by example in a public goods experiment with heterogeneity and incomplete information. Journal of Conflict Resolution 51 (5), 793–818.

Mahajan, N., Martinez, M. A., Gutierrez, N. L., Diesendruck, G., Banaji, M. R., Santos, L. R., 2011. The evolution of intergroup bias: Perceptions and attitudes in rhesus macaques. Journal of Personality and Social Psychology 100 (3), 387–405.

Manski, C. F., July 1977. The structure of random utility models. Theory and Decision 8 (3), 229–254.

Marlowe, F. W., Berbesque, J. C., Barrett, C., Bolyanatz, A., Gurven, M., Tracer, D., Jul. 2011. The 'spiteful' origins of human cooperation. Proceedings of the Royal Society B: Biological Sciences 278 (1715), 2159–2164.

Masclet, D., Noussair, C., Tucker, S., Villeval, M.-C., 2003. Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism. The American Economic Review 93 (1), 366–380.

Mathew, S., Boyd, R., Jul. 2011. Punishment sustains large-scale cooperation in prestate warfare. Proceedings of the National Academy of Sciences 108 (28), 11375–11380.

McElreath, R., Boyd, R., Mar. 2007. Mathematical Models of Social Evolution: A Guide for the Perplexed. University Of Chicago Press.

McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. in Frontiers in Econometrics, P. Zarembka (ed.), New York: Academic Press, 105–142.

McFadden, D., 1981. Econometric models of probabilistic choice. in Structural Analysis of Discrete Data with Econometric Applications, C.F. Manski and D. McFadden (eds.), Cambridge, MIT Press, 198–272.

McNamara, J. M., Leimar, O., Sep. 2010. Variation and the response to variation as a basis for successful cooperation. Philosophical Transactions of the Royal Society B: Biological Sciences 365 (1553), 2627–2633.

Mehra, Y. P., 2001. The wealth effect in empirical life-cycle aggregate consumption equations. Federal Reserve Bank of Richmond - Economic Quarterly 87/2.

Meier, S., Jul. 2006. A survey of economic theories and field evidence on prosocial behavior. Social Science Research Network Working Paper Series.

Mello, C. V., Vicario, D. S., Clayton, D. F., August 1992. Song presentation induces gene expression in the songbird forebrain. Proceedings of the National Academy of Sciences of the United States of America 89 (15), 6818–6822.

Messick, D., May 1999. Alternative logics for decision making in social settings. Journal of Economic Behavior & Organization 39 (1), 11–28.

Miguel, E., Gugerty, M. K., December 2005. Ethnic diversity, social sanctions, and public goods in kenya. Journal of Public Economics 89 (11-12), 2325–2368.

Nikiforakis, N., Mar. 2010. Feedback, punishment and cooperation in public good experiments. Games and Economic Behavior 68 (2), 689–702.

Nikiforakis, N., Normann, H.-T., December 2008. A comparative statics analysis of punishment in public-good experiments. Experimental Economics 11 (4), 358–369.

Noussair, C., Tucker, S., July 2005. Combining monetary and social sanctions to promote cooperation. Economic Inquiry 43 (3), 649–660.

Nowak, M. A., December 2006. Five rules for the evolution of cooperation. Science 314 (5805), 1560–1563.

Nowak, M. A., Sasaki, A., Taylor, C., Fudenberg, D., April 2004. Emergence of cooperation and evolutionary stability in finite populations. Nature 428 (6983), 646–650.

Nowak, M. A., Tarnita, C. E., Wilson, E. O., Aug. 2010. The evolution of eusociality. Nature 466 (7310), 1057–1062.

Oechssler, J., Riedel, F., Jan. 2001. Evolutionary dynamics on infinite strategy spaces. Economic Theory 17 (1), 141–162.

Ohtsuki, H., Iwasa, Y., Nowak, M. A., January 2009. Indirect reciprocity provides only a narrow margin of efficiency for costly punishment. Nature 457 (7225), 79–82.

Page, T., Putterman, L., Unel, B., Oct. 2005. Voluntary association in public goods experiments: Reciprocity, mimicry and efficiency. The Economic Journal 115 (506), 1032–1053.

Palfrey, T. R., Prisbrey, J. E., 1997. Anomalous behavior in public goods experiments: How much and why? The American Economic Review 87 (5), 829–846.

Parsons, T., Sep. 1967. Sociological Theory and Modern Society., first edition Edition. Free Press.

Pennisi, E., July 2005. How did cooperative behavior evolve? Science 309 (5731), 93+.

Plous, S., 1993. The Psychology of Judgment and Decision Making (McGraw-Hill Series in Social Psychology), 1st Edition. McGraw-Hill.

Precht, R. D., Oct. 2010. Die Kunst, kein Egoist zu sein: Warum wir gerne gut sein wollen und was uns davon abhält. Goldmann Verlag.

Rabin, M., 1993. Incorporating Fairness into Game Theory and Economics. The American Economic Review 83 (5), 1281–1302.

Range, F., Horn, L., Viranyi, Z., Huber, L., December 2008. The absence of reward induces inequity aversion in dogs. Proceedings of the National Academy of Sciences.

Reeve, H. K., Sherman, P. W., 1993. Adaptation and the Goals of Evolutionary Research. The Quarterly Review of Biology 68 (1).

Reuben, E., van Winden, Jan. 2005. Negative reciprocity and the interaction of emotions and fairness norms. Social Science Research Network Working Paper Series.

Robinson, G., Fernald, R., Clayton, D., November 2008. Genes and social behavior. Science 322 (5903), 896–900.

Rockenbach, B., Milinski, M., December 2006. The efficient interaction of indirect reciprocity and costly punishment. Nature 444 (7120), 718–723.

Rogers, D. S., Ehrlich, P. R., Mar. 2008. Natural selection and cultural rates of change. Proceedings of the National Academy of Sciences of the United States of America 105 (9), 3416–3420.

Rustagi, D., Engel, S., Kosfeld, M., Nov. 2010. Conditional cooperation and costly monitoring explain success in forest commons management. Science 330 (6006), 961–965.

Savage, L. J., June 1972. The Foundations of Statistics, 2nd Edition. Dover Publications.

Schotter, A., Oct. 1996. Fairness and survival in ultimatum and dictatorship games. Journal of Economic Behavior & Organization 31 (1), 37–56.

Selten, R., Ostmann, A., December 2000. Imitation Equilibrium. Tech. Rep. bgse16_2000, University of Bonn, Germany.

Siegfried, T., July 2005. In praise of hard questions. Science 309 (5731), 76–77.

Sigmund, K., De Silva, H., Traulsen, A., Hauert, C., July 2010. Social learning promotes institutions for governing the commons. Nature 466 (7308), 861–863.

Silk, J. B., Alberts, S. C., Altmann, J., November 2003. Social bonds of female baboons enhance infant survival. Science 302 (5648), 1231–1234.

Silk, J. B., Brosnan, S. F., Vonk, J., Henrich, J., Povinelli, D. J., Richardson, A. S., Lambeth, S. P., Mascaro, J., Schapiro, S. J., October 2005. Chimpanzees are indifferent to the welfare of unrelated group members. Nature 437 (7063), 1357–1359.

Simon, H., Egidi, M., Viale, R., Marris, R. L., March 2007. Economics, Bounded Rationality and the Cognitive Revolution. Edward Elgar Pub.

Simon, H. A., 1982. Models of Bounded Rationality. MIT Press Cambridge, Mass.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., Frith, C. D., Jan. 2006. Empathic neural responses are modulated by the perceived fairness of others. Nature 439 (7075), 466–469.

Sinha, A., February 2005. Not in their genes: Phenotypic flexibility, behavioural traditions and cultural evolution in wild bonnet macaques. Journal of Biosciences 30 (1), 51–64.

Smith, J. M., Mar. 1964. Group selection and kin selection. Nature 201 (4924), 1145–1147.

Soares, M. C., Bshary, R., Fusani, L., Goymann, W., Hau, M., Hirschenhauser, K., Oliveira, R. F., Sep. 2010. Hormonal mechanisms of cooperative behaviour. Philosophical Transactions of the Royal Society B: Biological Sciences 365 (1553), 2737–2750.

Sonnemans, J., Schram, A., Offerman, T., Jan. 1999. Strategic behavior in public good games: when partners drift apart. Economics Letters 62 (1), 35–41.

Stevens, J. R., Hauser, M. D., February 2004. Why be nice? psychological constraints on the evolution of cooperation. Trends in Cognitive Sciences 8 (2), 60–65.

Sutton, R. S., Barto, A. G., March 1998. Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning). The MIT Press.

Takahashi, J. S., Shimomura, K., Kumar, V., November 2008. Searching for genes underlying behavior: Lessons from circadian rhythms. Science 322 (5903), 909–912.

Tomasello, M., Warneken, F., August 2008. Human behaviour: Share and share alike. Nature 454 (7208), 1057–1058.

Tricomi, E., Rangel, A., Camerer, C. F., O/'Doherty, J. P., February 2010. Neural evidence for inequality-averse social preferences. Nature 463 (7284), 1089–1091.

van den Bergh, J. C. J. M., Gowdy, J. M., Oct. 2009. A group selection perspective on economic behavior, institutions and organizations. Journal of Economic Behavior & Organization 72 (1), 1–20.

Vohs, K. D., Mead, N. L., Goode, M. R., Nov. 2006. The psychological consequences of money. Science 314 (5802), 1154–1156.

von Neumann, J., Morgenstern, O., March 2007. Theory of Games and Economic Behavior (Commemorative Edition) (Princeton Classic Editions), 60th Edition. Princeton University Press.

Wade, M. J., 1978. A critical review of the models of group selection. The Quarterly Review of Biology 53 (2).

Wade, M. J., 1982. Group selection: Migration and the differentiation of small populations. Evolution 36 (5), 949–961.

Waibel, M., Floreano, D., Keller, L., 2011. A quantitative test of hamilton's rule for the evolution of altruism. PLoS Biol 9 (5).

West, S. A., Griffin, A. S., Gardner, A., Mar. 2007. Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. Journal of Evolutionary Biology 20 (2), 415–432.

West, S. A., Pen, I., Griffin, A. S., Apr. 2002. Cooperation and competition between relatives. Science 296 (5565), 72–75.

Whiten, A., Horner, V., de Waal, F. B. M., August 2005. Conformity to cultural norms of tool use in chimpanzees. Nature 437, 737–740.

Whitfield, C. W., Cziko, A.-M., Robinson, G. E., October 2003. Gene expression profiles in the brain predict behavior in individual honey bees. Science 302 (5643), 296–299.

Williams, G. C., May 1996. Adaptation and Natural Selection. Princeton University Press.

Wilson, D. S., 1977. Structured demes and the evolution of group-advantageous traits. The American Naturalist 111 (977).

Wilson, D. S., 1983. The group selection controversy: History and current status. Annual Review of Ecology and Systematics 14.

Wilson, D. S., May 2004. What is wrong with absolute individual fitness? Trends in Ecology & Evolution 19 (5), 245–248.

Wright, S., March 1943. Isolation by distance. Genetics 28 (2), 114–138.

Wu, J.-J. J., Zhang, B.-Y. Y., Zhou, Z.-X. X., He, Q.-Q. Q., Zheng, X.-D. D., Cressman, R., Tao, Y., Oct. 2009. Costly punishment does not always increase cooperation. Proceedings of the National Academy of Sciences of the United States of America 106 (41), 17448–17451.

Yukalov, V., Sornette, D., 2009. Processing information in quantum decision theory. Entropy 11, 1073–1120.

Yukalov, V., Sornette, D., 2010a. Decision theory with prospect interference and entanglement. Theory and Decision, http://arXiv.org/abs/0802.3597.

Yukalov, V., Sornette, D., 2010b. Mathematical structure of quantum decision theory. Advances in Complex Systems 13 (5), 659–698.

Zhou, W. X., Sornette, D., Hill, R. A., Dunbar, R. I. M., Feb. 2005. Discrete hierarchical organization of social group sizes. Proceedings of the Royal Society B: Biological Sciences 272 (1561), 439–444.