The Neural Basis of Altruistic Punishment

Dominique J.-F. de Quervain,[1]*† Urs Fischbacher,[2]* Valerie Treyer,[3] Melanie Schellhammer,[2] Ulrich Schnyder,[4] Alfred Buck,[3] Ernst Fehr[2,5]†

Many people voluntarily incur costs to punish violations of social norms. Evolutionary models and empirical evidence indicate that such altruistic punishment has been a decisive force in the evolution of human cooperation. We used $H_2{}^{15}O$ positron emission tomography to examine the neural basis for altruistic punishment of defectors in an economic exchange. Subjects could punish defection either symbolically or effectively. Symbolic punishment did not reduce the defector's economic payoff, whereas effective punishment did reduce the payoff. We scanned the subjects' brains while they learned about the defector's abuse of trust and determined the punishment. Effective punishment, as compared with symbolic punishment, activated the dorsal striatum, which has been implicated in the processing of rewards that accrue as a result of goal-directed actions. Moreover, subjects with stronger activations in the dorsal striatum were willing to incur greater costs in order to punish. Our findings support the hypothesis that people derive satisfaction from punishing norm violations and that the activation in the dorsal striatum reflects the anticipated satisfaction from punishing defectors.

The nature and level of cooperation in human societies is unmatched in the animal world. Humans cooperate with genetically unrelated strangers, often in large groups, with people they will never meet again, and when reputation gains are absent. Recent research indicates that strong reciprocity—the combination of altruistic punishment and altruistic rewarding—has been crucial in the evolution of human cooperation (*1*–*3*). People often reward others for cooperative, norm-abiding behaviors, and they punish violations of social norms (*4*, *5*). For thousands of years, human societies did not have the modern institutions of law enforcement—impartial police and impartial judges that ensure the punishment of norm violations such as cheating in an economic exchange, for example. Thus, social norms had to be enforced by other measures, and private sanctions were one of these means. Many norms are still enforced by private sanctions, even in contemporary Western societies. Such sanctions are altruistic if they involve costly acts that confer economic benefits on other individuals. If, for example, an individual sanctions a person who cheated in an economic exchange, the cheater's future interaction partners will benefit from this punishment because the cheater is now more aware that cheating will be punished. This knowledge is likely to deter future cheating (*3*).

Why do people punish violators of widely approved norms although they reap no offsetting material benefits themselves? We hypothesize that individuals derive satisfaction from the punishment of norm violators. Several sources suggest this hypothesis. First, recent models of social preferences (*6*–*8*) define utility functions that incorporate a motive to sanction violations of fairness and cooperation norms. These models predict actual behavior better than do models based on self-interested preferences, lending support to the idea that people are motivated to punish norm violations. Second, recent models of the evolution of human cooperation (*1*, *2*) indicate that altruistic punishment has deep evolutionary roots. This suggests that proximate mechanisms evolved that induce humans to bear the cost of punishing others. Because altruistic punishment is not an automatic response, such as the digestion of food, but rather is an action based on deliberation and intent, humans have to be motivated to punish. The typical proximate mechanism for inducing motivated action is that people derive satisfaction from the action. Most people seem to feel bad if they observe that norm violations are not punished, and they seem to feel relief and satisfaction if justice is established. Many languages even have proverbs indicating such feelings, for example, "Revenge is sweet."

A design to study the punishment of defectors. We examined the hypothesis that people derive satisfaction from the punishment of norm violations by combining an economic experiment involving real monetary payoffs with positron emission tomography (PET). Our hypothesis predicts that altruistic punishment is associated with the activation of brain areas related to reward processing. Single-neuron recording in nonhuman primates (*9*–*11*) and neuroimaging studies with humans using money as a reward medium (*12*–*16*) reliably indicate that the striatum is a key part of reward-related neural circuits. Moreover, if altruistic punishment occurs because the punisher anticipates deriving satisfaction from punishing, we should observe activation predominantly in those reward-related brain areas that are associated with goal-directed behavior. Single-neuron recording in nonhuman primates (*17*–*19*) provides strong evidence that the dorsal striatum is crucial for the integration of reward information and behavioral information in the sense of a goal-directed mechanism. A recent neuroimaging study also supports the view that the dorsal striatum is implicated in the processing of rewards that accrue as a result of a decision (*20*).

In our experiment, two human players, A and B, interact anonymously with each other (*21*). Both players know that they face a human player, and each of them is endowed with 10 money units (MUs). They can increase their income substantially if player A trusts B, and B acts in a trustworthy manner. More specifically, A makes the first decision. He can send his endowment of 10 MUs to B (case 1) or he can keep his endowment (case 2). If A trusts B and sends his endowment (case 1), the experimenter quadruples the amount sent so that B receives 40 MUs. At that moment, B has 50 MUs in total—his endowment plus the 40 units just received—and A has nothing. Then B has the choice of sending back nothing or half of the 50 MUs. Thus, if B acts trustworthily and sends back half, both players earn 25 MUs, but if B keeps all the money, he earns 50 MUs and A, who trusted B, earns nothing. In case 2, that is, if A does not trust B, both players keep their endowment of 10 MUs (*22*).

We hypothesized that if A trusts B, cooperation and fairness norms dictate that player B send back half the money. Therefore, if B is untrustworthy and keeps all the money, A interprets this as a norm violation, which we predict will evoke a desire to punish B. For

[1]Division of Psychiatry Research, University of Zurich, Lenggstrasse 31, 8029 Zurich, Switzerland. [2]Institute for Empirical Research in Economics, University of Zurich, Blümlisalpstrasse 10, 8006 Zurich, Switzerland. [3]PET Center, Nuclear Medicine, Department of Radiology, University Hospital, 8091 Zurich, Switzerland. [4]Psychiatric Department, University Hospital, Culmannstrasse 8, 8091 Zurich, Switzerland. [5]Collegium Helveticum, Schmelzbergstrasse 25, 8092 Zürich, Switzerland.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: quervain@bli.unizh.ch (D.D.); efehr@iew.unizh.ch (E.F.)

this reason, A receives the option of punishing B by assigning up to 20 punishment points to player B (*23*). After player A is informed about B's action, A has 1 min to deliberate and decide whether he wants to punish B and, if so, how many punishment points to assign. The experimenter asks player A for his decision at the end of the minute. Because we are interested in the neural basis of punishment, we scanned A's brain during this 1-min period. In total, player A was sequentially matched with seven different subjects in the role of B, that is, A played the experiment described above seven times. Because both players can earn considerably more money if A trusts B, and B is trustworthy, A has a strong incentive to trust B; in fact, all but one subject in the role of A trusted B in all seven trials. Player A faced a trustworthy opponent in three of the seven trials, but B kept all the money in the remaining four trials. Because we are interested in imaging altruistic punishment, and to keep radioactivity as low as possible, we scanned those trials in which B kept all the money, because A is expected to have a desire to punish B only in those trials. During a 10-min break between trials, A answered questionnaires in which he assessed the fairness of B's action in the previous trial and his desire to punish B on a seven-point Likert scale. Fifteen healthy, right-handed male subjects participated in the role of player A in our experiment. Because we are interested in A's response to the abuse of trust, our analysis includes the 14 subjects who trusted B.

Predicted brain activations across treatments. Player A experienced four different treatment conditions in the four trials in which B kept all the money. These conditions generate the contrasts necessary to measure the activation of reward-related brain areas during the punishment period. In the condition termed "intentional and costly" (IC), B himself decides whether to keep all the money or to send back money. Thus, if B keeps all the money, he intentionally abuses A's trust. In addition, the punishment is costly for both A and B. Every punishment point assigned to B costs one MU for A and reduces B's payoff by two MUs. In the condition termed "intentional and free" (IF), B also decides about the transfer himself, but the punishment is not costly for A. Every punishment point assigned to B costs nothing for A, whereas B's payoff is reduced by two MUs. In a third condition, which we call "intentional and symbolic" (IS), B again makes the decision, but punishment has only a symbolic meaning. Every punishment point assigned to B costs neither A nor B anything. Thus, A cannot reduce the payoff to B in this condition. Finally, there is a condition called "nonintentional and costly" (NC) in which a random device determines B's decision, removing the responsibility for it from player B. Punishment is again costly for both A and B; A loses one MU and B loses two MUs per punishment point assigned to B (*23*). To control for sequence effects, the sequence of treatment condition was randomly determined.

These conditions enable us to test our hypothesis by computing the differences in brain activation across relevant conditions. We predict, in particular, that the contrast IF-IS activates reward-related brain areas after A's trust has been abused. We predict that A has a desire to punish B both in the IF and the IS conditions because B intentionally abused A's trust, but A cannot really hurt B in the IS condition. Thus, the purely symbolic punishment in the IS condition is unlikely to be satisfactory because the desire to punish the defector cannot be fulfilled effectively, and in the unlikely case that symbolic punishment is satisfactory, we predict that it is less so than punishment in the IF condition.

The satisfaction from punishing effectively may have various psychological sources. Subjects who do not punish may feel bad because the defector gets away unpunished and has a much higher payoff than they themselves have; in this case, effective punishment prevents a negatively reinforcing outcome. Alternatively, effective punishment may be perceived as just and subjects may feel good about this; in that case, punishment is associated with a positively reinforcing outcome.

The IF-IS contrast is ideal for examining the satisfying aspects of effective punishment because, except for the difference in the opportunity to punish effectively, everything else is kept constant across conditions. If punishment is indeed satisfactory in the IF condition, we expect that subjects are also willing to incur cost to punish the defector. In fact, those subjects who show the strongest activation of reward-related areas in the IF condition should also be those who incur the largest cost of punishing in the IC condition. Moreover, if subjects reasonably weigh the costs and the satisfaction of punishing B, that is, if they punish as long as the marginal costs are below the marginal "benefits" of punishing, punishment in the IC condition should also be experienced as satisfactory. Thus, we predict that reward-related areas will also be activated in the IC-IS condition.

If B keeps all the money in the NC condition, he is not responsible for this action because a random device forced him to do so. We therefore predict that A does not view B's act as unfair and has no desire, or a strongly reduced desire, to punish B. If there is no desire to punish, punishment is unlikely to yield satisfaction. For this reason, we predict activations in reward-related areas in the IF-NC and the IC-NC contrasts. Finally, we can also compute the combined contrast (IC + IF) – (IS + NC). We should also observe activations in reward-related areas in this contrast, because there is the desire and the opportunity to punish both in the IC and the IF conditions, whereas there is no opportunity for punishment in the IS condition and there is no desire to do so in the NC condition. If either the opportunity or the desire to

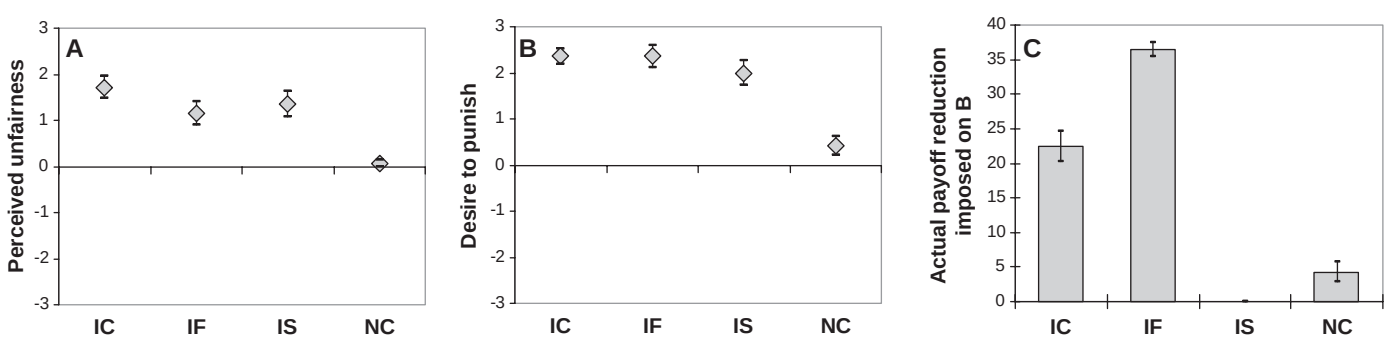


Fig. 1. Player A's feelings about player B and actual payoff reduction imposed on B. (**A**) Player A's perceived unfairness if B kept all the money. During the 10-min interval between PET scans, player A indicated on a seven-point Likert scale (from –3 to +3) whether he perceived B's action in the previous trial as fair or unfair. Maximal fairness is indicated by –3, maximal unfairness by +3. The figure shows the mean perception across subjects ± SE. (**B**) Player A's desire to punish B if the latter kept all the money. During the 10-min interval between PET scans, A indicated on a seven-point Likert scale (from –3 to +3) the strength of his desire to reward or to punish B. The maximal desire to reward is indicated by –3, the maximal desire to punish by +3. We show the mean desire to reward/punish ± SE. (**C**) Actual payoff reduction imposed on B if the latter keeps all the money. The figure shows the mean payoff reduction A imposed on B ± SE. In the IS condition, the economic payoff of B could not be reduced.

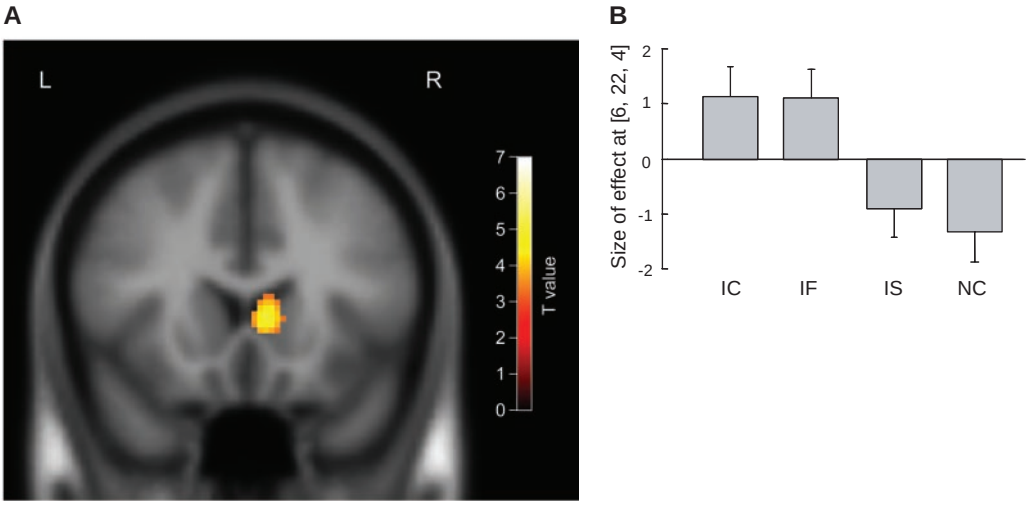


Fig. 2. (**A**) Activation in the caudate nucleus in conditions in which subjects indicated a strong desire to punish and could effectively do so (IC and IF) relative to conditions in which there is no effective punishment or the desire to punish is absent (IS and NC). (**B**) Effect sizes at the peak of blood-flow increase in the caudate nucleus. Bars indicate caudate activity in each condition relative to the mean brain activation ± SD.

punish effectively is absent, punishment can yield little or no satisfaction.

Questionnaire and behavioral results support these hypotheses (Fig. 1, A to C). Player A views B's act to keep all the money as very unfair in all three intentional conditions (IC, IF, and IS), whereas he views it as nearly neutral in the NC condition (Fig. 1A; sign test for equality of medians, $P < 0.002$ for all pair-wise comparisons of the NC condition with each intentional condition). Likewise, player A exhibits a strong desire to punish B in all three intentional conditions, but this desire is nearly absent in the NC condition (Fig. 1B; sign test for equality of medians, $P < 0.012$ for all pair-wise comparisons of the NC condition with each intentional condition). Moreover, player A imposes much higher payoff reductions on B in those conditions in which B intentionally abuses his trust, whereas almost no punishment is imposed on B in the NC condition (Fig. 1C; $P \leq 0.001$ for sign test comparing IC and NC and for the test comparing IF and NC). Twelve of 14 subjects punished B if he kept all the money in the IC condition, and all 14 subjects punished B in the IF condition. This contrasts with the NC condition, in which only 3 of 14 subjects reduced B's payoff, and those who did so punished only a little.

Does punishment activate reward-related brain circuits? Among the areas showing greater activation in the contrasts described above is the caudate nucleus (Table 1), which is activated in all five contrasts in which we predicted the activation of reward-related areas. For example, the peak activation in the contrast (IC + IF) – (IS + NC) is observed at the coordinates (6, 22, 4), the head of the caudate nucleus (Fig. 2A; $P < 0.05$, corrected for multiple comparisons). Moreover, effect-size analysis at the peak of the blood flow in the caudate (Fig. 2B) indicates that the different conditions contributed to this activation in the predicted way: We observe above-average activations in the IC and IF conditions, in which subjects express a strong desire to punish and can satisfy this desire; we observe below-average activations in the IS and NC conditions, in which subjects either cannot satisfy their desire to punish or feel no desire to punish. This pattern of caudate activation is also replicated in the individual contrasts IF-IS, IC-IS, IF-NC, and IC-NC (Table 1).

Table 1. PET results.

Contrast	Region (BA)	Coordinates			*Z* value
		x	*y*	*z*	
(IC + IF) – (IS + NC)	Caudate nucleus	6	22	4	5.11*
	Thalamus	22	–24	10	4.43*
IF-IS	Caudate nucleus	6	22	4	3.55
	Thalamus	22	–22	10	4.21
IC-IS	Caudate nucleus	6	24	2	3.70
	Thalamus	22	–22	10	4.15
IF-NC	Caudate nucleus	6	22	4	4.18
IC-NC	Caudate nucleus	6	22	4	4.23
IC-IF	Ventromedial prefrontal cortex (BA 10)	2	54	–4	4.59
	Medial orbitofrontal cortex (BA 11)	–4	52	–16	3.35

The table shows MNI coordinates (*x*, *y*, *z*) that locate the maxima of changes in blood flow. * indicates significant activations at the $P < 0.05$ level, corrected for multiple comparisons. Otherwise, the threshold for hypothesized brain regions is $P < 0.001$, uncorrected. For all activations at $P < 0.001$, see (*21*). BA denotes Brodmann area. IC is the intentional and costly condition, IF the intentional and free condition, IS the intentional and symbolic condition, and NC the nonintentional and costly condition. A negative value for the *x* coordinate indicates the left side of the brain. MNI denotes Montreal Neurological Institute.

The activation of the caudate in those conditions in which subjects expressed a strong desire to punish and could indeed punish is particularly interesting in light of this region's prominent role in the processing of rewards. In animals, this brain region has been associated with the processing of reward information by means of lesion experiments with rats (*24*) and single-cell recordings in nonhuman primates (*10*, *19*). Caudate activations in humans have been reported in several neuroimaging studies that investigated reward processing (*12*, *13*, *15*, *16*, *25*, *26*); in addition, caudate activations have been observed with reinforcers such as cocaine (*27*) and nicotine (*28*). Some neuroimaging studies even indicate that parametric increases in monetary rewards are positively correlated with caudate activations (*14*, *15*).

We also found increased blood flow in the thalamus (Table 1) in those conditions in which subjects expressed a strong desire to punish and could punish (IC and IF) relative to the symbolic punishment condition. No thalamus activation was found when IC and IF were compared with the NC condition, in which the desire to punish was absent. Activations in the thalamus have been reported in human neuroimaging studies investigating processing of monetary reward (*14*, *16*, *26*). Taken together, our findings suggest a prominent role of the caudate nucleus, with possible contributions of the thalamus, in processing rewards associated with the satisfaction of the desire to punish the intentional abuse of trust.

This result would be further supported if we were able to show that those subjects with a stronger caudate activation punish more strongly. We examined this question by computing

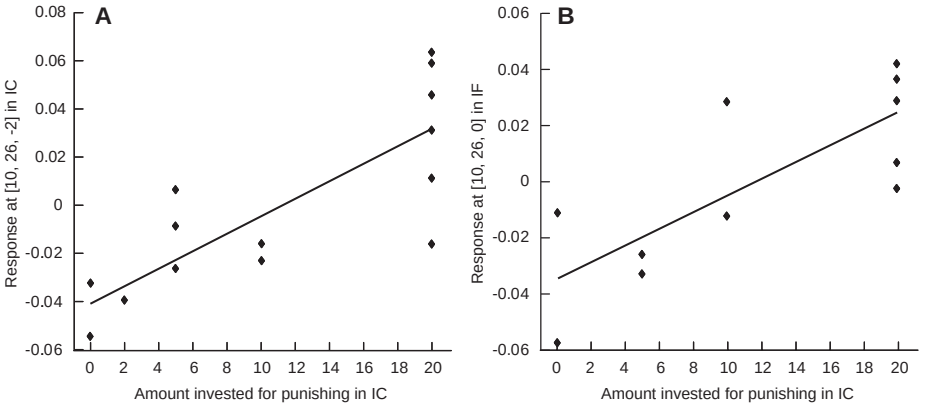
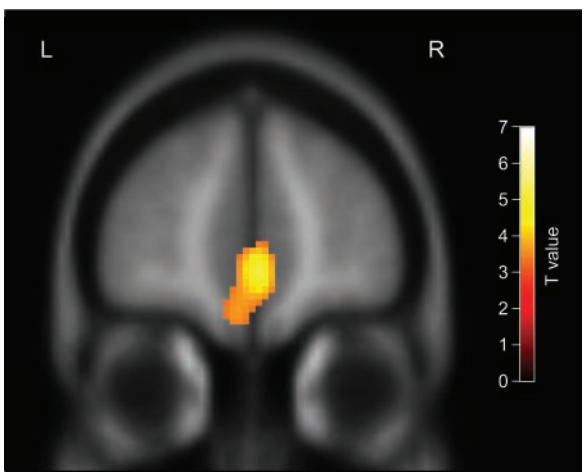


Fig. 3. (**A**) Positive correlation between caudate activation at coordinates [10, 26, –2] and the amount of money spent on punishment in the IC condition. Subjects with higher caudate activation in the IC condition spent more money on punishment in this condition. (**B**) Positive correlation between caudate activation at coordinates [10, 26, 0] in the IF condition in those subjects that punished maximally and the amount of money spent on punishment by these subjects in the IC condition. Subjects with higher caudate activation at the same (maximal) level of punishment in the IF condition spent more money on punishment in the IC condition.

Fig. 4. The role of the prefrontal cortex in integrating the benefits and costs of punishing. Activation of the ventromedial prefrontal cortex and the medial orbitofrontal cortex in the condition where subjects have a strong desire to sanction but where sanctioning is costly for the punisher (IC) relative to the condition where there is also a strong desire to sanction but sanctioning is costless for the punisher (IF).



the correlation between brain activation and the actual monetary punishment across subjects in the IC condition. We did indeed find a positive correlation between caudate activation (at coordinate position [10, 26, –2]; $P < 0.001$) and investments in punishment (Fig. 3A). This correlation can be interpreted in two ways. One interpretation is that a higher punishment could have induced stronger feelings of satisfaction, which suggests that stronger punishment causes stronger caudate activation. Alternatively, subjects who expected higher satisfaction from punishing a defector could have been willing to invest more money in punishment. If the second interpretation is true, the causality is reversed: A higher caudate activation reflects the greater expected satisfaction from punishment, which, in turn, causes higher investments in punishment. The second interpretation is particularly interesting in light of the caudate's oft-noted role in the integration of reward information and behavioral information in the sense of a goal-directed mechanism (*9*).

Brain activations and the decision to punish. Our data enable us to discriminate between the two interpretations. The key is to examine the caudate activations of the 11 subjects in the IF condition who punished maximally. Because these subjects imposed the same punishment on B, differences in their caudate activations in the IF condition cannot be due to differences in punishment. However, if caudate activation reflects the expected satisfaction from a given level of punishment, the differences in caudate activations across these subjects may reflect differences in expected satisfaction from a given level of punishment. If this interpretation is true, we should observe that subjects who exhibit a higher caudate activation in the IF condition, that is, subjects who expect a higher satisfaction from the same level of punishment, are willing to invest more money in punishment if it is costly to punish. In other words, this interpretation predicts that among the subjects who punished maximally in the IF condition, those with higher caudate activation in the IF invest more in punishment in the IC condition. This prediction is supported by a positive correlation between caudate activation in the IF and the amount invested in punishing in the IC condition (Fig. 3B; $P < 0.002$). This finding lends support to the hypothesis that the observed activations in the dorsal striatum reflect expected satisfaction from punishment, which is consistent with the view of the dorsal striatum as a key area involved in goal-directed, rewarding behavior.

If the punishment of intentional defectors is rewarding, player A faces a trade-off in the IC condition but not in the IF condition, because punishment is costly in the former. Player A has to weigh the emotional satisfaction of punishing against the monetary cost of punishing, which requires integration of separate cognitive operations in the pursuit of a behavioral goal. Much evidence indicates that the prefrontal and the orbitofrontal cortex are involved in integrating separate cognitive operations and decision making (*29*–*32*). Our behavioral data suggest that in the IC condition subjects face a decision problem because most subjects punish maximally in the IF condition, whereas the cost for the punisher reduces punishment significantly in the IC condition (Fig. 1C; sign test, $P = 0.039$). Therefore, we expected activations in the pre- and orbitofrontal cortex in the IC-IF contrast. The data show (Table 1 and Fig. 4) that the ventromedial prefrontal (BA 10) and the medial orbitofrontal cortex (BA 11) are activated in this contrast. The activation of BA 10 is interesting because this area has been associated with the integration of two or more separate cognitive operations in the pursuit of higher behavioral goals (*33*). The activation in the medial orbitofrontal cortex is also interesting because of this region's oft-noted involvement in difficult choices that require the coding of reward value (*34*, *35*). These activations also provide indirect support for the hypothesis that punishing defectors involves satisfaction, because if that were not the case, no benefits would have to be weighed against the costs of punishing and no integration would have to take place.

These results also illustrate the stark contrast between the biological and the psychological definitions of altruism (*4*). According to the biological definition, an act is altruistic if it is costly for the actor and confers benefits on other individuals. It is completely irrelevant for this definition whether the act is motivated by the desire to confer benefits on others, because altruism is solely defined in terms of the consequences of behavior. This contrasts with the psychological definition, which also requires that the act be driven by an altruistic motive that is not based on hedonic rewards (*36*). Thus, the punishment of defectors is an altruistic act in the biological sense because, typically, it is costly for the punisher and induces the punished individual to defect less in future interactions with others. However, our results suggest that it is not an altruistic act in the psychological sense.

Conclusions. Our study is part of recent attempts in "neuroeconomics" and the "cognitive neuroscience of social behavior" to understand the social brain and the associated moral emotions (*37–44*). However, this study sought to identify the neural basis of the altruistic punishment of defectors. The ability to develop social norms that apply to large groups of genetically unrelated individuals and to enforce these norms through altruistic sanctions is one of the distinguishing characteristics of the human species. Altruistic punishment is probably a key element in explaining the unprecedented level of cooperation in human societies (*1–3*). We hypothesize that altruistic punishment provides relief or satisfaction to the punisher and activates, therefore, reward-related brain regions. Our design generates five contrasts in which this hypothesis can be tested, and the anterior dorsal striatum is activated in all five contrasts, which suggests that the caudate plays a decisive role in altruistic punishment. Caudate activation is particularly interesting because this brain region has been implicated in making decisions or taking actions that are motivated by anticipated rewards (*17–20*). The prominent role of the caudate in altruistic punishment is further supported by the fact that those subjects who exhibit stronger caudate activation spend more money on punishing defectors. Moreover, our results also shed light on the reasons behind this correlation. Subjects who exhibit higher caudate activation at the maximal level of punishment if punishment is costless for them also spend more resources on punishment if punishment becomes costly. Thus, high caudate activation seems to be responsible for a high willingness to punish, which suggests that caudate activation reflects the anticipated satisfaction from punishing defectors. Our results therefore support recently developed social preference models (*6–8*), which assume that people have a preference for punishing norm violations, and illuminate the proximate mechanism behind evolutionary models of altruistic punishment.

References and Notes

1. R. Boyd, H. Gintis, S. Bowles, P. J. Richerson, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3531 (2003).
2. S. Bowles, H. Gintis, *Theor. Popul. Biol.* **65**, 17 (2004).
3. E. Fehr, S. Gächter, *Nature* **415**, 137 (2002).
4. E. Sober, D. S. Wilson, *Unto Others: The Evolution and Psychology of Unselfish Behavior* (Harvard University Press, Cambridge, MA, 1998).
5. E. Fehr, U. Fischbacher, *Nature* **425**, 785 (2003).
6. M. Rabin, *Am. Econ. Rev.* **83**, 1281 (1993).
7. E. Fehr, K. M. Schmidt, *Q. J. Econ.* **114**, 817 (1999).
8. C. F. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton University Press, Princeton, NJ, 2003).
9. W. Schultz, *Nature Rev. Neurosci.* **1**, 199 (2000).
10. P. Apicella, T. Ljungberg, E. Scarnati, W. Schultz, *Exp. Brain Res.* **85**, 491 (1991).
11. O. Hikosaka, M. Sakamoto, S. Usui, *J. Neurophysiol.* **61**, 814 (1989).
12. M. R. Delgado, V. A. Stenger, J. A. Fiez, *Cereb. Cortex* **14**, 1022 (2004).
13. B. Knutson, A. Westdorp, E. Kaiser, D. Hommer, *Neuroimage* **12**, 20 (2000).
14. C. Martin-Soelch, J. Missimer, K. L. Leenders, W. Schultz, *Eur. J. Neurosci.* **18**, 680 (2003).
15. M. R. Delgado, H. M. Locke, V. A. Stenger, J. A. Fiez, *Cognit. Affect. Behav. Neurosci.* **3**, 27 (2003).
16. B. Knutson, C. M. Adams, G. W. Fong, D. Hommer, *J. Neurosci.* **21**, RC159 (2001).
17. W. Schultz, R. Romo, *Exp. Brain Res.* **71**, 431 (1988).
18. R. Kawagoe, Y. Takikawa, O. Hikosaka, *Nature Neurosci.* **1**, 411 (1998).
19. J. R. Hollerman, L. Tremblay, W. Schultz, *J. Neurophysiol.* **80**, 947 (1998).
20. J. O'Doherty *et al.*, *Science* **304**, 452 (2004).
21. Materials and methods are available as supporting material on *Science* Online.
22. To maintain symmetry with case 1, player B can also give half of his money to A if A does not trust him. However, because all subjects (except one) in the role of A trusted B, this contingency almost never occurred.
23. In all conditions, both players received an additional endowment of 20 MUs after player B made his decision. This endowment allowed A to finance the cost of punishment in those conditions in which punishment was also costly for him. If A did not punish, both players kept the 20 MUs. If punishment was not costly for A, he kept the 20 MUs regardless of the number of assigned punishment points.
24. J. A. Salinas, N. M. White, *Behav. Neurosci.* **112**, 812 (1998).
25. M. J. Koepp *et al.*, *Nature* **393**, 266 (1998).
26. M. R. Delgado, L. E. Nystrom, C. Fissell, D. C. Noll, J. A. Fiez, *J. Neurophysiol.* **84**, 3072 (2000).
27. H. C. Breiter *et al.*, *Neuron* **19**, 591 (1997).
28. E. A. Stein *et al.*, *Am. J. Psychiatr.* **155**, 1009 (1998).
29. E. K. Miller, J. D. Cohen, *Annu. Rev. Neurosci.* **24**, 167 (2001).
30. A. D. Wagner, A. Maril, R. A. Bjork, D. L. Schacter, *Neuroimage* **14**, 1337 (2001).
31. D. C. Krawczyk, *Neurosci. Biobehav. Rev.* **26**, 631 (2002).
32. A. Bechara, H. Damasio, A. R. Damasio, *Cereb. Cortex* **10**, 295 (2000).
33. N. Ramnani, A. M. Owen, *Nature Rev. Neurosci.* **5**, 184 (2004).
34. R. Elliott, J. L. Newman, O. A. Longe, J. F. Deakin, *J. Neurosci.* **23**, 303 (2003).
35. F. S. Arana *et al.*, *J. Neurosci.* **23**, 9632 (2003).
36. C. D. Batson, J. Fultz, A. Schoenrade, A. Paduano, *J. Pers. Soc. Psychol.* **53**, 594 (1987).
37. R. Adolphs, *Curr. Opin. Neurobiol.* **11**, 231 (2001).
38. J. D. Greene, R. B. Sommerville, L. E. Nystrom, J. M. Darley, J. D. Cohen, *Science* **293**, 2105 (2001).
39. K. McCabe, D. Houser, L. Ryan, V. Smith, T. Trouard, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 11832 (2001).
40. J. K. Rilling *et al.*, *Neuron* **35**, 395 (2002).
41. T. Singer *et al.*, *Neuron* **41**, 653 (2004).
42. J. Moll *et al.*, *J. Neurosci.* **22**, 2730 (2002).
43. A. G. Sanfey, J. K. Rilling, J. A. Aronson, L. E. Nystrom, J. D. Cohen, *Science* **300**, 1755 (2003).
44. R. Adolphs, *Nature Rev. Neurosci.* **4**, 165 (2003).

45. We gratefully acknowledge support by the University of Zurich, the Swiss National Science Foundation, and the MacArthur Foundation Network on Economic Environments and the Evolution of Individual Preferences and Social Norms. We thank R. Adolphs, T. Singer, and L. Jäncke for helpful comments on earlier drafts of this paper.

Supporting Online Material

www.sciencemag.org/cgi/content/full/305/5688/1254/DC1
Materials and Methods
Table S1
References

26 May 2004; accepted 20 July 2004

Spatial Representation in the Entorhinal Cortex

Marianne Fyhn,[1] Sturla Molden,[1] Menno P. Witter,[1,2] Edvard I. Moser,[1]* May-Britt Moser[1]

As the interface between hippocampus and neocortex, the entorhinal cortex is likely to play a pivotal role in memory. To determine how information is represented in this area, we measured spatial modulation of neural activity in layers of medial entorhinal cortex projecting to the hippocampus. Close to the postrhinal-entorhinal border, entorhinal neurons had stable and discrete multipeaked place fields, predicting the rat's location as accurately as place cells in the hippocampus. Precise positional modulation was not observed more ventromedially in the entorhinal cortex or upstream in the postrhinal cortex, suggesting that sensory input is transformed into durable allocentric spatial representations internally in the dorsocaudal medial entorhinal cortex.

An extensive body of evidence suggests that the hippocampus is essential for fast encoding and storage of new episodic memories but has a more limited role in remote memory, which is thought to be stored primarily in the neocortex (*1–4*). Memory consolidation in the neocortex appears to be a slow and gradual process based on repeated interactions with the hippocampus (*2*, *3*). These interactions must be mediated largely through the entorhinal cortex, which interconnects the hippocampus with nearly all other association cortices (*5–8*). Understanding how information is processed in the entorhinal cortex is thus essential to resolving the interaction between the hippocampus and neocortex during encoding, consolidation, storage, and retrieval of memory.

However, little is known about how sensory input is represented in the entorhinal cortex. Although hippocampal memories are expressed at the neuronal level as representations with evident correlates to the spatial and nonspatial structure of the external environment (*6*, *9*, *10*), the functional correlates of entorhinal neurons

[1]Centre for the Biology of Memory, Medical-Technical Research Centre, Norwegian University of Science and Technology, 7489 Trondheim, Norway. [2]Research Institute Neurosciences, Department of Anatomy, VU University Medical Center, Amsterdam, Netherlands.

*To whom correspondence should be addressed. E-mail: edvard.moser@cbm.ntnu.no