Journal of Theoretical Biology 290 (2011) 88–103

Contents lists available at SciVerse ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Evolution of altruistic punishment in heterogeneous populations

Harmen de Weerd *, Rineke Verbrugge

Institute of Artificial Intelligence, Faculty of Mathematics and Natural Sciences, University of Groningen, PO Box 407, 9700 AK Groningen, The Netherlands

ARTICLE INFO

Article history:
Received 23 November 2010
Received in revised form
24 August 2011
Accepted 24 August 2011
Available online 1 September 2011

Keywords:
Agent-based models
Evolution of cooperation
Non-uniform costs and payoffs
Strong reciprocity
Emergent behavior

ABSTRACT

Evolutionary models for altruistic behavior typically make the assumption of homogeneity: each individual has the same costs and benefits associated with cooperating with each other and punishing for selfish behavior. In this paper, we relax this assumption by separating the population into heterogeneous classes, such that individuals from different classes differ in their ability to punish for selfishness. We compare the effects of introducing heterogeneity this way across two population models, that each represents a different type of population: the infinite and well-mixed population describes the way workers of social insects such as ants are organized, while a spatially structured population is more related to the way social norms evolve and are maintained in a social network.

We find that heterogeneity in the effectiveness of punishment by itself has little to no effect on whether or not altruistic behavior will stabilize in a population. In contrast, heterogeneity in the cost that individuals pay to punish for selfish behavior allows altruistic behavior to be maintained more easily. Fewer punishers are needed to deter selfish behavior, and the individuals that punish will mostly belong to the class that pays a lower cost to do so. This effect is amplified when individuals that pay a lower cost for punishing inflict a higher punishment.

The two population models differ when individuals that pay a low cost for punishing also inflict a lower punishment. In this situation, altruistic behavior becomes harder to maintain in an infinite and well-mixed population. However, this effect does not occur when the population is spatially structured.

1. Introduction

The question of how cooperation has evolved represents one of the more enduring puzzles in biology and social sciences, in which the role of many pieces is understood even if some pieces do not yet seem to fit together (Hamilton, 1964; Hardin, 1968; Axelrod and Hamilton, 1981; Sigmund, 2010; Gärdenfors, 2011). The paradox of cooperation is that although cooperation adds to the common good of a group of individuals, contributing to the common good generally bears a higher cost than the individual returns (Hardin, 1968; West et al., 2007). Individuals that enjoy the cooperation of others without being cooperative themselves are therefore at an evolutionary advantage. At first sight, a group of individuals thus seems to be destined to never cooperate, even if the combined benefit of every single individual cooperating outweighs the cost of contributing.

Even though cooperation seems to be destined to fail in theory, many social animals engage in cooperative action, ranging over a wide variety of activities (Wilkinson, 1984; Mulder and Langmore, 1993; Dugatkin, 1997; Crespi, 2001). To explain why cooperation stabilizes in many animal societies, a number of mechanisms have been proposed, varying in the assumptions they make on individual cognitive abilities (see among others Nowak, 2006; Gärdenfors, 2011). One of the mechanisms that may stabilize cooperation is punishment (Sigmund et al., 2001; Boyd et al., 2003; Brandt et al., 2003; Fowler, 2005). Punishment can provide the necessary incentive to stabilize cooperation in animal (Clutton-Brock and Parker, 1995; Monnin and Ratnieks, 2001) as well as in human societies (Ostrom, 2000; Fehr and Gächter, 2002). Experiments have shown that human subjects have a high willingness to sacrifice in order to punish selfish behavior, even when punishment is understood to yield no future benefits (Güth et al., 1982; Camerer and Thaler, 1995; Bolton and Zwick, 1995; Henrich et al., 2001; Fehr and Gächter, 2000).

An N-person extension of the prisoner's dilemma, known as the public goods game (Kagel et al., 1995; Fehr and Gächter, 2002), has been investigated in simulations of the evolution of cooperation. In the public goods game, the game is played by $N>2$ individuals, each of which receives an initial capital C. They may choose to keep that capital to themselves, or invest any part of it in a common pool. Once every player has decided how much to invest, the capital in the common pool is doubled, and divided equally among the players, irrespective of their investment. If every player invests their entire

* Corresponding author. Tel.: +31 50 363 4114; fax: +31 50 363 6687.
E-mail addresses: hdeweerd@ai.rug.nl (H. de Weerd), rineke@ai.rug.nl (R. Verbrugge).

0022-5193/$ - see front matter
doi:10.1016/j.jtbi.2011.08.034

capital, each will end up with $2C$ and therefore double their initial capital. However, each individual is faced with the temptation of exploiting the common pool. Since every individual investment is divided equally among all $N > 2$ individuals, the return on the individual investment is negative. The game-theoretical dominant strategy would therefore be to invest nothing. But if none of the players invests, each will end up with their initial capital C, which is half the capital they would have gained if everyone had invested. In experiments with volunteers with actual economic incentives, human players do tend to invest a reasonable sum. Typically, in the first round, participants choose to invest at least half their capital. When the game is repeated over several rounds, the amount invested quickly declines until nobody invests anything, unless there is an opportunity to punish individuals for low investments (Fehr and Gächter, 2002) or opt out of playing the public goods game (Orbell and Dawes , 1993; Semmann et al., 2003).

The models that have been proposed so far to explain why cooperation and punishment persist and would even be able to invade in a population of selfish individuals, commonly make the assumption of homogeneity. In a homogeneous society, individuals can use different strategies, but the payoffs of an encounter between two individuals depend only on the strategy the individuals adopt. Individuals have the same cost of punishing, and the same benefit of their partner cooperating. In this article, we propose to relax this assumption of homogeneity by allowing for populations that consist of two or more different sub-classes. By allowing the costs and benefits of cooperation to vary across sub-classes, individuals from different classes may have different opportunities. Empirical research shows that differences in marginal benefit from contributions to a public good changes the willingness to contribute and punish (Fisher et al., 1995; Reuben and Riedl, 2009). Subjects that enjoy a higher benefit not only tend to contribute more to the public good, but are also expected to do so, and are punished more severely by other players if they contribute less than their fair share.

In this research, we determine the effects of a heterogeneous population of individuals on the evolution of altruistic punishment, and the resulting structure of the population in a simulated environment. We adjust the model of the public goods game with voluntary participation introduced by Hauert et al. (2002, 2002) and further extended to include altruistic punishment (Fowler, 2005; Brandt et al., 2006) to allow for heterogeneous classes of individuals. Specifically, we investigate the effect of individual differences in the cost for punishing a co-player as well as the cost of being punished by another individual. We compare these effects across two different population models. In our first model, discussed in Section 2, the public goods game is played in an infinite size and well-mixed population, where individuals are assumed never to encounter each other more than once in the same setting. Section 3 describes the second model, which imposes a spatial structure on the population in the form of a lattice, such that individuals only interact with a small selection of close neighbors. For both population models, we present a model for a population that is divided into M classes of individuals, and show the results of an implementation of the model for the case of $M=2$ classes. The individuals we simulate share the knowledge that the population is heterogeneous, but not how this affects the rewards. Simulation results are presented separately for each model, while Section 4 summarizes these results and provides directions for further research.

2. Infinite population model

To determine the effect of heterogeneity of individuals on the evolution of altruistic punishment, we have constructed two model variations of the public goods game. In this section, we will discuss a model based on the assumption of an infinite sized, well-mixed population of individuals. This model can be used to represent any sufficiently large population in which individuals are very unlikely to encounter the same co-player twice in the setting of a public goods game over the course of their lifetime. The infinite population model may therefore describe the public goods game in a large colony of social insects, such as ants, bees or wasps. In these societies, workers generally exhibit altruistic behavior by sacrificing most or all of their direct reproduction to help rear the offspring of the queen (Oster and Wilson, 1979). Interestingly, in some species of social insects, infertile workers can still lay haploid eggs destined to be males (Wenseleers et al., 2005). There is an evolutionary incentive to do so when the queen is mated to more than two males, in the sense that workers are more related to their own sons than to sons of their queen mother and sons of their sister workers (Trivers and Hare, 1976; Wenseleers et al., 2004). The reward for such behavior is therefore an increase in their inclusive fitness, that is the probability of their genes surviving. Natural selection would therefore favor the social insect that lays its own eggs. However, workers lay eggs at the expense of performing their duties to the colony. Punishment for this selfish behavior takes the form of queen and worker policing (Monnin and Ratnieks, 2001). Through this mechanism, worker-laid eggs are destroyed, effectively removing all benefits from the selfish behavior.

Even though workers, queens and males in a colony of social insects represent morphologically different castes that perform different tasks, the homogeneous infinite population model can be used to model the interactions between the workers of large colonies. However, some colonies of social insects exhibit a further subdivision of the worker caste (Oster and Wilson, 1979) up to a point where a heterogeneous infinite population model would fit the situation better. For example, leaf-cutting ant workers exhibit a 200-fold variation in body mass (Wilson, 1980), while in weaver ants of the genus *Oecophylla*, workers show a clear bimodal size distribution, with almost no overlap in size between minor and major workers (Hölldobler and Wilson, 1990). In cases like these, morphologically different workers typically perform different tasks depending on their physical traits. In general, the minor workers stay in the nest to tend to the queen and her brood, while major workers perform the more dangerous tasks of foraging and defending the colony (Hölldobler and Wilson, 1990).

In the remainder of this section, we will discuss how heterogeneity between individuals affects the evolution of altruistic punishment in the infinite population model. As a starting point, we use the model introduced by Brandt et al. (2006), which already allows for voluntary participation. This model is extended in the present work by dividing the population into M heterogeneous classes of individuals. For the simulation results, we restrict ourselves to the case $M=2$.

2.1. Infinite population model: methods

In the infinite population model, we follow Brandt et al. (2006). Their model is an extension of the basic public goods model, in which players may choose not to share in the public good and instead receive a fixed payoff. We further extend their model to allow for heterogeneous groups within the population. In our case, the population is assumed to consist of M classes of individuals, which occur at a fixed ratio within the population. That is, although evolutionary dynamics affect the frequencies at which the different strategies occur within each class, this has no effect on the relative frequency of the different classes within the population. In effect, this means there is no genetic basis that determines the individual membership to a class. Each class of individuals occurs at a constant frequency $0 < f_i < 1$ $(1 \le i \le M)$, such that $\sum_i f_i = 1$.