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.