Chapter 14
The Neurobiology of Altruistic Punishment: A Moral Assessment of its Social Utility

Rebekka A. Klein

14.1 Introduction

The article deals with the experimental model of altruistic punishment and social norm enforcement which has recently been designed in the fields of experimental economics and neuroeconomics. By using this model, neurobiologists and economists investigate the close relationship between neurobiological mechanisms in the brain and the enforcement of cooperation norms in human social behavior. They have shown experimentally that the implementation of a costly punishment tool in social dilemma experiments provides strong evidence for the impact of altruistic and prosocial behaviors at the level of group interaction and cooperation. The biological and behavioral interpretation of this evidence will be critically questioned in this article from the point of view of moral philosophy. The following argument will be presented: an exclusive concern for biological motivational mechanisms and behavioral outcomes of punishment fails to discriminate between good and bad punishment in a moral and legal sense, because it does not provide us with an appropriate criterion by which to evaluate the social utility of punishment. Hence, the moral aspects of this behavior have to enter the picture in order to allow us to arrive at a full judgment on its social utility.

R.A. Klein (✉)
Institute for Systematic Theology, University of Halle-Wittenberg,
06099 Halle/S., Germany
e-mail: kleinrebekka@hotmail.com

K.S. Plaisance and T.A.C. Reydon (eds.), *Philosophy of Behavioral Biology*,
Boston Studies in the Philosophy of Science 282, DOI 10.1007/978-94-007-1951-4_14,

14.2 The Study of Altruism in Experimental Economics and Neuroeconomics

The understanding of altruism in experimental economics is based on a consideration with regard to the economy of human behavior. It says: if a human being is altruistic he will incur personal costs in order to increase the benefit of other individuals.[1] Hence, the economic notion of altruism differs from the biological concept by remaining basically on the individual level[2], whereas biologists account for altruistic behavior in terms of Darwinian fitness, group selection and the number of offspring. In addition, the economic notion of altruism can be distinguished from the notion of altruism in psychology and philosophy because it does not deal with the motives, e.g. the beliefs, desires, and reasons behind actual behavior which are crucial for calling a behavior altruistic in psychology and philosophy.[3] Instead, the economic notion of altruism focuses on the outcomes of behavior and measures them in terms of costs and benefits to the individual.

Recently, economists have applied this concept of altruism to the study of human social behavior in experiments that have been conducted according to behavioral game theory[4]. Their observation of social interactions and transactions in social dilemma games was guided by an interest in modeling the social preferences of individuals. In economic theory, preferences are used to measure people's choices and their valuations of certain goods such as food, money, prestige, etc.[5] To determine the actual preferences of people, economists observe people's choices in an experimental environment in which real money is at stake. They do so because they particularly focus on the monetary outcomes of behavior. With regard to social preferences, these outcomes have to have a relation to other individuals' outcomes, and thus are referred to as 'social.'

However, the growing interest on the part of experimental economists in the study of altruism and social preferences is a rather provocative enterprise within their own discipline, because the standard approach in economics, the neoclassical theory of human behavior, usually does not take into account non-selfish, altruistic or even social preferences. Instead, they assume that human rational behavior is exclusively exhibited in the form of egoistic rational choices (the homo economicus model of human agency). This rather narrow understanding of human behavior, which reduces it to the economic principle of self-interested profit-maximization, is due to the habit of neoclassical theory to found its explanation of human behavior

[1] See Fehr & Fischbacher (2003), 785.

[2] See a more detailed analysis of the three different notions of altruism in biology, psychology, and economics in Clavien & Klein (2010). The authors investigate the contribution of experimental economics and neuroeconomics to the debate on psychological altruism, and point out that so far there is neither evidence for nor against psychological altruism in economic experiments.

[3] See the difference between biological and psychological altruism in Sober & Wilson (1998).

[4] An introduction in behavioral game theory can be found in Camerer (2003).

[5] See Camerer & Fehr (2004), 55.

on simple axiomatic approaches.[6] This has led to a model of human agency which is not at all convincing as regards human psychology. To show that the neoclassical approach does not provide proper tools to account for actual human behavior, some experimental economists have begun to systematically test these axiomatic assumptions about human agency in the field and in laboratory experiments using a game-theoretical framework. Their work, which has been done in conjunction with anthropologists and ethnologists, has shown that the assumption of a purely selfish rational agent is not appropriate in most human societies across the globe.[7]

In addition to the behavioral experiments, some experimental economists have also tried to reinforce their view of human social behavior by naturalizing human agency, and investigating the biological roots of people's choices. To this end, they integrated new methodologies and research strategies from the natural sciences into their experimental framework, and participated in the foundation of the trans-disciplinary research approach of 'neuroeconomics.'[8] This approach allows the combination of the methodological tools of neuroscience and those of experimental economics in a shared experimental environment.[9] It can help to uncover the psychological motivational mechanisms behind people's choices, and is very useful in terms of integrating psychological parameters into the economic model of human behavior. Nonetheless, neuroeconomics as a behavioral and brain science does not claim that neoclassical economics is wrong as a whole, but that all its theoretical assumptions and predictions of human behavior can be verified or falsified by empirical research. One of the main objectives of neuroeconomics thus is to provide "…an alternative theoretical approach for predicting behavior and a methodology for testing those theories."[10]

Beyond the engagement with its own discipline, neuroeconomics participates in the major endeavor of explaining the nature of human altruism and the evolution of cooperation across human species. From the point of view of evolutionary anthropology, human cooperation not only differs from non-human mammalian species with respect to intensity and frequency, but 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.[11] As a consequence,

[6] Glimcher et al. (2009) give a short introduction into the history and development of neurobiological studies in economics and refer to the axiomatic approach of neoclassical economics as one of the main causes of this development.

[7] An overview of the field experiments on social preferences can be found in Henrich et al. (2004). This book documents a global study on the validity of cooperation and fairness norms in social exchange practices. It shows that the economic assumption that individuals exhibit purely selfish preferences in their behavior is violated in all of the fifteen small-scale societies that have been investigated.

[8] See Glimcher et al. (2009) for how wide-spread the approach of neuroeconomics is and the different research questions it can be applied to.

[9] See Gintis (2007).

[10] Glimcher et al. (2009), 6.

[11] See Henrich & Henrich (2006), 223-224.

humans live in large-scale societies which are built on anonymous encounters between genetically non-related individuals.[12] Human cooperation flourishes in these societies in spite of anonymity and non-relatedness, because group interaction is based on social norms. Stability and coordination in social interaction among humans is, therefore, established through the enforcement of norms.

In modern societies, this enforcement is done in two ways. In the case of legal norms, these norms are maintained because their violation is formally sanctioned by the law and penalty system of society. In the case of social norms, which back up the enforcement of legal norms by providing an informal basis for them, enforcement takes place in an autonomous and self-organized process of monitoring and control in local communities, as has been shown by Elinor Ostrom's field studies in the 1990s.[13] Ostrom studied independent systems of social monitoring and control in several long-standing common property regimes, including Swiss grazing pastures, Japanese forests, and irrigation systems in Spain and the Philippines. She could show that the establishment of cooperative institutions in these regimes is organized by the resource users themselves. Hence, the maintenance of social norms and their adaptation as rules of behavior is not secured through formal sanctions by state policy, but through self-governance such as, for instance, social monitoring and interpersonal sanctioning in local communities and (ethnic) groups.

Starting from this insight, the sciences of experimental economics and neuroeconomics have developed a wide range of experimental tools to study the relevant behavioral patterns of social norm enforcement. For obvious reasons, they account for norms in terms of social preferences and individual choices.[14] Thus, they investigate the maintenance of norms as a "second-order public good"[15] in social interaction. By definition, goods are referred to as 'public' in experimental economics if every individual participating in the interaction has a benefit from them "…including those who did not pay any costs of providing the good."[16] Thus, public goods such as natural resources or social infrastructure in human societies are prone to be exploited by free-riders and have to be protected by social norms which govern their use. But norms cannot be chosen by people in the same way as material goods are. Rather, they have to be established and monitored as stable behavioral patterns through the initiative of individuals. Thus, they are not given in advance, but are constituted in social interaction ('second-order public goods'). The behavior of altruistic punishment, which will be focused upon in this article, has been proven to be one of the key patterns for the maintenance of social norms in human interaction.

[12] See Fehr & Fischbacher (2004), 185.

[13] See Ostrom (1990) and Ostrom et al. (1992).

[14] For a philosophical concept of social norms which is in accordance with game theory, see Bicchieri (2006). Bicchieri also integrates various psychological dispositions in her model of norms as preferences of the individual. Thus, her account might also be very valuable for the study of norms in neuroeconomics.

[15] Fehr & Gächter (2002), 137.

[16] Ibid.

14.3 The Correlation of Norm Enforcement and Altruistic Punishment

Several experimental studies on cooperation and prosociality in economics have shown that altruistic punishment plays a key role in understanding the evolution of norm enforcement in human societies.[17] Altruistic punishment does not directly benefit the welfare of an individual person, but society as a whole. Therefore, it is referred to as a 'prosocial' behavior. The term 'prosociality' is used in experimental economics to indicate a behavior that does not directly benefit others (as does cooperation), but the well-being of group interaction as a whole.[18] The behavioral pattern of altruistic punishment has been clearly shown to be of great significance for the study of prosociality in a series of behavioral experiments in economics and neuroeconomics.[19] These have been conducted in different behavioral laboratories since the first study on altruistic punishment was published by Ernst Fehr and Simon Gächter in 2002.[20]

In this study, altruistic punishment is defined as a non-selfish act of punishment which "[provides] …a material benefit for the future interaction partners of the punished subject but not for the punisher."[21] In an experimental setup with 240 participants[22] at the University of Zurich, Fehr and Gächter tested their subject's individual willingness to punish altruistically in a 'public goods' experiment. In this type of experiment, several people have the option of investing a certain amount of money in a group project. Afterwards, the sum of all contributions is to be shared among the group members equally. The experiment in Zurich was conducted in twelve sessions and the group composition was changed after each session. The latter guaranteed that none of the subjects could again meet the same subjects during the experiment. This ensured that the subjects' 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 at the end of each session. In order to test whether the subjects' willingness to punish did include the willingness to suffer personal cost, the punishment was not only costly for the free-rider, but also for the punishing subject himself, because he had to pay for it from his own gain.

---

[17] The claim that social reciprocity (prosocial norm enforcement) provides the best explanation for the evolution of punishing behaviors has been defended in Carpenter et al. (2004).

[18] A definition of the distinction between prosociality and cooperation can be found in Henrich & Henrich (2006). For a model explaining the cultural evolution of prosociality and cooperation see Gintis (2003).

[19] Fehr & Gächter (2002); Fehr & Fischbacher (2003); Fehr & Rockenbach (2003); De Quervain et al. (2004). An assessment of the evolutionary origin of altruistic punishment can be found in Boyd et al. (2003).

[20] Fehr & Gächter (2002).

[21] Ibid., 139.

[22] All of the participants in the experiment were undergraduate students from the University of Zurich.

The results of the experiment were as follows: over twelve sessions, 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.[23] A minority of 9.3% of the subjects punished more than ten times. Thus, the experimental results provide strong evidence that altruistic punishment is a stable behavioral pattern among humans. Additionally, a significant effect of altruistic punishment was shown in the later sessions of the experiment. After having been punished, the punished subjects invested a higher amount of money in the group project and changed from non-cooperative to cooperative behaviors in the following sessions. Thus, altruistic punishment caused a substantial increase in terms of the average cooperation level of the group over time. This was highly correlated with the subject's investment strategies and can, therefore, be considered among the facilitating conditions of the evolution of human cooperation. Hence, the remarkable result of the study by Fehr & Gächter (2002) was that the opportunity to punish free-riders altruistically has a significant impact on the maintenance of the norm of cooperation and equity, even in anonymous encounters.

With regard to the interpretation of this evidence, the experimenters suggested that the evolution of social norms has to be explained further in terms of the level of the individual's preferences. Thus, the experimenters asked how the willingness to punish might be triggered on a psychological level. As a suggestion, they hypothesized that the subjects' negative emotions concerning the free-riding behavior of others might be the source of their decision to punish. Emotions such as anger and outrage could provide a proximate mechanism of altruistic punishment.[24]

To elicit the correlation between punishment and the individual's emotions, the experimenters prepared a questionnaire which was given to the subjects after the experiment, and asked them to indicate their intensity of anger concerning the free-riding behavior on a seven-point scale. As a result, 47% of the subjects indicated the highest intensity of anger. Hence, the experimenters concluded that these emotions might be a psychological trigger for punishment. This led them to seek a research tool to further investigate this correlation, which in turn led them to engage in a new research field investigating the neurobiology of prosocial and cooperative behaviors in humans.

14.4 The Neurobiological Explanation of Altruistic Punishment

In a follow-up study[25] to the first experiment on altruistic punishment in 2004, economists Ernst Fehr and Urs Fischbacher started to work together with neuropsychologists for the first time. They added a neuroimaging tool to the experimental

[23] See Fehr & Gächter (2002), 137.

[24] A definition of proximate causes of evolution can be found in Mayr (1961), 1503.

[25] De Quervain et al. (2004).

setup of their study on social norm enforcement, and observed the neurological foundations of people's choices. The idea of combining experiments on norm enforcement with the neurological investigation of the human mind had already come up in a study in 2003 when neuroscientists Alan Sanfey, James Rilling and colleagues adapted an experimental design from economics, and started to investigate the neural substrates of the cognitive and emotional processes involved in decision-making concerning altruistic punishment. After they brain-scanned the subjects with functional magnetic resonance imaging (fMRI), they found an increased activity in the 'anterior insula'—a brain area associated with negative emotional feelings. Hence, they concluded that emotions might be the psychological and neurological driving force behind this behavior, a view which was still consistent with the 2002 findings of Fehr and Gächter.

However, the follow-up study by Fehr, Fischbacher and de Quervain in 2004 led to a rather different neurological finding. The procedure of this experiment was as follows: the subjects were brain-scanned during their decision to punish free-riding behavior by using positron emission tomography (PET). They were placed in a scanner immediately after the interaction with another player was over. The scanning started when subjects learned about the free-riding behavior of the other participant and it finished when they had determined the punishment. In the observation of the neural circuits of the subjects' brains, it could be shown that not the 'interior insula,' but a brain area linked to the anticipation of reward—the 'caudate nucleus'—played a prominent role when people decided to punish. Subjects who exhibited stronger activation of the 'caudate nucleus' were ready to incur more personal costs to punish a free-rider in comparison with subjects who exhibited low caudate activation. Hence, the experimenters interpreted the finding as evidence of the anticipation of "hedonic rewards"[26] being the benefit that altruistic punishers weigh against the costs of punishing. The punishing subjects seemed to feel relief when the violated social norm was established again through an act of retributive justice.

Thus, experimenters concluded that, according to the underlying neurological processes, the subjects' decision-making was driven by hedonic motivation. Hedonic motivation is one of the key features in an evolutionary explanation of behavior, because there is natural selection for avoiding pain and unpleasantness. Therefore, the correlation between hedonic motivation and altruistic punishment might function as a proximate mechanism of the evolution of human cooperation. But this has to be explored further in future research, and cannot be concluded from a single study.

In my view, a much more pressing question with regard to the interpretation of the result of the neuroeconomics study concerns the assignment of psychological motivational states to the neurological findings, and their validity for determining the social utility of this behavior. My question is whether it is really justifiable to conclude from the consequentialist and neurobiological explanation of punishment in neuroeconomics that punishment is a prosocial and thus beneficial act in terms of the welfare of human society. In the following sections of the paper, I will try to cast

[26] Ibid., 1257.

some doubt on this conclusion. I will show that the behavioral and neurobiological explanation of punishment might lead to a shortened (reduced) judgment when it comes to determining the social utility of this behavior. Thus, external reasoning about its motivation and consequences has to be integrated into the picture in order to form a judgment about the purely positive evaluation of its prosocial outcomes.

As we have seen, the behavioral pattern of altruistic punishment as investigated in economics is different from that of reciprocal (direct) and reputation-based (indirect) altruism as investigated in evolutionary biology. Its manifestation in human behavior is dependent on the revealed preference of an individual to incurring personal costs which are never likely to be recovered, in order to sanction another for his norm violation or social free-riding behavior. Thus, the punisher is referred to as an altruistic person in a consequentialist sense which means that his personal motivation for the decision to punish does not enter the picture. The study by Fehr and Gächter (2002) has shown that this kind of altruistic behavior has a remarkable effect on human interaction: it increases the average cooperation level of group interaction in the long run. From the perspective of neuroeconomics, altruistic punishment is among the proximate (individual) causes of the evolution of human cooperation and is due to a neural mechanism which explains why the human species maintains such a high degree of cooperation among non-relative individuals, which is different in kind from that of all other species.[27]

But the investigation of the neural mechanism underlying altruistic punishment has also shown that there is not only cost but also benefit to the punisher: he anticipates a strong feeling of satisfaction when expecting the free-rider to be punished and the norm of cooperation and equity being re-established. Thus, a behavior which is altruistic in the consequentialist sense seems to be motivated by hedonic reward anticipation on the psychological level. Thus, the study is ambivalent in its result: the individual's motivation for altruistic punishment is obviously self-concerned in the first place. During decision-making, the punishing subject anticipates his own state of mind which will occur after the punishment is carried out.[28] Hence, as several interpretations of the neuroeconomic study of de Quervain and colleagues (2004) have shown, it is not absolutely clear from the neurological findings whether the punishing subject's feeling of satisfaction is primarily related to the (indirect) establishment of the cooperation norm, or whether it is primarily related to a desire for revenge—longing for a compensation of the cost he has suffered as a result of the initial social free-riding. In other words: is the motivation for altruistic punishment grounded in a desire for social norm enforcement or a desire for revenge?

Unfortunately, no further neuroeconomic research has been done to answer this question concerning the psychological motive underlying altruistic punishment.[29]

---

[27] See Fehr & Fischbacher (2003); Fehr & Fischbacher (2005).

[28] Knutson (2004) has already pointed towards this ambivalence of the study's results. The claim that there is no evidence explaining the causal chain of motivation behind the behavior is developed further in Clavien & Klein (2010).

[29] For a distinction between motive and motivation see the article on "Altruistic Emotional Motivation" by Christine Clavien in this volume.

And from the point of view of neuroeconomics, the question might also be irrelevant because the outcome of both of these motives, the enforcement of a social norm, is the same. On purely consequentialist grounds, it doesn't matter that the enforcement of a social norm is merely a secondary (instrumental) motive or even an unintended outcome of people's choices. The only thing that matters is whether the causal chain that leads to this outcome works reliably. Whether it is grounded in self-concern, or even selfish motivational states or motives, does not influence the evaluation of the prosocial outcomes of punishment behavior. But the disregard of the issue of motivation makes the use of the term 'altruism' with respect to punishment behaviors in economics highly questionable.

In contrast to the view that the neurobiological investigation of motivational states is sufficient to judge on the social utility of punishment behavior, I will point out now that the neuroeconomic approach is too short-sighted. In the following section I will show that—in contradiction to the neuroeconomic interpretation—the proof that punishers act out of hedonic motivation is the crucial point when it comes to the moral assessment of the consequences of punishment behavior. My thesis will be based on the argument that the motivation for a punitive act, and the motive behind that act, are not negligible in an assessment of its consequences. Hence, I have to clarify in what sense the questions of motive and motivation are crucial concerning the distinction between good and bad punishment in a moral and legal sense.

14.5 Moral Philosophical Assessment of Altruistic Punishment

In this section, I will introduce the moral perspective of judgment on social behaviors as a supplementary approach to the behavioral and brain sciences. The moral philosophical approach adds certain crucial aspects to the experimental study of behavior, whose understanding and explanation will improve the evaluation of its social utility and will help to avoid misjudgments concerning its overall prosocial consequences. The moral assessment of human behaviors not only deals with the question of whether certain behaviors have a prosocial or antisocial outcome concerning the common good or society's welfare, but also concerning the welfare of a single individual. Hence, it judges the social utility of human behaviors, not only in terms of 'general others,' which are represented by the anonymous social structures and institutions of society, but also in terms of 'concrete others,' who are affected in their individual well-being by the actions of others.

In this regard, the moral motive behind a punitive act shapes the social character and outcome of this behavior in a twofold sense: (a) it marks the boundary of the punitive act as regards its consequences for the well-being of concrete others, and (b) it prevents punishment from becoming an act of sheer violence which goes awry in the sense that it is extended beyond the scope of the moral and legal measures of social interaction. Hence, the motive or intention behind punishment is crucial for determining how it is acted out with respect to others as regards their individual

right to well-being and intactness (a), and their individual right to the adequacy of punishment of their offence (b). Thus, the empirical question of who is harmed, the extent of such harm and whether this can count as a prosocial or antisocial act, cannot be answered from a moral perspective without taking into account the intention and motivation accompanying the punitive act on the part of the punisher.

The moral question concerning punishment becomes even more pressing when we recognize that in every act of punishment—whether it is legally justified or not—there is individual leeway with regard to how the one who imposes the sanction can strengthen or weaken its consequences for others. Sometimes this leeway is acted out by the individual in the form of a very subtle psychological mistreatment of the other, and sometimes it is done in a very offensive and exposing way, involving dehumanization. Nonetheless, both modes can count among the varieties of human cruelty, insofar as they violate the individual's well-being and intactness with lasting effect.

To consider an example for the question I have in mind, we can see how the behavioral pattern of prosocial and altruistic punishment is demarcated from the sadistic behavior that was exhibited in Abu Ghraib Prison in Iraq in 2004. In Abu Ghraib, the societal institution of penalty became an excuse and a means for an excess of sheer violence.[30] The imprisoned criminals were held in a kind of lawless state. They were physically tortured and sexually abused by their prison guards. Although this treatment violated the norms of prisoner treatment outlined in the 'Geneva Convention' (1949), it was well known and accepted among the military police authorities and in the U.S. government.[31] The guards could, therefore, rely on official tolerance or, rather, official neglect of their behavior.

Abu Ghraib represents punishment which is certainly not beneficial to society because the legal institution of punishment is turned into its opposite—a violation of legal norms. Although the imprisoned criminals of war might have legally deserved punishment in terms of imprisonment, they received a much harder (physical) punishment than the one that would have been legally imposed on them—including acts of debasement and dehumanization. What is interesting about the case in the context of my argument is the following: the unlegislated punishment became possible because the prison guards established a social norm among their group members, considering it acceptable to punish the prisoners in order to nourish their own sadistic appetites. Maybe their behavior was rationalized afterwards by arguing that the prisoners deserved this kind of punishment because they are criminals. Hence, the prison guards considered it as a collective goal to maximize their pleasure at the cost of others who do not share their religious, national and ethnic background and who have failed to be respected as human beings in terms of their human dignity.

The incidents in Abu Ghraib show how important it is to safeguard the notion that the purpose of punishment in society is to enforce social norms which do not

[30] See Taguba (2004). The *Taguba Report* on the torture scandal in Abu Ghraib judges the behavior of the prison guards from the point of view of the *Geneva Convention Relative to the Treatment of Prisoners of War* (1949).

[31] See the discussion in Denner (2004).

violate the moral or legal norms of egalitarian cooperation. This is because the latter are also established precisely in order to protect individual's rights. In the case of Abu Ghraib, norms were not officially established. Hence, a form of self-governance took place among the group members. In this regard, the situation in Abu Ghraib is similar but not identical to the paradigmatic case study of social norm enforcement in the 'public goods' experiments. The crucial difference between the 'public goods' situation and the situation in Abu Ghraib in terms of societal welfare is that it was not a prosocial but an antisocial norm[32] that evolved out of the lawless state people were placed in. However, one could argue that this was due to the circumstance that prison guards and prisoners were placed in an environment where the one side had executive power over the other, whereas in 'public goods' situations, all individuals belong to the same group and therefore start from a level of egalitarian interaction. I think this argument points in the right direction, and can be substantiated by experimental evidence.

In a large-scale field experiment with fifteen native societies around the world, it has already been shown that differences of culture matter a lot when it comes to social norm enforcement.[33] Furthermore, there is evidence from a field experiment in Papua New Guinea (Bernhard et al. 2006) that the enforcement of norms across the boundaries of culture, nation and group membership seems to work less effectively than in 'public goods' situations, where people belonging to the same group establish sanctioning behaviors which protect the commons that their collective life is dependent on. Hence, especially in intercultural and inter-group interactions on the local and global level, it is important to safeguard the idea that punishment, as a means of norm enforcement, should not have antisocial or inhuman side effects for individuals. Accordingly, one could argue that the solution to this problem might be that social norm enforcement has to be prevented from violating the superordinate norms of justice and equity which are universally held in all human societies. This means that the evolution of 'particularistic norms' among social groups has to be governed by the maintenance of universal 'societal norms' such as equity, fairness and reciprocity.

Focusing on the distinction between particularistic and universal norms would lead us in the end to a solution of the conflict between prosocial and antisocial consequences of punishment on the level of transnational political and legal institutions. The latter was certainly not under consideration when behavioral economists designed their experimental research tools to study the self-governance of social norm enforcement among individuals. In contrast to a non-individual, state-governed or cosmopolitan solution, they have proposed that it is not the responsibility of centralized institutions alone to govern the evolution of social norms. Rather, individuals and groups can develop a system of monitoring and controlling the maintenance of

[32] The norm is antisocial only with respect to the wider group of people that includes the guards as well as the prisoners. With respect to the population of the guards alone, the norm is actually prosocial, because it increases their status. Hence, the fact that a particular action is prosocial with respect to a limited peer group does not say that it is morally unproblematic in general.

[33] See Henrich et al. (2004).

norms for themselves. Hence, to take the experimental economist's research work seriously, and to account for the distinction of socially beneficial and non-beneficial punishment on the individual level, we again have to look more closely at the individual rationale and psychological motivation of the punishing subjects.

Beyond the institutional level it is the responsibility of individuals to prevent social norm enforcement through punishment developing into acts with antisocial side effects. Coming back to the case of Abu Ghraib we can ask: did the prison guards' desire to satisfy their own sadistic appetites simply override their rational faculties with regard to weighing the costs and benefits of punishment against each other? Or did they establish an individual rationale for their behavior which made it reasonable to establish an antisocial norm, promoting collective fulfillment of their sadistic appetites at the cost of others? Undoubtedly, there were personal costs to the prison guards in Abu Ghraib: they risked prosecution and eventually lost their jobs and were accused of breaking international law. But undoubtedly, there were some benefits for them as well: the punishers expected reputational gain from their fellows when they abused prisoners while also fulfilling their own sadistic appetites.

Although it represents a different kind of punishment than the one that was in mind in the economic experiments on altruistic punishment, the torture scandal in Abu Ghraib is a good example of the dangers of highlighting the prosocial consequences of punishment in a purely consequentialist sense.[34] The paradigmatic case shows how the establishment of a social norm among group members can turn into a norm violation itself: the violation of human rights. Of course, such a situation as in Abu Ghraib was not modeled in the neuroeconomic experiment presented earlier (De Quervain et al. 2004). The experimental setting only allowed for financial punishment which means that the degree of harm which could be imposed on a non-cooperative subject was limited and controlled externally, ensuring that the punishing subject could not overrun the given conditions of punishment. This means that the punishment in the experiment was not shaped by the punishing subject's individual preferences determining the mode of punishment (psychological, physical, financial, etc.) and its heaviness was not independent of the experimenter's setting. Thus, an immoral excess of punishment, i.e. a punitive act which overrides the boundaries which the moral sense of the other imposes on punishment, could not even be modeled.

Furthermore, the argument could be raised that the experimental study by De Quervain et al. (2004) neglected to pose the morally crucial question of how the motivational states of punishing subjects might shape and influence the different modes of punishment, as well as the difference between excessive and limited punishment. The study simply did not take into account the fact that the motivational states of the subjects might make a difference with regard to the act of punishment itself. However,

[34] See the experiments related to punishment in prison in Milgram (1963). As far as I can see, the experimental economic study of punishment has not been related to this social psychology study of the excess of physical punishment.

the neurobiological investigation of these motivational states has positively shown that the punishing subjects are looking for some personal benefit besides the prosocial effect of their punishment behavior. They weigh the material costs of punishment (personal loss) against its being the cause of a feeling of reward. But it remains rather unclear in the neuroeconomic study whether the anticipated reward is an appraisal of the social utility of the punishment (norm enforcement) or a cause for personal satisfaction (revenge). Nonetheless, experimental economics and neuroeconomics claim that the objective utility of this behavior which can be observed in its outcome (increasing the average cooperation level over time) is sufficient to appraise punishment as a social utility tool in human societies.

Contrary to this position, I have claimed that a critical moral evaluation of the social utility of punishment has to start from the negative observation that the punishment of people who deserve it in terms of their preceding antisocial behavior is not in itself a socially valuable act. The moral costs of punishment may outweigh the social benefit, because punishment always involves someone being harmed—either physically, psychologically or financially. This raises the moral question of a possible violation of an individual's rights associated with, or even inherent to, punishment—a problem which is more or less concealed in the euphemistic term 'altruistic punishment.' Since it is the major distinction between a liberal and a dictatorial concept of society that cooperation and 'public goods' are not maintained to the disadvantage of individual's rights, we have to ask for a justification of any cost the punisher imposes upon others. In order to consider punishment as 'prosocial,' it is not enough—as the neuroeconomic concept of altruism claims—that we ensure that the punisher obtains no reputational or financial benefit from the punishment, especially if there is evidence of a hedonic reward mechanism governing his decision-making. The fact that there is a material cost to the punisher does not safeguard that his behavior will not have intolerable antisocial side effects in terms of the outcome, for example when punishment is acted out in order to satisfy a sadistic appetite.

Hence, the moral assessment of punishment requires external reasoning about the motivation and intentions of the punishing subject. This reasoning can, of course, be substantiated or falsified by neurological findings, but it is, in principle, indispensable when it comes to an evaluation of punishment behaviors from a moral perspective. The moral question concerning the motivation of punishment is whether the punishing subject is still concerned with the individual welfare of the other. This moral concern for the other should occur even when the punished subject deserves punishment, because it prevents the punitive act from going awry, i.e. turning out to have ambiguous, prosocial and antisocial consequences at the same time.

14.6 Punishment and the Welfare of a Just Society

In the preceding section, I have shown why the question of motive or intention behind punishment is not insignificant. I have argued that the distinction between justified punishments and acts of unjustified violence shall be upheld by external

reasoning about the motivation for, and motive behind, a punitive act. By considering neuroeconomic findings about the motivational causes of a punitive act, I have pointed out that to harm someone for a good reason can include the motive of revenge as well as the motive of preventing further norm violations. The first one is a selfish motive leading (instrumentally) to prosocial consequences, whereas the second is a purely prosocial motive. In this section, I will argue that the only way to ensure that norm enforcement by punishment has purely prosocial outcomes on both the individual and the societal level, is not to show that it is altruistic, but to prove that it is primarily driven by an egalitarian motive. Hence, the moral assessment of punishment has to distinguish between (a) punishment as a means of establishing egalitarian cooperation and (b) punishment as an excess of sheer violence (retaliation, revenge, sadistic appetite and the like).

Thus, the following steps associated with the assessment of punitive acts have to guide the evaluation of its social utility: (a) determine the (neurobiological) motivational causes of the punitive act, (b) look for a moral concern included in these motivational causes, (c) clarify the intention or motive behind the punishment behavior, (d) consider the conformity of the motive to moral and legal norms of a just society's welfare, (e) evaluate the prosocial or antisocial consequences of a punitive act. It should be obvious from these assessment steps that the investigation of the neurobiological motivational causes of behavior alone has not clarified the prosociality of the intention or motive behind punishment behavior. A moral assessment of these motivational causes is needed in order to provide a valid judgment of its social utility. But nor is this enough. One more step has to follow: the consequences of punishment behavior as well as its motivational causes have to be assessed from a moral perspective. Let me point out briefly how this can be done by bringing together the experimental economist's and the moral philosopher's approaches.

Instead of using the term 'prosociality,' moral and social theory account for the welfare of a just society in terms of a high level of egalitarian cooperation, because they conceive of justice as the equal distribution of the liberties and rights of individuals in societal cooperation. Thus, egalitarian cooperation is cooperation which aims at producing social and economic equity in society without violating the rights of the individual. Hence, moral and social theory judges the social utility of punishment behavior in terms of its contribution to the maintenance and increasing of egalitarian cooperation. In order to determine what this contribution is, moral and social theory requires the justification of an agent's decision to punish, by investigating his underlying 'egalitarian motives.'[35] In the behavioral study of prosocial decision-making, egalitarian motives can be represented by modeling the consequences of behavior in terms of an increase in equality and a decrease in inequality in the aggregate level of distribution.[36] Hence, egalitarian motives are correlated to

[35] See the behavioral experiment on egalitarian motives in Dawes et al. (2007). For future research, it would be necessary to investigate the neurobiological underpinnings of this behavioral model of egalitarian motives.

[36] See Masclet & Villeval (2008).

the financial outcome of social decision-making. They can be inferred from the equal distribution of an exchanged good. In the experimental study by de Quervain et al. (2004), which investigated the biological motivational causes of punishment, no egalitarian motives were in play. The punishing subjects did not align the other subject's outcome to their own by punishment. Rather, they decreased both outcomes equally in order to harm the social free-rider. Hence, their punishment was not grounded in egalitarian motives which, at very least, casts a severe doubt on its social utility from the perspective of moral philosophy.

In this article, I have demonstrated how punishment and norm enforcement go awry when a moral concern is missing in the motivational causes of this behavior. I have shown how judgments on the moral nature of motivational causes change the interpretation of altruistic punishment as a socially beneficial behavior. Furthermore, I have argued that even the consequences of punishment cannot be judged to be socially valuable from the perspective of moral philosophy unless they are evaluated in the light of their underlying egalitarian motive. From these arguments, it can be concluded that the legitimacy of social norms is not bound to their effective enforcement, but requires some external reasoning about the motivation of, and motive behind, enforcement on the individual level. Otherwise, their legitimacy could not be demarcated from their misuse in a dictatorial system of social disciplinary power, which would extinguish the norms' crucial function of providing the 'breeding ground' of egalitarian cooperation within modern and democratic societies.

In a society's penal system, the political power of law is enforced by sanctions which are bound by the law. Thus, legal punishment is distinguished from illegal punishment by its conformity to the law, and not to some other subjective rationale. In contrast, punishment is not simply bound by the law as part of the society's system of informal social norm enforcement. Rather, in this context, bad punishment is demarcated from good punishment by its impact on the welfare of a just society. Hence, a society's system of informal norm enforcement is referred to as 'just' if it establishes welfare in terms of its 'public goods' as well as in terms of its individual's rights. Thus, there is a positive and a negative condition for the prosociality of punishment: in order to be 'prosocial' the punitive act has to (a) increase the average cooperation level of social interaction and it has (b) to do so by not violating the rights of individuals. Both conditions refer to the moral aspects of this behavior on the individual and at the societal level.

14.7 Summary of the Argument

In the experimental study of the motivational causes of punishment behavior, an exclusive concern for biological motivational mechanisms and behavioral outcomes of behavior fails to discriminate between good and bad punishment in a moral sense. The article has substantiated this claim concerning a recent study of the neurobiological underpinnings of altruistic punishment (De Quervain et al. 2004). This study has revealed the biological motivational causes of punitive acts with a background

in social norm violation. It has shown that the punishing subjects are looking for some personal benefit besides the prosocial effect of their behavior. Although this provides valuable new insights into the psychological motivation of altruistic subjects who anticipate satisfaction from the punishment of norm violation, it challenges the claim for the purely altruistic and socially beneficial nature of this behavior. The subject's anticipation of reward can refer to a selfish (e.g. retaliation, revenge) as well as a prosocial motive for punishment (norm enforcement). To handle this ambiguity, the article has claimed that external reasoning about the moral concern of punishment is required to judge its social utility. The use of the terms 'altruism' and 'prosociality' with respect to punishment in economics does merely conceal this ambiguity. Hence, the consideration of moral motives and intentions should enter the neurobiological and behavioral explanation of punishment behavior. The article concludes with the argument that the assessment of the intention or motive behind punishment behavior is not insignificant for the question as to whether it has a positive or a negative impact on the welfare of a just society.

References

Bernhard, H., Fischbacher, U., and Fehr, E. (2006): Parochial Altruism in Humans. *Nature* 442: 912–915.

Bicchieri, C. (2006): *The Grammar of Society. The Nature and Dynamics of Social Norms*. New York: Cambridge University Press.

Boyd, F.R., Gintis, H., Bowles, S., and Richerson, P.J. (2003): The Evolution of Altruistic Punishment. *Proceedings of the National Academy of Sciences of the United States of America* 100: 3531–3535.

Camerer, C.F. (2003): *Behavioral Game Theory. Experiments in Strategic Interaction*. Princeton, NJ: Princeton University Press.

Camerer, C.F., and Fehr, E. (2004): Measuring Social Norms and Preferences Using Experimental Games: A Guide for Social Scientists. In *Foundations of Human Sociality*, ed. J. Henrich et al. Oxford: Oxford University Press, 55–95.

Carpenter, J.P., Matthews, P.H., and Ong'ong'a, O. (2004): Why Punish? Social Reciprocity and the Enforcement of Prosocial Norms. *Journal of Evolutionary Economics* 14: 407–429.

Clavien, C., and Klein, R.A. (2010): Eager for Fairness or for Revenge? Psychological Altruism in Economics. *Economics & Philosophy* 26: 267–290.

Dawes, C.T., Fowler, J.H., Johnson, T., Mc Elreath, R., and Smirnov, O. (2007): Egalitarian Motives in Humans. *Nature* 446: 794–796.

De Quervain, D.J.F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., and Fehr, E. (2004): The Neural Basis of Altruistic Punishment. *Science* 305: 1254–1258.

Denner, M. (2004): *Torture and Truth: America, Abu Ghraib, and the War on Terror*. New York: New York Review Books.

Fehr, E., and Fischbacher, U. (2003): The Nature of Human Altruism. *Nature* 425 (2003): 785–791.

Fehr, E., and Fischbacher, U. (2004): Social Norms and Human Cooperation. *Trends in Cognitive Sciences* 8: 185–190.

Fehr, E., and Fischbacher, U. (2005): Human Altruism – Proximate Patterns and Evolutionary Origins. *Analyse & Kritik* 27: 6–47.

Fehr, E., and Gächter, S. (2002): Altruistic Punishment in Humans. *Nature* 415: 137–140.

Fehr, E., and Rockenbach, B. (2003): Detrimental Effects of Sanctions on Human Altruism. *Nature* 422: 137–140.

Gintis, H. (2003): Solving the Puzzle of Prosociality. *Rationality and Society* 15: 155-187.

Gintis, H. (2007): A Framework For the Unification of the Behavioral Sciences. *Behavioral Brain Science* 30: 1–61.

Glimcher, P.W., Camerer, C.F., Fehr, E., and Poldrack, R.A. (2009): Introduction: A Brief History of Neuroeconomics, in *Neuroeconomics. Decision-Making and the Brain*, ed. P. Glimcher et al. Amsterdam: Elsevier/Academic Press, 1–12.

Henrich, J., and Henrich, N. (2006): Culture, Evolution and the Puzzle of Human Cooperation. *Cognitive Systems Research* 7: 220–245.

Henrich, J., Boyd, R., Bowles, S., Camerer, C.F., Fehr, E., and Gintis, H. (2004): *Foundations of Human Sociality. Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. Oxford: Oxford University Press.

Knutson, B. (2004): Sweet Revenge? *Science* 305, 1246–1247.

Masclet, D., and Villeval, M.-C. (2008): Punishment, Inequality, and Welfare: A Public Good Experiment. *Social Choice and Welfare* 31: 475–502

Mayr, E. (1961): Cause and Effect in Biology. *Science* 134: 1501–1506.

Milgram, S. (1963): Behavioral Study of Obedience. *Journal of Abnormal and Social Psychology* 67: 371–378.

Ostrom, E. (1990): *Governing the Commons. The Evolution of Institutions for Collective Action*. New York: Cambridge University Press.

Ostrom, E., Walker, J., and Gardner, R. (1992): Covenants With and Without A Sword: Self-Governance Is Possible. *American Political Science Review* 86, 404–417.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. (2003): The Neural Basis of Economic Decision-Making in the Ultimatum Game. *Science* 300: 1755–1758.

Sober, E., and Wilson, D.S. (1998): *Unto Others. The Evolution and Psychology of Unselfish Behavior*. Cambridge, Mass.: Harvard University Press.

Taguba, A.M. (2004): The "Taguba Report" On Treatment Of *Abu Ghraib* Prisoners In Iraq. ARTICLE 15-6 Investigation of the 800th Military Police Brigade, May 2004, http://news.findlaw.com/hdocs/docs/iraq/tagubarpt.html (28.6.2008).