

The evolutionary significance of costly punishment is still to be demonstrated

Costly punishment represents an evolutionary puzzle because it involves an individual paying a cost to harm another individual (1, 2). A recent study by Wu et al. (3) confirmed earlier conclusions (4) that costly punishment is mostly maladaptive and may have evolved for other reasons. We welcome their emphasis on cultural differences in experimental games (5), but we argue that these first tests (3, 4) are not yet conclusive with regard to the functional significance of costly punishment. There are three problems: (i) the definition of “defection” in their Prisoner’s Dilemma is problematic for studies on punishment, (ii) a closer examination of their payoff matrices reveals them to be nonconducive to punishment, and (iii) their specific game structure imposes an additional opportunity cost that further reduces the effectiveness of punishment.

In Wu et al.’s experiment (3), individuals play a repeated game in pairs and can choose between three options: “cooperate” (C; giving their partner a benefit $+b$ at a personal cost $-c$), “defect” (D; inflicting a cost of $-d$ on their partner while gaining $+d$ for themselves), and “punish” (P; paying a cost of $-\alpha$ to inflict a greater cost of $-\beta$ on their partner). The corresponding payoff matrix is:

$$\text{Actor} \begin{matrix} & \text{Recipient} \\ & \begin{matrix} \text{C} & \text{D} & \text{P} \end{matrix} \\ \begin{matrix} \text{C} \\ \text{D} \\ \text{P} \end{matrix} & \begin{pmatrix} b-c & -c-d & -c-\beta \\ b+d & d-d & -\beta+d \\ b-\alpha & -\alpha-d & -\alpha-\beta \end{pmatrix} \end{matrix}$$

In their experiments, $\alpha = 1$, $\beta = 4$, $d = 1$, $c = 1$, and b is either 2 or 3, which correspond to the following two payoff matrices:

$$\text{Actor} \begin{matrix} & \text{Recipient} \\ & \begin{matrix} \text{C} & \text{D} & \text{P} \end{matrix} \\ \begin{matrix} \text{C} \\ \text{D} \\ \text{P} \end{matrix} & \begin{pmatrix} 1 & -2 & -5 \\ 3 & 0 & -3 \\ 1 & -2 & -5 \end{pmatrix} \end{matrix}$$

and

$$\text{Actor} \begin{matrix} & \text{Recipient} \\ & \begin{matrix} \text{C} & \text{D} & \text{P} \end{matrix} \\ \begin{matrix} \text{C} \\ \text{D} \\ \text{P} \end{matrix} & \begin{pmatrix} 2 & -2 & -5 \\ 4 & 0 & -3 \\ 2 & -2 & -5 \end{pmatrix} \end{matrix}$$

The most parsimonious definition of defection is $d = 0$. Defection is then simply the absence of cooperation. With $d > 0$, however, a new element is introduced, where “defection” then means taking something ($-d$) from the other person.

This is problematic in studies on punishment because it is now unclear whether punishment in the experimental game is a reaction to the absence of cooperation or to the immediate loss that defection has inflicted.

If the participants of the experimental game did not bother about the immediate loss inflicted by defection but simply interpreted defection as the absence of cooperation, the payoff matrix simplifies to:

$$\text{Actor} \begin{matrix} & \text{Recipient} \\ & \begin{matrix} \text{C} & \text{D} & \text{P} \end{matrix} \\ \begin{matrix} \text{C} \\ \text{D} \\ \text{P} \end{matrix} & \begin{pmatrix} b-c & -c & -c-\beta \\ b & 0 & -\beta \\ b-\alpha & -\alpha & -\alpha-\beta \end{pmatrix} \end{matrix}$$

We can rederive the original payoff matrix of Wu et al. (3) by setting $c = 2$, $\alpha = 2$ and $\beta = 3$ in both of their matrices, and $b = 3$ in the first and $b = 4$ in the second. The efficiency ratio of punishment is then $\alpha:\beta = 2:3$ which is far less conducive to punishment than the $\alpha:\beta = 1:4$ punishment ratio assumed by the authors (3, 4).

A further problem with the experimental set-up is that players can only choose between cooperating, defecting, or punishing. A “defection” can then only be punished in the next round of simultaneous choices, i.e., choosing to punish results in the individual forfeiting a potential act of cooperation or “defection.” Punishment seems more likely to promote cooperation if punishing, or not punishing, is a separate decision between two rounds of cooperating or not cooperating.

In conclusion, the game settings that have been used in recent studies (3, 4) on the evolutionary significance of costly punishment are problematic. It can therefore not yet be concluded that costly punishment is unlikely to evolve to increase cooperation.

Daniel J. Rankin^{a,b,1}, Miguel dos Santos^{a,c}, and Claus Wedekind^c
^aDepartment of Biochemistry, University of Zürich, Building Y27, Winterthurststrasse 190, CH-8057 Zürich, Switzerland; ^bSwiss Institute of Bioinformatics, Quartier Sorge Bâtiment Génopode, CH-1015 Lausanne, Switzerland; and ^cDepartment of Ecology and Evolution, Biophore, University of Lausanne, CH-1015 Lausanne, Switzerland

1. Sigmund K, Hauert C, Nowak MA (2001) Reward and punishment. *Proc Natl Acad Sci USA* 98:10757–10762.
2. Herrmann B, Thöni C, Gächter S (2008) Antisocial punishment across societies. *Science* 319:1362–1366.
3. Wu J-J, et al. (2009) Costly punishment does not always increase cooperation. *Proc Natl Acad Sci USA* 106:17448–17451.
4. Dreber A, Rand DG, Fudenberg D, Nowak MA (2008) Winners don’t punish. *Nature* 452:348–351.
5. Henrich J, et al. (2006) Costly punishment across human societies. *Science* 312:1767–1770.

Author contributions: D.J.R., M.d.S., and C.W. wrote the paper.

The authors declare no conflict of interest.

¹To whom correspondence should be addressed at: Department of Biochemistry, University of Zürich, Building Y27, Winterthurststrasse 190, CH-8057 Zürich, Switzerland. E-mail: d.rankin@bioc.uzh.ch.