THE (NEURO-)BIOLOGY OF ALTRUISTIC PUNISHMENT

A Philosophical Investigation of a Concept of Human Social Behavior

Rebekka A. Klein
University of Zurich
University Research Priority Project: Foundations of Human Social Behavior
Subproject: Philosophical Problems of Natural Altruism
Address: University of Zurich, Bluemlisalpstrasse 10, CH-8006 Zurich
Contact: +41-44-63 45094, klein@iew.uzh.ch

Abstract

This paper deals with the experimental model of altruistic punishment and social norm enforcement which has been created in the research field of neuroeconomics recently. By use of this model, neurobiologists and economists investigate the close relationship between neurobiological mechanisms in the brain and specific patterns of human social behavior. They have experimentally shown that the implementation of a punishment tool in social interaction experiments gives empirical evidence for the great impact of non-selfish behavior on social group interaction and individual strategies of cooperation, competition and collective action. The interpretation of this evidence and their impact on social theory is critically questioned in this paper from a philosophical point of view.

Outline

1 The study of altruism as behavioral pattern in neuroeconomics

2 The correlation of norm enforcement and altruistic punishment in humans

3 The neurobiological explanation of altruistic behavior and its use in economics

4 Summary: The significance of altruistic punishment

5 Connecting Questions and Philosophical Assessment: The Paradigm of ›Abu Ghraib‹

1 The study of altruism as behavioral pattern in neuroeconomics

The biological understanding of altruism is based on a consideration about the economy of human nature. It says: If a human being is altruistic he/she will incur personal costs to increase the material benefit or fitness of another human being. Thus, although the biological concept of altruism is applied to humans, it does not capture their psychological motivation, e.g. their beliefs, desires, and reasons appearing behind actual behavior.[1] However, the biological concept of altruism concentrates on significant outcomes and proper evolutionary function of human social behaviors. It is useful to explore consequences of behavior rather than psychological motivation. Hence, it is especially helpful to study the behavioral traits of social exchange practices.[2]

The latter has recently been detected by experimental economists who pursue a social science research strategy. They apply the concept of biological altruism in experiments which are conducted to study human social behavior according to a game

[1] Cf. the difference of biological and psychological altruism in: Elliott Sober and David S. Wilson, *Unto Others. The Evolution and Psychology of Unselfish Behavior* (Cambridge, Mass.: Harvard University Press, 1998).

[2] Colin F. Camerer and Ernst Fehr, "Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists," in *Foundations of Human Sociality*, ed. Joseph Henrich et al. (Oxford: Oxford University Press, 2004), 55-95.

theoretical framework. Their economically inspired study of social interactions and transactions among humans is guided by an interest in modeling people's social preferences. Preferences are used in economic theory to measure people's choices and valuations of certain goods, as for example, food, money, prestige, etc. To determine the preference relations of people, economists observe their actual choices and decision-making in an experimental environment where real money is at stake. They do so, because they specially focus their research on outcomes of behavior. In case of social preferences, these outcomes of behavior involve a relation to other individuals and are referred to as social.

However, the growing interest of experimental economics in biological altruism also is a rather provocative enterprise within their own discipline, because standard economic models of human social behavior do not take into account non-selfish, altruistic or even social preferences. Instead, they assume that human rational behavior is mainly exhibited in self-interested, egoistic choices (*homo oeconomicus* model). To show that this is wrong, some experimental economists began to reinforce their view of the economic agent by naturalizing human sociality and investigating the biological foundations of human interaction strategies. For this purpose, they integrated new methodologies and research strategies from the natural sciences into their experimental framework and founded the trans-disciplinary research approach of neuroeconomics. This approach allows combining the methodological tools of neurobiology and experimental economics in a shared experimental environment.[3]

Neuroeconomics as a behavioral science does not claim that classical economic game theory is wrong as a whole, but that certain predictions and principles of it can be falsified by empirical research. One of the main objectives of neuroeconomics is to

[3] Herbert Gintis, "A framework for the unification of the behavioral sciences," *Behav Brain Sci* 30 (2007): 1-61.

provide empirical evidence for the claim that the concept of human behavior as represented by the *homo oeconomicus* model is too simplified to capture the complexity and diversity of people's real life behavior. In addition to this, neuroeconomics participates in the major endeavor of explaining the nature of human altruism and the evolution of cooperation across human species.

In the view of evolutionary anthropology, human cooperation does not only differ from non-human mammalian species, it rather is of a different kind: It shows a great variability in scale and domain and was probably developed in a non-genetic evolutionary process which cannot be observed in other species.[4] One example of this might be that humans live in large-scale societies which are built up on the basis of anonymous encounters between genetically non-related individuals. Human cooperation is flourishing in these societies in spite of anonymity and non-relatedness, because group interaction is based on behavioral standards/social norms. Stability and coordination in human social interaction is, therefore, established through the creation and maintenance of social norms. The maintenance of norms and their adaptation as rules of behavior is secured through enforcement strategies such as, for instance, social monitoring and interpersonal sanctioning.[5] Starting from this insight, the science of neuroeconomics has developed a wide range of experimental tools to study the relevant behavioral patterns. Altruistic punishment has proven to be one of the key patterns among them.

[4] Cf. Joseph Henrich and Natalie Henrich, "Culture, evolution and the puzzle of human cooperation," *Cognitive Systems Research* 7 (2006): 223-224.
[5] This has been shown first in the field studies of Elinor Ostrom. Cf. Elinor Ostrom, *Governing the Commons. The Evolution of Institutions for Collective Action* (New York: Cambridge University Press, 1990).

2 The correlation of norm enforcement and altruistic punishment in humans

Neuroeconomics has recently argued that one of the best examples for the unique character of human prosociality is the behavioral pattern of altruistic punishment. Unlike cooperation, the term prosociality is used in this context as an umbrella term for patterns of behavior that do not directly benefit to others (as cooperation does), but to the well-being of group interaction as a whole. Prosociality is, therefore, a term that covers a broader understanding of social behaviors than cooperation. It can be used to model the benefit of an interaction beyond the self-other relationship and takes into account the social dimension of ›the third‹. In sociological theories, ›the third‹ is representative of something that is beyond ›me‹ (ego) and ›you‹ (alter), although it is part of the relationship of the two.[6] This can be a third person as well as a social institution, which has a significant influence on the interaction of self and other. The different manifestations of ›the third‹ come into play when co-presence of self and other opens up for a triadic relationship, which allows recognition of social differences and inequalities between the two. With recognition of these differences and it's overcoming in human society and law system starts the evolution of mankind.

Several experimental studies on cooperation and prosociality in neuroeconomics have shown that altruistic punishment plays a key role in understanding the evolution of human social behavior. For altruistic punishment does not directly benefit to the welfare of an individual person, but to society as a whole. It is, therefore, a necessary condition for the maintenance of norms within a social group. Norms in this case are representatives of the point of view of ›the third‹, which allows evaluating the relationship of self to another as driven by a positive or negative concern for the welfare of others, e.g. prosocial or antisocial preference.

[6] Cf. an overview of different notions of ›the third‹ in social philosophy: Thomas Bedorf, Dimensionen des Dritten. Sozialphilosophische Modelle zwischen Ethischem und Politischen (München: Fink 2003).

The behavioral pattern of altruistic punishment has been clearly shown to be of great significance for the study of human prosociality by a set of behavioral experiments. These have been conducted in different behavioral labs since the first study on altruistic punishment was published in 2002 by Ernst Fehr and Simon Gächter.[7] In this study, the two economists define altruistic punishment as a non-selfish act of punishment which "provide[s] a material benefit for the future interaction partners of the punished subject but not for the punisher."[8] In an experimental setup at the University of Zurich, Fehr and Gächter tested their subject's individual willingness to punish in a public goods experiment. In this type of experiment, a bunch of people has the option of investing a certain amount of money into a group project. Afterwards, the sum of all contributions is shared among the group members equally. The experiment in Zurich was conducted in six sessions and group composition was changed each session. The latter guaranteed that none of the subjects could meet again with the same subjects during the experiment.[9] This secured on the methodological level that the subject's decisions and behaviors were not based on a preference for reputation building among group members. The opportunity to punish group members who did not invest in the group project, but benefited from its gain, was offered in the end of each session.

Results of the experiment were: The opportunity to punish social free-riding behavior was taken by 84.3% of the subjects at least once and even 34.3% of the subjects punished more than five times during the six sessions of the experiment.[10] Thus, the experimental results provide strong evidence that the punishment tool is helpful to model a clear pattern of human social behavior. Additionally, an effect of altruistic punishment could be shown in the later sessions of the experiment. After having been punished, the

[7] Ernst Fehr and Simon Gächter, "Altruistic Punishment in Humans," *Nature* 415 (2002): 137-140.
[8] Ernst Fehr and Simon Gächter, "Altruistic Punishment in Humans," *Nature* 415 (2002): 139.
[9] The total number of participants in the experiment was 240, all of them undergraduate students from the University of Zurich.
[10] Ernst Fehr and Simon Gächter, "Altruistic Punishment in Humans," *Nature* 415 (2002): 137.

punished subjects invested a higher amount of money into the group project and changed from non-cooperative to cooperative behaviors. Thus, altruistic punishment caused a substantial increase of the average cooperation level of the group over time. It has a strong correlation to the subject's investment strategies and might, therefore, be a facilitating condition of the evolution of human cooperation.

However, the remarkable result of the study was that some "altruistic" subjects took the opportunity to punish free-riders, although it was costly for them and had neither direct nor indirect benefit to them. In the interpretation of this evidence, experimenters proposed that the willingness of subjects to punish shall be explained by investigating their human nature. They suggested, moreover, that the subject's negative emotions might be the proximate source of their willingness to punish. These emotions would function as a proximate mechanism of altruistic punishment on the individual level. Later, this suggestion was confirmed and amplified by neurobiological evidence.

3 The neurobiological explanation of altruistic behavior and its use in economics

The neurobiological investigation of altruistic punishment can give insights into the basic structure of the motivational systems that underlie an individual's protection of shared group interests. Thus, experimental economists developed the design for an experiment where both could be done: an observation of behavioral responses toward the rules and institutions of experimental setting and an investigation of the correlating brain activations. The aim of this combined research strategy was to be able to predict how social efficiency of the outcomes of an interaction could be sustained.

In a follow-up study[11] to the first punishment-experiment, the experimenters added the following neurobiological tool: The punishing subjects were brain-scanned by using positron emission tomography (PET). The procedure was the following: The subjects were placed into a scanner immediately after the interaction with another player was over. The scanning started when subjects learned about the free-riding and non-cooperative behavior of the other player, and it was finished when they had determined the punishment. In the observation of neural circuits of the subject's brain, it could be shown that negative as well as positive emotions are in the play when people decide whether they want to take personal costs to punish a free-rider. The finding of the observation was the following: First, the free-rider's behavior caused anger to the punishing subject. But during the period when subjects decided whether they want to punish, their anticipation of the consequences of punishment was associated with a strong feeling of reward. Hence, experimenters interpreted the finding as evidence for emotional satisfaction as being the benefit altruistic punishers weigh against the costs of punishing.[12]

This conclusion about the motivation of the punishers could be drawn, because brain areas (*nucleus caudate*) which are responsible for reward-related considerations were activated during decision-making. Thus, experimenters concluded that the subject's decision-making was driven by hedonic motivation – according to the neurobiological processes underlying this motivation. Hedonic motivation is one of the key figures in an evolutionary explanation of behavior, because there is natural selection for avoiding pain and bodily injuries. Hedonism might be one of the motivational mechanisms which are generated by an evolutionary framework. Therefore, the correlation of hedonic

[11] Dominique J.-F. de Quervain et al., "The Neural Basis of Altruistic Punishment," *Science* 305 (2004): 1254-1258.
[12] Dominique J.-F. de Quervain et al., "The Neural Basis of Altruistic Punishment," *Science* 305 (2004): 1257.

motivation and altruistic punishment might function as a proximate mechanism of the evolution of human prosociality: But this has to be explored further in future research.

4 Summary: The significance of altruistic punishment

As we have seen, the pattern of altruistic punishment is different from reciprocal (direct) and reputation-based (indirect) altruism as investigated in evolutionary biology. Its actualization in human behavior is dependent on the willingness of an individual to incur personal costs in order to sanction another for his/her norm violation or social free-riding behavior. This cost is never likely to be recovered. The punisher is, therefore, referred to as an altruistic person (in a biological sense), because his/her behavior increases the average cooperation level of group interaction in the long run. In the perspective of neuroeconomics, altruistic punishment is among the proximate (individual) causes of human evolution. It is due to a neural mechanism which explains why human species has such a unique high degree of cooperation which is missing among all other species. The investigation of the neural mechanism of altruistic punishment disclosed that there is not only cost, but also benefit to the punisher: he experiences a strong and emotional feeling of satisfaction when expecting the free-rider to be punished. Thus, the understanding of altruism deployed in economics does only work in terms of the consequences of behavior. It cannot be referred to as an account to altruism in a psychological sense!

5 Connecting Questions and Philosophical Assessment: The Paradigm of ›Abu Ghraib‹

To understand the kind of altruism that is unique in humans, neuroscientists and experimental economists have developed a trans-disciplinary experimental research design which allows studying the major behavioral patterns of social exchange situations. But can, for instance, evidence for an evolutionary mechanism behind altruistic punishment and norm enforcement account for such complex society systems and cultural institutions as the law and penal system? To give an example: How is the behavioral pattern of altruistic punishment being demarcated from the behavior that was exhibited in Abu Ghraib when the institution of penalty by law became an excuse for penalty against law? To answer this question we have to ask first: In what sense is the individual act of altruistic punishment and norm enforcement different from spitefulness and ›schadenfreude‹?

Although it is clear that the paradigm of ›Abu Ghraib‹ does not fit perfectly to the pattern of altruistic punishment, it provides a good scenario to question the value of neuroeconomic predictions. However, the moral problem of Abu Ghraib is reflected in the question: For what purpose did the prison guards punish the prisoners? Did they want to maintain the law and order system in a somewhat extreme sense or did they merely want to satisfy their own sadistic appetite? Undoubtedly, there were personal costs to the punishers: Some of them lost their job, others have been cautioned. But undoubtedly as well, there were some benefits such as social reputation among the other guards and so on.

From a moral point of view, it is wrong that social order is maintained by deploying people's spitefulness and willingness to harm others – even if they harm others on their own cost. As well, the proof that punishers do so out of a hedonic

motivation is not the crucial point that helps any further. Rather, it is the assessment of context and situation that can give a hint whether a certain act of punishment has a positive or a negative impact on the welfare of society. We cannot investigate the behavioral patterns of social norm enforcement without taking into account that they are always interrelated and correlated to cultural patterns of behavior. And the most important cultural pattern is, of course, human communication, which opens up for freedom within human behavior. It is not enough to grasp the real life behavior of human beings by naturalizing their motivational forces, although it is a very interesting and valuable task.