S Supporting Information

S.1 Stability condition for behavioral equilibrium

In order for a candidate behavioral equilibrium, a_1^* and a_2^*, that solves the equations in (5) to represent a local maximum for each individual's objective function and to be locally stable under the behavioral dynamics given in (4), the candidate equilibrium must satisfy two conditions. The first guarantees that a_1^* and a_2^* represent local maxima and is given by two equations:

$$\begin{aligned} \left.\frac{\partial^2 x_1}{\partial a_1^2}\right|_{\substack{a_1=a_1^* \\ a_2=a_2^*}} &< 0 \\ \left.\frac{\partial^2 x_2}{\partial a_2^2}\right|_{\substack{a_1=a_1^* \\ a_2=a_2^*}} &< 0 \,. \end{aligned} \tag{S1}$$

The second condition is found by analyzing the conditions for the local stability of a rest point for the gradient dynamics given in (4). The local stability of this rest point is determined by the eigenvalues, λ_1 and λ_2 of the Jacobian matrix defined by

$$J = \begin{pmatrix} \frac{\partial^2 x_1}{\partial a_1^2} & \frac{\partial^2 x_1}{\partial a_1 \partial a_2} \\ \frac{\partial^2 x_2}{\partial a_1 \partial a_2} & \frac{\partial^2 x_2}{\partial a_2^2} \end{pmatrix} \tag{S2}$$

when J is evaluated at a_1^* and a_2^*. A sufficient condition for the stability of this rest point is that the real parts of both λ_1 and λ_2 are less than zero [1]; this condition holds when

$$\mathrm{Tr}(J) = \frac{\partial^2 x_1}{\partial a_1^2} + \frac{\partial^2 x_2}{\partial a_2^2} < 0$$

and

$$|J| = \frac{\partial^2 x_1}{\partial a_1^2}\frac{\partial^2 x_2}{\partial a_2^2} - \frac{\partial^2 x_1}{\partial a_1 \partial a_2}\frac{\partial^2 x_2}{\partial a_1 \partial a_2} > 0 \,, \tag{S3}$$

where $\mathrm{Tr}(J)$ and $|J|$ are evaluated at a_1^* and a_2^*. The equations in condition (S1) guarantee that $\mathrm{Tr}(J) < 0$, so our second condition is equation (S3).

Moreover, we can specify more stringent conditions that guarantee the uniqueness and global stability of the behavioral equilibrium defined by (5). Let $\mathcal{A}$ be the set of allowable actions, i.e. $(a_1, a_2) \in \mathcal{A} \times \mathcal{A}$ where $\times$ denotes the Cartesian product. Suppose that $\mathcal{A}$ is convex. Then if the local stability conditions (S1) and (S3) hold not only at the equilibrium point but also hold for all $(a_1, a_2) \in \mathcal{A} \times \mathcal{A}$, the behavioral equilibrium defined by (5) is unique and globally stable under the dynamic given in (4). A proof of this result would follow Rosen [2] noting that Rosen's result guarantees that the equilibrium is unique and globally stable when $J + J^{\mathsf{T}}$ is negative definite, where $^{\mathsf{T}}$ denotes matrix transpose; J is stable (i.e. all eigenvalues of J have negative real parts) if and only if $J + J^{\mathsf{T}}$ is negative definite [3, p. 160]

S.2 Derivation of the first and second order ESS conditions

First, we establish that one can express the behavioral equilibrium actions a_1^* and a_2^* as functions of β_1 and β_2. Remember that the behavioral equilibrium is given by the equations in (5), which say that the derivative of both individuals' objective functions with respect to their own actions vanish at the behavioral equilibrium. Note also that the objective function of an individual is a function of the payoffs, and consequently, of actions, but also depend on the genetically determined trait β of each individual. Thus, we can take the functions $x_1(a_1, a_2)$ and $x_2(a_1, a_2)$ as instances of a family of functions parametrized by β, $X(a_1, a_2, \beta)$. Moreover, since both the payoff and the objective functions are symmetric with respect to the two individuals, we can write, $x_1(a_1, a_2) = X(a_1, a_2, \beta_1)$ and $x_2(a_1, a_2) = X(a_2, a_1, \beta_2)$. In this way, we can express the behavioral equilibrium conditions as:

$$\left.\frac{\partial x_1}{\partial a_1}\right|_{\substack{a_1=a_1^* \\ a_2=a_2^*}} = \left.\frac{\partial X(a_1, a_2, \beta_1)}{\partial a_1}\right|_{\substack{a_1=a_1^* \\ a_2=a_2^*}} = 0 \tag{S4}$$

$$\left.\frac{\partial x_2}{\partial a_2}\right|_{\substack{a_1=a_1^* \\ a_2=a_2^*}} = \left.\frac{\partial X(a_2, a_1, \beta_2)}{\partial a_2}\right|_{\substack{a_1=a_1^* \\ a_2=a_2^*}} = 0\,. \tag{S5}$$

Thus, we have two equations with four variables (a_1^*, a_2^*, β_1 and β_2) and we can use the implicit function theorem to express two of them as functions of the other two. Specifically, the implicit function theorem says that, for J in (S2), as long as $|J| \neq 0$, we can solve for a_1^* and a_2^* in terms of β_1 and β_2. Furthermore, the first derivatives of a_1^* and a_2^* with respect to β_1 and β_2 exist. Now, we take the total derivatives of equations (5) with respect to β_1 at the behavioral equilibrium:

$$\begin{aligned}
\frac{d}{d\beta_1}\left[\frac{\partial x_1}{\partial a_1}\right] &= \frac{\partial^2 x_1}{\partial a_1^2}\frac{\partial a_1^*}{\partial \beta_1} + \frac{\partial^2 x_1}{\partial a_1 \partial a_2}\frac{\partial a_2^*}{\partial \beta_1} + \frac{\partial^2 x_1}{\partial \beta_1 \partial a_1} = 0\\
\frac{d}{d\beta_1}\left[\frac{\partial x_2}{\partial a_2}\right] &= \frac{\partial^2 x_2}{\partial a_1 \partial a_2}\frac{\partial a_1^*}{\partial \beta_1} + \frac{\partial^2 x_2}{\partial a_2^2}\frac{\partial a_2^*}{\partial \beta_1} + \frac{\partial^2 x_2}{\partial \beta_1 \partial a_2} = 0\,,
\end{aligned}$$

In these equations, all partial derivatives of x_1 and x_2 are evaluated at the behavioral equilibrium (a_1^*, a_2^*).

From these equations, one can solve for $\partial a_1^*/\partial \beta_1$ and $\partial a_2^*/\partial \beta_1$ to obtain

$$\frac{\partial a_1^*}{\partial \beta_1} = -\left[\frac{1}{|J|}\frac{\partial^2 x_2}{\partial^2 a_2}\frac{\partial^2 x_1}{\partial \beta_1 \partial a_1}\right]_{\substack{a_1=a_1^*\\ a_2=a_2^*}} \tag{S6}$$

$$\frac{\partial a_2^*}{\partial \beta_1} = \left[\frac{1}{|J|}\frac{\partial^2 x_2}{\partial a_1 \partial a_2}\frac{\partial^2 x_1}{\partial \beta_1 \partial a_1}\right]_{\substack{a_1=a_1^*\\ a_2=a_2^*}}\,, \tag{S7}$$

which uses the fact that $\frac{\partial^2 x_2}{\partial \beta_1 \partial a_2} = 0$. These derivatives give us the necessary ingredients to write down the change in an individuals fitness when that individual's β is increased or decreased. As explained in the text, in order for β^* to be an ESS, it must be the case that no mutant individual with a value of $\beta \neq \beta^*$ can obtain a higher fitness than a resident individual with the evolutionarily stable value β^* when the population is nearly fixed for β^*. To write the ESS condition, we adopt the convention that individual 1 is always the mutant individual, and individual 2 is the resident, i.e. $\beta_1 = \beta_m$ and $\beta_2 = \beta_r$. Substituting the derivatives (S6) and (S7), we can write for the ESS condition (equation 7):

$$\frac{dw_m}{d\beta_m} = \frac{\partial u_1}{\partial a_1}\frac{-1}{|J|}\frac{\partial^2 x_2}{\partial a_2^2}\frac{\partial^2 x_1}{\partial \beta_1 \partial a_1} + \frac{\partial u_1}{\partial a_2}\frac{1}{|J|}\frac{\partial^2 x_2}{\partial a_1 \partial a_2}\frac{\partial^2 x_1}{\partial \beta_1 \partial a_1} = 0\,.$$

Rearranging, we get:

$$\frac{dw_m}{d\beta_m} = \frac{1}{|J|}\frac{\partial^2 x_1}{\partial \beta_1 \partial a_1}\left(\frac{\partial u_1}{\partial a_2}\frac{\partial^2 x_2}{\partial a_1 \partial a_2} - \frac{\partial u_1}{\partial a_1}\frac{\partial^2 x_2}{\partial a_2^2}\right) = 0\,. \tag{S8}$$

Again, all partial derivatives with respect to a_1 and a_2 are evaluated at the behavioral equilibrium. For this equation to hold, the term in the parentheses has to be zero since $\frac{\partial^2 x_1}{\partial \beta_1 \partial a_1} \neq 0$, so we have the condition in equation (8).

The second order condition for the candidate ESS is that the w_m is a fitness maximum, and not a minimum, at $\beta_m = \beta_r = \beta^*$. Using the implicit function theorem again and the first order condition (8), this second order condition can be shown to be (after considerably more algebra):

$$\frac{d^2 w_m}{d\beta_m^2} = \left(\frac{\frac{\partial^2 x_1}{\partial \beta_1 \partial a_1}}{|J|} \right)^2$$
$$\left[\frac{\partial^2 x_2}{\partial a_2^2}, -\frac{\partial^2 x_2}{\partial a_1 \partial a_2} \right] \cdot \left[H(u_1) + \frac{\left(\frac{\partial u_1}{\partial a_1} \frac{\partial^2 x_1}{\partial a_2 \partial a_1} - \frac{\partial u_1}{\partial a_2} \frac{\partial^2 x_1}{\partial a_1^2} \right)}{|J|} H(\frac{\partial x_2}{\partial a_2}) \right] \cdot \left[\frac{\partial^2 x_2}{\partial a_2^2}, -\frac{\partial^2 x_2}{\partial a_1 \partial a_2} \right]^{\mathsf{T}} < 0\,, \quad \text{(S9)}$$

where $H(\cdot)$ is a Hessian matrix with respect to a_1 and a_2 of the function that is its argument. Likewise, the convergence stability condition is [4–6]

$$\frac{d^2 w_m}{d\beta_m^2} + \frac{d^2 w_m}{d\beta_m d\beta_r} < 0\,, \quad \text{(S10)}$$

where the first term on the right hand side is given by equation (S9) and the second term is given by

$$\frac{d^2 w_m}{d\beta_m d\beta_r} = \frac{\frac{\partial^2 x_1}{\partial \beta_1 \partial a_1} \frac{\partial^2 x_2}{\partial \beta_2 \partial a_2}}{|J|^2}$$
$$\left[\frac{\partial^2 x_2}{\partial a_2^2}, -\frac{\partial^2 x_2}{\partial a_1 \partial a_2} \right] \cdot \left[H(u_1) + \frac{\left(\frac{\partial u_1}{\partial a_1} \frac{\partial^2 x_1}{\partial a_2 \partial a_1} - \frac{\partial u_1}{\partial a_2} \frac{\partial^2 x_1}{\partial a_1^2} \right)}{|J|} H(\frac{\partial x_2}{\partial a_2}) \right] \cdot \left[-\frac{\partial^2 x_1}{\partial a_2 \partial a_1}, \frac{\partial^2 x_1}{\partial a_1^2} \right]^{\mathsf{T}}$$
$$+ \frac{\frac{\partial^2 x_1}{\partial \beta_1 \partial a_1} \left(\frac{\partial u_1}{\partial a_1} \frac{\partial^2 x_1}{\partial a_2 \partial a_1} - \frac{\partial u_1}{\partial a_2} \frac{\partial^2 x_1}{\partial a_1^2} \right)}{|J|^2} \left[\frac{\partial^2 x_2}{\partial a_2^2}, -\frac{\partial^2 x_2}{\partial a_1 \partial a_2} \right] \cdot \left[\frac{\partial^3 x_2}{\partial \beta_2 \partial a_2^2}, \frac{\partial^2 x_2}{\partial \beta_2 \partial a_1 \partial a_2} \right]^{\mathsf{T}} \quad \text{(S11)}$$

S.3 Complementarity and mutual regard

In this section, we find sufficient conditions for $\beta > 0$ to be evolutionarily stable when individuals have generic other-regarding objectives and the payoff functions characterize a social interaction similar in spirit to a prisoner's dilemma. Let x_2 be the objective of individual 2. We assume that individual 2 is both "other-regarding" and "self-regarding", so $x_2 = G(u_1, u_2)$, for some function G, is increasing in both u_1 and u_2 and that x_2 is bounded and concave with respect to the payoffs u_1 and u_2; i.e., all pure second derivatives of x_2 with respect to u_1 and u_2 are negative. These conditions are given in the first row of (S12). Payoffs u_1 and u_2 are decreasing functions of the action of the focal individual and increasing functions of the partner's action. We also assume that the actions of individual 2 incur accelerating costs to individual 2 itself and yield benefits with diminishing returns to individual 1. The conditions on payoffs are given in the second and third rows of (S12).

$$\begin{array}{llll} \frac{\partial x_2}{\partial u_1} > 0 & \frac{\partial x_2}{\partial u_2} > 0 & \frac{\partial^2 x_2}{\partial u_1^2} < 0 & \frac{\partial^2 x_2}{\partial u_2^2} < 0 \\ \frac{\partial u_1}{\partial a_1} < 0 & \frac{\partial u_1}{\partial a_2} > 0 & \frac{\partial^2 u_1}{\partial a_2^2} < 0 & \\ \frac{\partial u_2}{\partial a_1} > 0 & \frac{\partial u_2}{\partial a_2} < 0 & \frac{\partial^2 u_2}{\partial a_2^2} < 0 & \end{array} \tag{S12}$$

The response coefficient ρ is given by:

$$\rho = -\left(\frac{\partial^2 x_2}{\partial a_1 \partial a_2} \Big/ \frac{\partial^2 x_2}{\partial a_2^2} \right)_{\substack{a_1 = a_1^* \\ a_2 = a_2^*}} .$$

The necessary condition for x_2 to be an ESS objective function is that ρ is positive, which means that there is positive feedback between the actions of the individuals. We will show that ρ is positive when,

$$\frac{\partial^2 x_2}{\partial u_1 \partial u_2} > 0 \tag{S13}$$

$$\frac{\partial^2 u_1}{\partial a_1 \partial a_2} \geq 0 \quad \frac{\partial^2 u_2}{\partial a_1 \partial a_2} \geq 0\,. \tag{S14}$$

The inequality in condition (S13) says that the payoffs have to be complementary inputs into the objective x_2. This is exactly how we defined "conditional regard" in equation (3) for the other-regarding objective function given in equation (2). The two inequalities in condition (S14) say that the actions have to be complementary inputs into payoffs.

To show $\rho > 0$, we will first calculate $\frac{\partial^2 x_2}{\partial a_1 \partial a_2}$ and show that its positive. If the objective x_2 is a function of u_1 and u_2, we can write using the chain rule:

$$\begin{aligned} \frac{\partial^2 x_2}{\partial a_1 \partial a_2} =& \frac{\partial x_2}{\partial u_1} \frac{\partial^2 u_1}{\partial a_1 \partial a_2} + \frac{\partial x_2}{\partial u_2} \frac{\partial^2 u_2}{\partial a_1 \partial a_2} \\ &+ \frac{\partial u_1}{\partial a_1} \left(\frac{\partial^2 x_2}{\partial u_1^2} \frac{\partial u_1}{\partial a_2} + \frac{\partial^2 x_2}{\partial u_1 \partial u_2} \frac{\partial u_2}{\partial a_2} \right) \\ &+ \frac{\partial u_2}{\partial a_1} \left(\frac{\partial^2 x_2}{\partial u_1 \partial u_2} \frac{\partial u_1}{\partial a_2} + \frac{\partial^2 x_2}{\partial u_2^2} \frac{\partial u_2}{\partial a_2} \right) . \end{aligned} \tag{S15}$$

Given our previous assumptions in (S12), the conditions (S13) and (S14) guarantee that $\frac{\partial^2 x_2}{\partial a_1 \partial a_2} > 0$. Next, we calculate the denominator, $\frac{\partial^2 x_2}{\partial a_2^2}$, which must be negative for the behavioral equilibrium to be stable (see equation (S1)):

$$\begin{aligned}\frac{\partial^2 x_2}{\partial a_2^2} =& \frac{\partial x_2}{\partial u_1}\frac{\partial^2 u_1}{\partial a_2^2} + \frac{\partial x_2}{\partial u_2}\frac{\partial^2 u_2}{\partial a_2^2} \\ &+ \frac{\partial^2 x_2}{\partial u_1^2}\left(\frac{\partial u_1}{\partial a_2}\right)^2 + \frac{\partial^2 x_2}{\partial u_2^2}\left(\frac{\partial u_2}{\partial a_2}\right)^2 + 2\frac{\partial^2 x_2}{\partial u_1 \partial u_2}\frac{\partial u_1}{\partial a_2}\frac{\partial u_2}{\partial a_2}\,.\end{aligned} \tag{S16}$$

Our assumptions in (S12) and the condition (S13) guarantee that $\frac{\partial^2 x_2}{\partial a_2^2} < 0$.

We find that the additive payoff functions given in equation (1) and multiplicative objective given in equation (2) meet both our assumptions in (S12) and our complementarity conditions in (S13) and (S14). Other simple objective functions do not necessarily generate the positive feedback of actions between individuals necessary for mutual regard. For example, suppose the objective function of individual 2 is given by

$$x_2(a_1, a_2) = u_2(a_1, a_2) + \beta_2 u_1(a_1, a_1)\,. \tag{S17}$$

In this case, x_2 is an other-regarding objective for positive β_2, which can still be thought of as a measure of the strength of that other-regard. However, using the same payoff functions defined in equation (1), equation (S15) shows that $\rho = 0$. The first two terms of (S15) are zero since the mixed second derivatives of the payoff functions are zero, and the linearity of the objective function means that the two terms in parentheses in (S15) are also zero. Thus, the additive objective function in (S17) generates no positive feedback and cannot support an evolutionarily stable level of mutual regard $\beta > 0$.

S.4 Some results about Pareto efficiency at the behavioral equilibrium and ESS outcome

First, we write the condition for an outcome to lie on the Pareto boundary, which is the set of payoff pairs that are Pareto efficient, i.e. payoff pairs for which no player can increase its payoff without decreasing the payoff of another player. This means that for a fixed u_1, u_2 is maximized on the Pareto boundary, and likewise u_1 is maximized for a fixed u_2. If we set $u_2(a_1, a_2) = \chi$ where

χ is a constant, the Implicit Function Theorem says that there exist some $g(a_1, \chi) = a_2$ for which $u_2(a_1, g_1(a_1, \chi)) = \chi$. We maximize u_1 at this value by differentiating

$$\frac{du_1(a_1, g(a_1, \chi))}{da_1} = \frac{\partial u_1}{\partial a_1} + \frac{\partial u_1}{\partial a_2}\frac{\partial g}{\partial a_1} = 0\,. \tag{S18}$$

Now, we can get $\frac{\partial g}{\partial a_1}$ by differentiating u_2, which by construction of the function g, must be constant:

$$\frac{du_2(a_1, g(a_1, \chi))}{da_1} = \frac{\partial u_2}{\partial a_1} + \frac{\partial u_2}{\partial a_2}\frac{\partial g}{\partial a_1} = 0\,. \tag{S19}$$

Substituting $\frac{\partial g}{\partial a_1}$ from (S19) into (S18) and rearranging gives:

$$\frac{\partial u_1}{\partial a_1}\frac{\partial u_2}{\partial a_2} = \frac{\partial u_1}{\partial a_2}\frac{\partial u_2}{\partial a_1} \tag{S20}$$

Pareto-efficiency implies concordance of objectives

First, we need to specify precisely what we mean by concordance of objectives. Remember that $\frac{\partial x_1}{\partial a_1}$ measures how much and in which direction individual 1 changes its action. Now consider the quantity $\frac{\partial x_2}{\partial a_1}$, which by the same interpretation, measures how much and in which direction individual 2 *would* change a_1, if only it had control over it. Thus, the difference $\frac{\partial x_1}{\partial a_1} - \frac{\partial x_2}{\partial a_1}$ can be thought of as measuring how much individuals 1 and 2 differ in their preference over a_1. Similarly, the difference $\frac{\partial x_2}{\partial a_2} - \frac{\partial x_1}{\partial a_2}$ is the difference between individuals in preferences over a_2. If $\frac{\partial x_1}{\partial a_1} - \frac{\partial x_2}{\partial a_1} = \frac{\partial x_2}{\partial a_2} - \frac{\partial x_1}{\partial a_2} = 0$, the two individuals do not differ in their preferences over which direction and how much they should adjust their actions; their objectives are concordant. The concordance conditions can hold locally at a single action pair, or globally over the entire action space. Because we focus near the behavioral equilibrium, we will deal with local concordance of objectives.

Consider now the following class of objectives that are functions of u_1 and u_2: $x_1 = x_1(u_1, u_2)$, $x_2 = x_2(u_1, u_2)$. The first-order conditions for the behavioral equilibrium is:

$$\frac{\partial x_1}{\partial a_1} = \frac{\partial x_1}{\partial u_1}\frac{\partial u_1}{\partial a_1} + \frac{\partial x_1}{\partial u_2}\frac{\partial u_2}{\partial a_1} = 0 \tag{S21}$$

$$\frac{\partial x_2}{\partial a_2} = \frac{\partial x_2}{\partial u_1}\frac{\partial u_1}{\partial a_2} + \frac{\partial x_2}{\partial u_2}\frac{\partial u_2}{\partial a_2} = 0 \tag{S22}$$

Now, from (S20), we can substitute $\frac{\partial u_1}{\partial a_1} = \frac{\partial u_1}{\partial a_2}\frac{\partial u_2}{\partial a_1} / \frac{\partial u_2}{\partial a_2}$ into (S21), and get:

$$\frac{\partial x_1}{\partial u_1}\frac{\partial u_1}{\partial a_2}\frac{\partial u_2}{\partial a_1} + \frac{\partial x_1}{\partial u_2}\frac{\partial u_2}{\partial a_1}\frac{\partial u_2}{\partial a_2} = 0 \tag{S23}$$

Canceling $\frac{\partial u_2}{\partial a_1}$, we end up with:

$$\frac{\partial x_1}{\partial u_1}\frac{\partial u_1}{\partial a_2} + \frac{\partial x_1}{\partial u_2}\frac{\partial u_2}{\partial a_2} = \frac{\partial x_1}{\partial a_2} = 0 \tag{S24}$$

Thus, at a Pareto optimal behavioral equilibrium, individual 1's objective is (locally) concordant with respect to individual 2's action a_2. Note that this also means that individual 1's objective function x_1 is a local peak (subject to second order conditions) at this behavioral equilibrium. One can show using the same argument that it must be $\frac{\partial x_2}{\partial a_2} = \frac{\partial x_2}{\partial a_1} = 0$, such that the individuals' objectives are concordant over a_1 as well at the Pareto optimal equilibrium.

Conditions for a Pareto-efficient ESS outcome and objectives

Now, we focus our attention specifically to the Pareto-efficient behavioral equilibrium of two individual that have an evolutionarily stable objective function. The game is symmetric ($u_1(a_1, a_2) = u_2(a_2, a_1)$). We further assume that the individuals' objective functions are functions of their payoffs: $x = x(u_1, u_2)$. Even though both individuals have the same objectives, we concentrate on individual 2 with its objective x_2 for consistency; the analysis is equally valid for x_1. Because the game is symmetric, the behavioral equilibrium in a monomorphic ESS also has to be symmetric, i.e. $a_1^* = a_2^* = a^*$. We have three equations that this outcome has to satisfy. First, the Pareto efficiency in a symmetric outcome implies that it maximizes $u_2(a, a) = u_1(a, a)$:

$$\frac{du_2(a,a)}{da} = \left[\frac{\partial u_2}{\partial a_1} + \frac{\partial u_2}{\partial a_2}\right]_{a_1=a_2=a^*} = \left[\frac{\partial u_1}{\partial a_1} + \frac{\partial u_1}{\partial a_2}\right]_{a_1=a_2=a^*} = 0 \tag{S25}$$

We denote the value of a satisfying equation (S25) by $\hat{a}$; i.e., $a^* = \hat{a}$. The Pareto-efficient action $\hat{a}$ must also satisfy the behavioral equilibrium conditions, which for individual 2 reads:

$$\left[\frac{\partial x_2}{\partial a_2}\right]_{a_1=a_2=\hat{a}} = \left[\frac{\partial x_2}{\partial u_1}\frac{\partial u_1}{\partial a_2} + \frac{\partial x_2}{\partial u_2}\frac{\partial u_2}{\partial a_2}\right]_{a_1=a_2=\hat{a}} = 0 \tag{S26}$$

Finally, we have the ESS condition:

$$\left[\frac{\partial u_1}{\partial a_1} - \left(\frac{\partial^2 x_2}{\partial a_1 \partial a_2} \bigg/ \frac{\partial^2 x_2}{\partial a_2^2}\right) \frac{\partial u_1}{\partial a_2}\right]_{a_1=a_2=\hat{a}} = 0 \tag{S27}$$

The Pareto efficiency condition (S25) and the symmetry of the payoff functions yield the following relations at the Pareto optimal outcome:

$$\frac{\partial u_1}{\partial a_1} = -\frac{\partial u_1}{\partial a_2} = -\frac{\partial u_2}{\partial a_1} = \frac{\partial u_2}{\partial a_2} \tag{S28}$$

With these relations, we can immediately see that the response coefficient $\rho = -\frac{\partial u_1}{\partial a_1} \big/ \frac{\partial u_1}{\partial a_2} = 1$. This is a slight generalization of the result from André and Day [7], who showed that the response coefficient at the payoff-maximizing equilibrium is equal to 1 for the continuous iterated Prisoner's Dilemma game. With $-\frac{\partial u_1}{\partial a_1} \big/ \frac{\partial u_1}{\partial a_2} = 1$, the ESS condition can be re-arranged and expanded using the chain rule:

$$\begin{aligned}
&-\Bigg[\frac{\partial^2 x_2}{\partial u_1^2}\left(\frac{\partial u_1}{\partial a_2}\right)^2 + \frac{\partial^2 x_2}{\partial u_1 \partial u_2}\frac{\partial u_1}{\partial a_2}\frac{\partial u_2}{\partial a_2} + \frac{\partial x_2}{\partial u_1}\frac{\partial^2 u_1}{\partial a_2^2} \\
&\qquad\qquad + \frac{\partial^2 x_2}{\partial u_2^2}\left(\frac{\partial u_2}{\partial a_2}\right)^2 + \frac{\partial^2 x_2}{\partial u_1 \partial u_2}\frac{\partial u_1}{\partial a_2}\frac{\partial u_2}{\partial a_2} + \frac{\partial x_2}{\partial u_2}\frac{\partial^2 u_2}{\partial a_2^2}\Bigg]_{a_1=a_2=\hat{a}} \\
&= \Bigg[\frac{\partial^2 x_2}{\partial u_1^2}\frac{\partial u_1}{\partial a_1}\frac{\partial u_1}{\partial a_2} + \frac{\partial^2 x_2}{\partial u_1 \partial u_2}\frac{\partial u_1}{\partial a_2}\frac{\partial u_2}{\partial a_1} + \frac{\partial x_2}{\partial u_1}\frac{\partial^2 u_1}{\partial a_1 \partial a_2} \\
&\qquad\qquad + \frac{\partial^2 x_2}{\partial u_2^2}\frac{\partial u_2}{\partial a_1}\frac{\partial u_2}{\partial a_2} + \frac{\partial^2 x_2}{\partial u_1 \partial u_2}\frac{\partial u_1}{\partial a_1}\frac{\partial u_2}{\partial a_2} + \frac{\partial x_2}{\partial u_2}\frac{\partial^2 u_2}{\partial a_1 \partial a_2}\Bigg]_{a_1=a_2=\hat{a}}
\end{aligned} \tag{S29}$$

Using the relations in (S28), we can cancel the terms involving the second derivatives of x_2 with respect to u_1 and u_2. We can also see using these relations that (S26) reduces to: $\frac{\partial x_2}{\partial u_1} = \frac{\partial x_2}{\partial u_2}$. Thus, equation (S29) can be re-written as:

$$-\left[\frac{\partial x_2}{\partial u_1}\left(\frac{\partial^2 u_1}{\partial a_1 \partial a_2} + \frac{\partial^2 u_2}{\partial a_1 \partial a_2}\right)\right]_{a_1=a_2=\hat{a}} = \left[\frac{\partial x_2}{\partial u_1}\left(\frac{\partial^2 u_1}{\partial a_2^2} + \frac{\partial^2 u_2}{\partial a_2^2}\right)\right]_{a_1=a_2=\hat{a}}, \tag{S30}$$

or, equivalently

$$\left[\frac{\partial x_2}{\partial u_1}\left(\frac{\partial^2 u_1}{\partial a_1 \partial a_2} + \frac{\partial^2 u_1}{\partial a_2^2} + \frac{\partial^2 u_2}{\partial a_1 \partial a_2} + \frac{\partial^2 u_2}{\partial a_2^2}\right)\right]_{a_1=a_2=\hat{a}} = 0\,. \tag{S31}$$

This equation can only be satisfied if the terms in the parentheses add up to 0, or if $\frac{\partial x_2}{\partial u_1} = 0$. Since the payoff function in a game is exogenously specified, the former case is non-generic and not likely to hold. Thus, we conclude that at an Pareto-efficient behavioral equilibrium whose objective functions are evolutionarily stable, the objective function must have a critical point in the u_1 direction. By repeating the same argument, but substituting $\frac{\partial x_2}{\partial u_1}$ instead of $\frac{\partial x_2}{\partial u_2}$, we can conclude that x_2 must also have a critical point in the u_2 direction at $(\hat{a}, \hat{a})$. With the further assumptions that $\frac{\partial^2 x_2}{\partial u_1^2} < 0$ and $\frac{\partial^2 x_2}{\partial u_1^2} < 0$ from section S.3, we can conclude that the objective function x_2 has to have a local peak at the Pareto efficient behavioral equilibrium, if it is to be an ESS.

References

[1] Edelstein-Keshet L (2005) *Mathematical Models in Biology* (Society for Industrial and Applied Mathematics, Philadelphia, PA).

[2] Rosen JB (1965) Existence and uniqueness of equilibrium points for concave N-person games. *Econometrica* 33:520–534.

[3] Marcus M, Minc H (1992) *A Survey of Matrix Theory and Matrix Inequalities* (Dover, New York, NY).

[4] Eshel I, Motro U (1981) Kin selection and strong evolutionary stability of mutual help. *Theor Popul Biol* 19:420–433.

[5] Taylor PD (1989) Evolutionary stability in one-parameter models under weak selection. *Theor Popul Biol* 36:125–143.

[6] Christiansen FB (1991) On conditions for evolutionary stability for a continuously varying character. *Am Nat* 138:37–50.

[7] André JB, Day T (2007) Perfect reciprocity is the only evolutionarily stable strategy in the continuous iterated prisoner's dilemma. *J Theor Biol* 247:11–22.

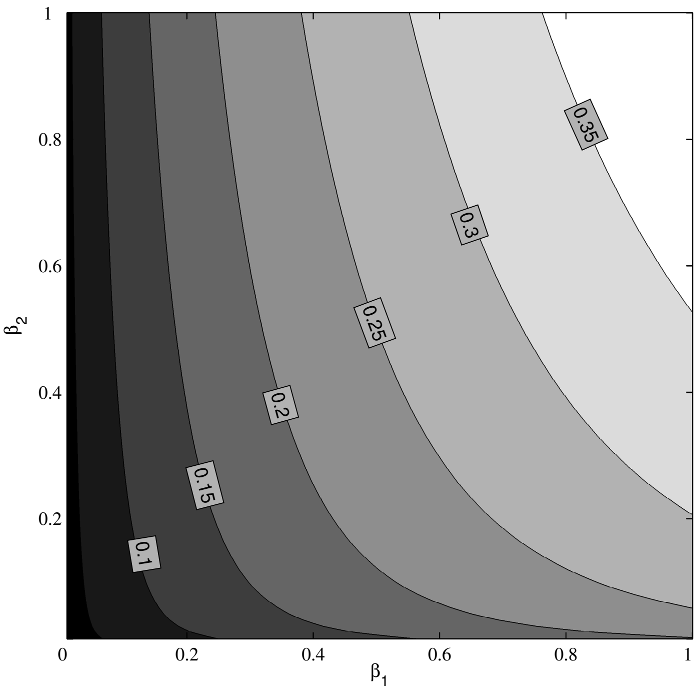

Figure S1: Contour plot of the action of individual 1 at the behavioral equilibrium, a_1^*, which is obtained by plugging the payoff functions in (1) and the objective functions in (2) into the equations in (5) and solving numerically for a_1^* and a_2^*. The value of a_1^* is given by the contour labels. Since the payoff and objective functions are symmetric, the plot of a_2^* can be obtained by simply rotating this plot about the $\beta_1 = \beta_2$ line. Using the resource sharing example, a_1^* can be thought of as the donation level of individual at the behavioral equilibrium. This level is an increasing function of both β_1, the level of other-regard the focal individual 1 has for its partner, and β_2, the level of other-regard the partner has for the focal individual. $\frac{\partial a_1^*}{\beta_1}$ is generally larger than $\frac{\partial a_1^*}{\beta_2}$, as one might expect given that the two individuals do not perfectly mirror each other's actions, though this difference decreases as both β_1 and β_2 increase.