A theory for the evolution of other-regard integrating proximate and ultimate perspectives

Erol Akçay[1,2], Jeremy Van Cleve[1,3], Marcus W. Feldman, and Joan Roughgarden

Department of Biology, 371 Serra Mall, Stanford University, Stanford, CA 94305

Edited by Simon A. Levin, Princeton University, Princeton, NJ, and approved September 15, 2009 (received for review April 21, 2009)

Although much previous work describes evolutionary mechanisms that promote or stabilize different social behaviors, we still have little understanding of the factors that drive animal behavior proximately. Here we present a modeling approach to answer this question. Our model rests on motivations to achieve objectives as the proximate determinants of behavior. We develop a two-tiered framework by first modeling the dynamics of a social interaction at the behavioral time scale and then find the evolutionarily stable objectives that result from the outcomes these dynamics produce. We use this framework to ask whether "other-regarding" motivations, which result from a kind of nonselfish objective, can evolve when individuals are engaged in a social interaction that entails a conflict between their material payoffs. We find that, at the evolutionarily stable state, individuals can be other-regarding in that they are motivated to increase their partners' payoff as well as their own. In contrast to previous theories, we find that such motivations can evolve because of their direct effect on fitness and do not require kin selection or a special group structure. We also derive general conditions for the evolutionary stability of other-regarding motivations. Our conditions indicate that other-regarding motivations are more likely to evolve when social interactions and behavioral objectives are both synergistic.

Animal behavior is determined both by proximate mechanisms that dictate an animal's actions in real time and by evolutionary forces that shape these proximate mechanisms. Even though the evolutionary dynamics of social behavior have been extensively studied (1–4), proximate mechanisms of behavior and how they interface with evolutionary forces remain poorly understood (4). In recent years, some models have integrated a proximate mechanism with an evolutionary analysis (5, 6). Furthermore, an explicitly two-tiered approach with potentially cooperative behavioral dynamics embedded in an evolutionary dynamic has been proposed (7) as necessary to understand the evolution of social behavior. We contribute to this literature by developing a unified framework for modeling the evolution of a specific type of behavioral interaction based on a well-defined proximate mechanism.

Our proximate mechanism is based on the notion that animals are motivated to achieve certain objectives. Goal-seeking behavior has been a recurring theme in animal behavior and has been an integral part of earlier ethological thinking (e.g. 8, 9). However, this idea lost its prominence after the emergence of modern behavioral ecology, which focuses mainly on the fitness consequences of behavior (see, for example, page 6 of ref. 10). In addition, proximate models of behavior based on goal-seeking have focused mostly on nonsocial behaviors such as foraging (9) and have rarely considered social interactions. Here, we study goal-seeking behavior in the context of a social interaction by developing a model of a pair of interacting animals whose motivations derive from their internal reward sensations, which they aim to maximize. These reward sensations, which could be encoded in specific neural circuits such as dopamine pathways involved in learning (11), are represented in our model as objective functions.

We consider a specific class of objectives functions that are based on the payoffs both individuals receive from the interaction. When the objective function of a focal individual positively weights the payoff of its social partner, the focal individual is said to exhibit an other-regarding preference for its partner (12). We focus on such other-regarding objectives for three reasons. First, the existence of other-regarding preferences has received substantial support recently from laboratory experiments that show a capacity in some nonhuman primates for unsolicited food sharing even when the recipient cannot reciprocate (13–15). Second, explanations for the evolution of such preferences are still in dispute (3, 4, 16) and often rely on costly punishment and reproductive differences between groups (12) or on indirect selection on kin (13, 17) instead of on direct selection on the actions of focal individuals. Third, from a conceptual perspective, an other-regarding preference is a simple way in which the behavioral objectives of two interacting individuals can be brought into concordance even when their payoff interests diverge. In this way, we can clearly delineate "altruistic" motivations driving a specific behavior from the underlying fitness consequences of such behavior.

We integrate the proximate model of behavioral objectives with an analysis of the selection pressures acting on those objectives and find two new results. First, we show that other-regarding objectives, and thus motivations, can evolve through direct selection on the fitness effects of individual behaviors. Second, we show that synergism in the payoffs from the social interaction and synergism in individuals' objectives promote the evolution of other-regarding objectives. These synergisms are directly related to how the benefits and costs of different behaviors are turned into payoffs and how individuals convert information about the payoffs from the social interaction into reward sensations.

Results

The Behavioral Model. We begin by developing a model of a social interaction in which two individuals share resources with each other. Imagine, for example, two capuchin monkeys, one having been given apples, the other carrots, as in the experiment by de Waal (18). Both individuals need the sugar in the apple and the β-carotene in the carrot, so each would do best to exchange some of its holdings. For simplicity, we assume that costs and benefits of sharing are the same for both individuals. We label the donation that a focal individual, individual 1(l1), makes to its partner, individual 2(l2), by a_1, and the donation that l2 makes to l1 by a_2. We term a_1 and a_2 individuals' "actions" and assume that $0 \leq a_1, a_2 \leq 1$. In the dynamical behavioral model below, the actions can be thought of as the rates at which individuals exchange donations. Suppose that the marginal benefit a food

Author contributions: E.A. and J.V. designed research; E.A. and J.V. performed research; and E.A., J.V., M.W.F., and J.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]E.A. and J.V. contributed equally to this work.

[2]To whom correspondence should be sent at the present address: National Institute for Mathematical and Biological Synthesis, 1534 White Ave, University of Tennessee, Knoxville, TN 37996. E-mail: erol@nimbios.org.

[3]Present address: Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501.

This article contains supporting information online at www.pnas.org/cgi/content/full/0904357106/DCSupplemental.

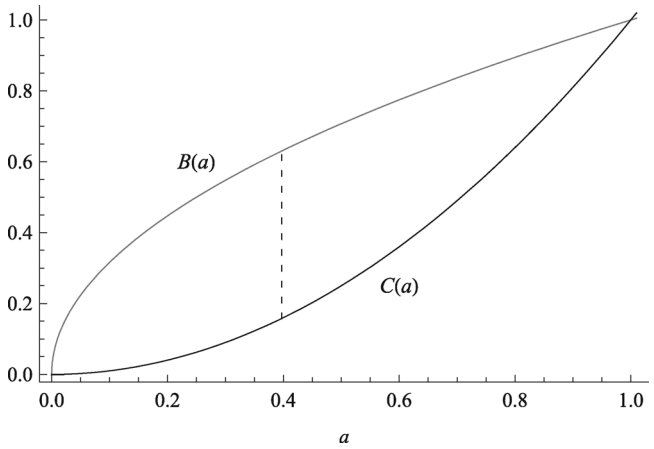

Fig. 1. Benefit and cost functions for the payoffs defined in Eq. **1** where $0 \leq a \leq 1$. For the payoff function defined in Eq. **1**, the maximum payoff obtainable in a monomorphic population is $u_1(0.397, 0.397) = 0.472$, which is denoted by the dashed line.

item has for an individual diminishes as the quantity the individual receives of that food item increases. Conversely, the marginal cost of giving up a food item increases as the amount of it the individual has decreases. The total costs and benefits to an individual of donating and receiving are denoted by the functions $C(a)$ and $B(a)$, respectively. The total payoff to an individual is then the sum of these costs and benefits, which we denote by $u_1(a_1, a_2)$ for l1, and $u_2(a_1, a_2)$ for l2. In particular, we assume

$$u_1(a_1, a_2) = B(a_2) - C(a_1) = \sqrt{a_2} - a_1^2$$
$$u_2(a_1, a_2) = B(a_1) - C(a_2) = \sqrt{a_1} - a_2^2. \quad [1]$$

These benefit and cost functions are plotted in Fig. 1. In the long run, the accumulation of these payoffs to an individual, be they food resources, territory, or some other quantity, will increase the fitness of that individual.

What an individual donates in the interaction is determined by its motivations, which we assume to be directed at maximizing some objective function. The objective function can be seen to describe the internal reward sensation of an individual given a state of the world; i.e., how much an individual "likes" a given outcome (e.g. a pair of donation rates a_1 and a_2). We assume that the objective function of an individual depends on its own payoff and possibly the payoff of its partner. Because the payoffs are functions of the actions, the value of the objective function is also determined by the actions. We denote the objective functions of l1 and l2 by $x_1(a_1, a_2)$, and $x_2(a_1, a_2)$, respectively. Just as the structure of the specific neural pathways responsible for internal reward sensations is influenced by genetic factors, we assume that the shape of an individual's objective function can be determined by its genotype. In particular, we assume that how an individual's objective function depends on its own payoff and its partner's payoffs is determined by a single-locus continuous trait that we denote by β. Each individual is characterized by a value of β (β_1 and β_2 for the focal individual and its partner, respectively).

The effect that β has on the shape or form of the objective function determines the behavioral interpretation of β. Here we are interested in objective functions in which β determines the existence and strength of the other-regarding preference of a focal individual for its partner. In this analysis, we assume that the objective functions take the following forms:

$$x_1(a_1, a_2) = u_1 u_2^{\beta_1} = \left(\sqrt{a_2} - a_1^2\right)\left(\sqrt{a_1} - a_2^2\right)^{\beta_1}$$
$$x_2(a_1, a_2) = u_1^{\beta_2} u_2 = \left(\sqrt{a_2} - a_1^2\right)^{\beta_2}\left(\sqrt{a_1} - a_2^2\right), \quad [2]$$

where β_1 and β_2 are nonnegative. Here, β_1 in Eq. **2** determines of how much l2's payoff is weighted in l1's objective. Thus, β_1 can be seen as a measure of the degree of the other-regarding preference that l1 has for l2. If $\beta_1 = 0$ and $x_1(a_1, a_2) = u_1(a_1, a_2)$, then l1 is motivated purely to increase its own payoff and has no other-regarding preference for l2. If, on the other hand, $\beta_1 = 1$, then l1 has equal regard for its partner's payoff (u_2) and its own payoff (u_1) when determining its actions, and it aims to maximize the product of the two, $u_1 u_2$. When $\beta_1 = \beta_2 = 1$, both individuals aim to maximize $u_1 u_2$, which is similar to the "team-play" dynamic proposed by Roughgarden et al. (7). This behavioral equilibrium, which maximizes the product of the payoffs, coincides with the Nash bargaining solution (NBS) of the game (19).

Although there are other kinds of objective functions that can be other-regarding, we focus on the objectives in Eq. **2** because these display a property that we call "conditional regard": When $\beta_1 > 0$, l1's motivation to increase u_2 is positively related to how well l1 is doing (i.e., the magnitude of u_1). Mathematically, this property can be expressed as

$$\frac{\partial^2 x_1}{\partial u_1 \partial u_2} > 0. \quad [3]$$

In *Synergism and Other-Regarding Objectives*, we show that the conditional-regard property plays an important role in the evolution of other-regarding objectives; specifically, certain objectives that do not satisfy Eq. **3**, such as an objective that sums the payoffs of the two individuals (see also the *SI Appendix*), cannot be evolutionarily stable (ES).

Within the time scale of the social interaction, the behavioral dynamics describe how individuals act and react to each other. At any given time, the actions of the individuals (i.e., the donation rates), result in a level of reward sensation for each that is captured by the value of the objective functions x_1 and x_2. Individuals then adjust their actions to increase this reward sensation, i.e. they increase the value of their objective functions. We model this kind of behavioral interaction mechanistically with the following gradient dynamic:

$$\frac{da_1}{dt} = \frac{\partial x_1}{\partial a_1}$$
$$\frac{da_2}{dt} = \frac{\partial x_2}{\partial a_2}. \quad [4]$$

Here the partial derivative of an individual's objective function with respect to its own action can be viewed as measuring the motivation of that individual to increase or decrease its action. Note that the only requirement for individuals to follow these behavioral dynamics is that they can sense their objective function locally (or myopically); this does not imply that individuals are aware of the global shape of their objective functions, which would allow longer-term calculations based on that information (see *Discussion* for more on this issue). The behavioral dynamics in Eq. **4** come to an equilibrium if both individuals are at local maxima of their objective functions, which happens when

$$\frac{\partial x_1}{\partial a_1} = \frac{\partial x_2}{\partial a_2} = 0. \quad [5]$$

When the behavioral dynamics reach an equilibrium, the donation rates a_1 and a_2 stay constant; we denote their values at the behavioral equilibrium by a_1^* and a_2^*. At this behavioral equilibrium, l1 and l2 accrue payoffs at rates $u_1(a_1^*, a_2^*)$ and $u_2(a_1^*, a_2^*)$, respectively. Notice that because the objectives x_1 and x_2 are parametrized by β_1 and β_2, respectively, the behavioral equilibrium values of the actions will also depend on β_1 and β_2, which means that the payoffs that the focal individual and its partner receive at the behavioral equilibrium are each a function of both β_1 and β_2. Thus, the resource donations l1 and l2 make at the behavioral equilibrium, a_1^* and a_2^*, in the resource-sharing example above are both

functions of the levels of other-regard, β_1 and β_2. In this case, both a_1^* and a_2^* are increasing in both β_1 and β_2 (see Fig. S1 in the *SI Appendix*), which means that, although our model of the behavioral dynamics in Eq. **4** is mechanistic and does not allow individuals to make strategic actions at each time step of the social interaction, the behavioral equilibrium does represent a kind of negotiated outcome. In fact, the behavioral equilibrium given by the solution to Eq. **5** might also be obtained through an explicit negotiation process or other kind of strategic interaction.

Given that the fitness of each individual is an increasing function of the payoffs at the behavioral equilibrium, the genetically determined trait β specifically links the behavioral and evolutionary outcomes by affecting the equilibrium actions a_1^* and a_2^*. With this link, we can now take an ultimate perspective by asking how β evolves and whether values of $\beta > 0$ can be ES.

Evolutionary Stability Analysis. For the evolutionary analysis, we need to specify the fitness of a focal individual with trait β_1 when interacting with a partner that has a trait value of β_2. We denote the fitness of 1l by w_1, which is a function of both β_1 and β_2. We assume that individuals interact and reproduce within a panmictic population of infinite size. In such a population, there is only direct selection and no kin selection. To simplify matters, we assume that social partners are randomly assigned in the population and that each individual has only one social interaction. We also assume that individuals quickly negotiate to the behavioral equilibrium, where they accumulate payoff at rate $u_1(a_1, a_2)$, and $u_2(a_1, a_2)$. In that case, we can take the fitness to be equal to the payoff at the behavioral equilibrium and write

$$w_1(\beta_1, \beta_2) = u_1(a_1^*, a_2^*) = u_1(a_1^*(\beta_1, \beta_2), a_2^*(\beta_1, \beta_2)), \qquad [6]$$

where the dependence on β_1 and β_2 is through the behavioral equilibrium actions a_1^* and a_2^*.

In order to find out whether other-regarding motivations with $\beta > 0$ are ES, we perform a standard evolutionarily stable strategy (ESS) analysis (20, 21) where β can be interpreted as a strategy in the evolutionary game. The ES value of β, which we denote by β^*, is the value that guarantees that no mutant individual with a value of $\beta \neq \beta^*$ can attain a higher fitness than a resident individual with the ES value β^* when the population is nearly fixed for β^*. In *The Behavioral Model*, we referred to the two positions individuals can occupy in the interaction as 1 and 2. In the evolutionary analysis, however, we need to distinguish between individuals with resident and mutant genotypes. Therefore, in the following analysis, we use a subscript m to indicate an individual with a mutant value of $\beta = \beta_m$ and subscript r to denote a resident individual with value of $\beta = \beta_r$. Because the game is symmetric, we can adopt the convention that the mutant individual is always in position 1, such that the fitness of a mutant, w_m, interacting with a resident is given by Eq. **6** with $\beta_1 = \beta_m$ and $\beta_2 = \beta_r$. Assuming that mutant values of β are close to the resident value, we can look for an ESS value of β by differentiating the fitness of a focal mutant individual, w_m with respect to β_m; this derivative will be zero at the ESS when the focal mutant has $\beta_m = \beta_r = \beta^*$. With our definition of fitness given in Eq. **6**, we can write this first order ESS condition for the focal mutant individual as

$$\frac{\partial w_m}{\partial \beta_m} = \frac{\partial u_1}{\partial a_1}\frac{\partial a_1^*}{\partial \beta_m} + \frac{\partial u_1}{\partial a_2}\frac{\partial a_2^*}{\partial \beta_m} = 0, \qquad [7]$$

where the partial derivatives of u_1 are evaluated at a_1^* and a_2^* and $\beta_m = \beta_r = \beta^*$. Given a candidate ESS β^* that satisfies the condition in Eq. **7**, we also need a second-order condition to show that the candidate ESS is ES (20, 21); this condition is given in Eq. **S9** in the *SI Appendix*. An additional second-order condition given in the *SI Appendix* (Eq. **S11**) allows convergence stability, which means that a population can approach a candidate ESS through a succession of mutant invasions that sweep to fixation (22). Such an ESS is called a continuously stable strategy (CSS) and can be seen as the phenotypic end point of long-term evolutionary dynamics (23).

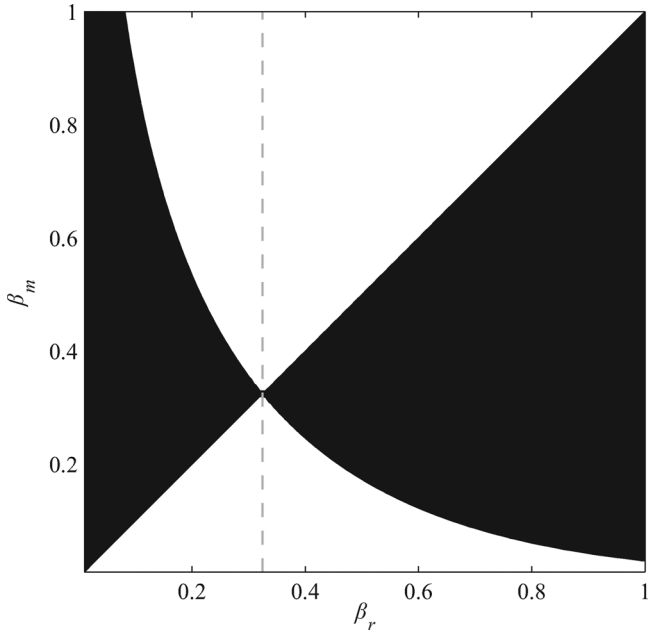

Fig. 2. This frame shows a pairwise invasibility plot for β using the fitness function given in Eq. **6**, the payoff functions defined in Eq. **1**, and the objective functions given in Eq. **2**. In the black regions, rare mutants with a value of $\beta = \beta_m$ attain a higher fitness than resident individuals with $\beta = \beta_r$ when both individuals interact with residents; these rare mutants invade the resident population. In the white regions, mutants have a lower fitness and cannot invade. The ESS $\beta^* = 0.3246$, is denoted by a gray dashed line. If $\beta_r > \beta^*$, mutants invade when β_m is slightly smaller than β_r because these points lie in the black region to the right of the dashed line. Likewise, mutants invade when $\beta_r < \beta^*$ and $\beta_m > \beta_r$ for β_m close to β_r. Therefore, rare mutants invade a resident population only when their β_m is closer to β^* than β_r, which means that $\beta^* = 0.3246$ is convergence stable.

We solve Eq. **7**, and obtain a candidate ESS value of β by substituting the objective functions in Eq. **2** into the equations for the behavioral equilibrium in Eq. **5**. The resulting expressions are then solved numerically for a_1^* and a_2^*, given β_m and β_r. By using the values of a_1^* and a_2^* as functions of β_m and β_r, we find a candidate ESS by setting $\beta_m = \beta_r = \beta^*$ in Eq. **7** and again solving the system of equations numerically. By using this method, we find a single candidate ESS, $\beta^* = 0.3246$, which is both evolutionarily and convergence stable and is shown graphically in Fig. 2 with a pairwise invasibility plot (24). In the black regions of Fig. 2, the fitness of a focal mutant with β_m is higher than the fitness of a focal resident with β_r when both individuals interact with residents. When rare, these mutants increase in frequency and invade the population. In the white region, mutant focal individuals with β_m attain a lower fitness than focal residents when both individuals interact with other residents; such mutants cannot invade. A resident population with $\beta_r = \beta^*$ is indicated by the gray dashed line; because the dashed line lies completely in white regions, no mutant with $\beta_m \neq \beta^*$ can invade, and β^* is ES. The ESS $\beta^* = 0.3246$ is also convergence stable, as described in Fig. 2.

Fig. 2 also reveals that resident population with β_r very small can be invaded by mutants with β_m slightly larger. When $\beta_r = 0$, however, residents maximize their individual payoff and set their donation rate at the behavioral equilibrium to zero ($a_2^* = 0$) regardless of what their partners do. Because of the conditional-regard property of the objectives in Eq. **2**, a mutant with $\beta_m > 0$ interacting with such a resident will also set its donation rate to zero. Thus, any mutant with $\beta_m > 0$ will have the same fitness as a resident individual with $\beta_r = 0$ and can invade the resident population through genetic drift. Once a positive β initially becomes

fixed through drift, larger values of β can evolve because β^* is both evolutionarily and convergence stable.

Finally, we should note that the fitness of individuals in a population fixed for the ESS β^* is higher than that of individuals in a population fixed for $\beta < \beta^*$ but lower than the maximum attainable fitness in a monomorphic population, i.e. the NBS, which occurs in a population fixed for $\beta = 1$. However, Fig. 2 indicates that $\beta = 1$ can be invaded by any other mutant β; thus, the NBS outcome will not be ES within the class-objective functions we consider here (see section S.4 of the *SI Appendix* for more on Pareto efficient outcomes like the NBS).

We have shown that a certain class of other-regarding objectives, given in Eq. **2**, can be ES because of their individual effects on fitness. We would also like to know more generally what types of social interactions, encapsulated in the payoff functions u and behavioral objectives x, make other-regarding preferences ES. In the next section, we take up this issue.

Synergism and Other-Regarding Objectives. Here we present conditions on generic payoff and objective functions that can make other-regarding objectives ES. The first-order ESS condition in Eq. **7** is shown in the *SI Appendix* (see Eq. **S8**) to be equivalent to

$$\left[\frac{\partial u_1}{\partial a_1} - \left(\frac{\partial^2 x_2}{\partial a_1 \partial a_2} \Big/ \frac{\partial^2 x_2}{\partial a_2^2}\right)\frac{\partial u_1}{\partial a_2}\right]_{\substack{a_1=a_1^* \\ a_2=a_2^*}} = 0 \qquad [8]$$

for generic payoff and objective functions. Here, x_2 is the objective function of the social partner of a focal mutant individual. All derivatives are evaluated at a_1^* and a_2^* corresponding to $\beta_1 = \beta_2 = \beta^*$. The first term on the left-hand side of this equation gives the direct effect of changing the focal l1's action on its own payoff, whereas the second term gives the indirect effect through the feedback of l1 on l2's action. Thus, the condition Eq. **8** stipulates that these two effects need to exactly cancel out at a behavioral equilibrium corresponding to an ESS, so that a mutant l1 cannot increase its fitness by changing its action. The strength of the feedback is measured by the response coefficient ρ, given by:

$$\rho = -\left(\frac{\partial^2 x_2}{\partial a_1 \partial a_2} \Big/ \frac{\partial^2 x_2}{\partial a_2^2}\right)_{\substack{a_1=a_1^* \\ a_2=a_2^*}}. \qquad [9]$$

In order to further analyze Eq. **8**, we must first make some assumptions about $\frac{\partial u_1}{\partial a_1}$ and $\frac{\partial u_1}{\partial a_2}$. We are generally interested in dyadic interactions characterized by (i) the presence of conflict between the payoff interests of the two individuals, and (ii) the possibility of mutually beneficial outcomes. Formally, we suppose that $\frac{\partial u_1}{\partial a_1} < 0$ and $\frac{\partial u_2}{\partial a_2} < 0$ at some behavioral equilibrium, meaning that a mutant focal l1 could increase its payoff by decreasing its action, if the action of its partner, the resident l2, was held constant. Suppose also that $\frac{\partial u_1}{\partial a_2} > 0$ and $\frac{\partial u_2}{\partial a_1} > 0$ so that the action of an individual actually helps its partner. Both of these conditions are satisfied in the payoff functions given in Eq. **1**. Finally, in order to determine when other-regarding objectives can be ES, we will assume that the objective function x_2 is other-regarding, i.e. $\frac{\partial x_2}{\partial u_2} > 0$ as well as $\frac{\partial x_2}{\partial u_1} > 0$, and then look for any additional conditions that need to be met for evolutionary stability.

Under the assumptions in the preceding paragraph, the ESS condition Eq. **8** can only be satisfied if $\rho > 0$; in other words, a resident l2 responds to changes in a_1 by adjusting a_2 in the same direction. The stability of the behavioral equilibrium requires that the denominator in the right-hand side of Eq. **9** is negative (see Eqs. **S1** and **S16** in the *SI Appendix*). Thus, we have a positive response coefficient when the numerator is positive, i.e.

$$\frac{\partial^2 x_2}{\partial a_1 \partial a_2} > 0. \qquad [10]$$

Recall that the objective function x_2 is related to the actions a_1 and a_2 through the payoffs u_1 and u_2. Therefore, we can expand the derivative in Eq. **10** using the chain rule (see Eq. **S15** in the *SI Appendix*). The sufficient conditions for the inequality in Eq. **10** to be satisfied are

$$\frac{\partial^2 u_1}{\partial a_1 \partial a_2} \geq 0 \quad \frac{\partial^2 u_2}{\partial a_1 \partial a_2} \geq 0 \qquad [11]$$

and

$$\frac{\partial^2 x_2}{\partial u_1 \partial u_2} > 0. \qquad [12]$$

The common feature in all three conditions is that they require cross-derivatives of their respective functions to be positive.

Looking at these conditions more closely, we can interpret the inequalities in Eq. **11** based on how payoff functions operate in a social interaction. The payoff function in a social interaction can be thought of as similar to the production of goods by a factory using inputs such as raw material and labor. Thus, the cross-derivatives of the payoff with respect to its arguments acquire a meaning connected to the economics of production: They measure whether the inputs (i.e., the actions a_1 and a_2) are substitutes or complements for each other. If inputs are substitutes, having more of one input (i.e., an increase in one individual's action) decreases the marginal value of the other input and the cross-derivative is negative; conversely, with complementary inputs, having more of one input increases the marginal value of the other, and the cross derivative is positive. The latter might happen, for example, if the actions stand for donating food items with different essential nutrients. Complementarity is related to synergism, as defined by Queller (25), which involves a similar kind of positive nonadditivity. Synergism at the fitness level has long been known to promote the evolution of cooperation (25), and the inequalities in Eq. **11** indicate that complementarity in payoffs plays a similarly important role in making other-regarding objectives ES.

The meaning of this result is more readily apparent when considering a special subset of payoff functions where the payoff derives from a benefit produced as a result of the individuals' investments minus the private costs each individual incurs to invest:

$$\begin{aligned} u_1 &= B(a_1, a_2) - C(a_1) \\ u_2 &= B(a_2, a_1) - C(a_2). \end{aligned} \qquad [13]$$

These payoff functions include the functions in Eq. **1**, as well as public good games (26) and the snowdrift game (27). For this class of functions, the condition in Eq. **11** reduces to

$$\frac{\partial^2 B}{\partial a_1 \partial a_2} \geq 0, \qquad [14]$$

meaning that the production of benefits involves complementary investments by the two individuals. Therefore, the biology of how the benefits are produced determines whether mutual regard will evolve or not. For example, complementarity effects can be expected when individuals are specialized to contribute different resources (such as different food types) or different skills (such as nest defense vs. foraging) or when the benefit involves accelerating returns to investment for some range of investments.

Fig. 3 illustrates the above result. It depicts the ESS level of other-regard, β^*, for a family of benefit functions $B(a_1, a_2) = \sqrt{a_1 + a_2 + \nu a_1 a_2}$, where ν is a parameter measuring how synergistic the individuals' actions are in producing benefits. The cost function is quadratic as in Eq. **1**. For low values of ν, the actions are substitutes in the benefit function (i.e., the condition in Eq. **11** is not met), and no positive β is ES. As ν increases, the degree of complementarity in producing benefits increases. Consequently, β^* first becomes positive and then increases with

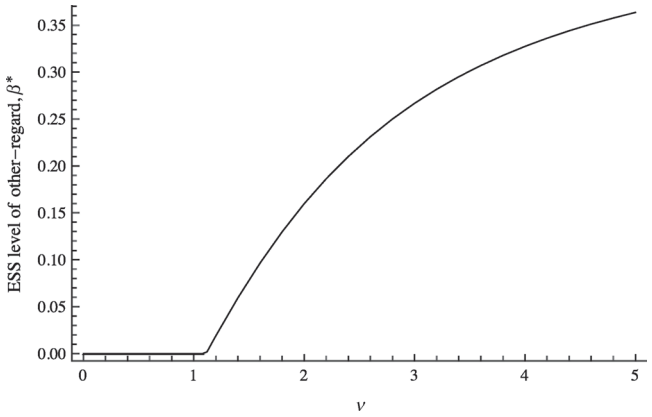

Fig. 3. The change in the ESS β^* as the payoff function underlying the interaction becomes more complementary, regulated by the parameter on the *x*-axis, ν. Specifically, the benefit function is given by $B(a_1, a_2) = \sqrt{a_1 + a_2 + \nu a_1 a_2}$, and the cost function is quadratic as in Eq. **1**. As ν increases, the actions of the individuals become more complementary (i.e. the cross-derivative of B becomes more positive), and consequently the ES value of β increases.

increasing ν. In general, the ES level of other-regarding preference (β^*) will be higher when the benefit function has a greater degree of complementarity. This relationship can be used to make empirical predictions about how the ecology of the social interaction affects the social traits of individuals in an interaction (see *Discussion*).

The condition in Eq. **12**, on the other hand, is identical to the definition of the "conditional regard property" given in Eq. **3**, which means that objective functions that are of the product form in Eq. **2** are more conducive to the evolution of other-regard than other functional forms, such as an additive form like $x_2 = \beta u_1 + u_2$. Therefore, this condition predicts that other-regarding objectives are likely to display the conditional regard property as well.

Discussion

In this paper, we present a simple model that explicitly integrates a proximate mechanism of behavior—motivations to achieve certain behavioral objectives—with the ultimate mechanism, natural selection. Integrating the two dynamical tiers at which behavior is determined leads to two new results.

Our first new result shows that other-regarding objectives (or preferences) can evolve without kin selection or selection acting between groups, and can even evolve in interactions where individuals face a strict payoff conflict (i.e. increasing one individual's payoff necessarily decreases the other's). This result rests on the fact that in our model, behavioral actions are determined dynamically in real time. These dynamics set the stage for behavioral feedbacks between individuals' actions, which are encapsulated by the response coefficient ρ in Eq. **8**. The behavioral feedback generates a positive association between the genotype of the focal individual and the actions of its partner, which has been long recognized as the central condition for the evolution of cooperation (25). Both individuals increase their actions because of this behavioral feedback, which generates a mutually beneficial outcome. Therefore, in terms of inclusive fitness theory, other-regarding objectives are selected for their direct fitness effects (1, 4). Previously, other-regarding preferences were thought to be associated with group selection (12, 16), or cooperative breeding and hence kin selection (17). Our results show that neither scenario is required for other-regarding preferences to evolve and are consistent with recent studies that show other-regarding preferences in a noncooperatively breeding primate (14, 15).

This result also relates to the issue of whether and how natural selection can bring the objectives of unrelated individuals into concordance, as occurs in team-play dynamics (7). Other-regarding objectives, as they evolve in our model, represent an intermediate stage between entirely "selfish" objectives (aimed at maximizing one's own payoff) and entirely concordant objectives (both individuals maximizing the same objective) and can be seen as a first step in the evolution team play. Our results indicate that this first step can occur through selection on the direct fitness effects of the social behavior. The evolution of the capacity for coordinated action and entirely concordant objectives remain as unresolved issues, although we can offer some speculation by noting the connection between concordant objectives and Pareto-efficient behavioral outcomes, which are the ones that exhaust all mutually beneficial possibilities in an interaction. In section S.4 of the *SI Appendix*, we show that Pareto efficiency of a behavioral equilibrium implies that individuals' objectives are concordant at that behavioral equilibrium. We also show that Pareto-efficient objectives can be ES and give additional analytical conditions that such objectives must satisfy. The results of André and Day (28) and Dekel et al. (29) further suggest that any such Pareto-efficient equilibrium will not only be ES, but will be the only convergence-stable equilibrium in a finite population. The next step in this work is to determine plausible objective functions that can yield a Pareto-efficient outcome.

Our second result has two components: First, we show that the more individuals' actions are complementary in producing benefits, the more likely other-regarding objectives are to evolve and the higher the level of other-regard will be. Thus, we predict that the incidence of other-regarding objectives and the level of regard should correlate positively with ecological scenarios that create complementarity. For example, if the production of benefit (such as hunting or raising offspring) requires extensive coordination between partners or if partners specialize to different tasks, the payoff function will exhibit complementarity, and partners will have other-regarding objectives. Second, we show that some forms of objective functions are more conducive to the evolution of other-regarding preferences. In particular, conditional regard, which results in synergism in the production of reward sensations, promotes the evolution of other-regarding objectives. Conditional regard amplifies the level of regard an actor has for its partner's payoff as a function of the actor's payoff; this effect can be measured in an experimental setting with choice trials or tests of inequity aversion. Our result suggests that other-regarding objectives and conditional regard should correlate, thus providing another hypothesis for comparative research in social species.

It is useful to compare our model with the one introduced by McNamara et al. (5) and Wahl and Nowak (6), which also has a two-tier approach to modeling behavioral and evolutionary dynamics. In that model, individuals respond to each other by using linear-response rules, the slopes of which are determined genetically and are subject to evolution. Recently, André and Day (28) provided a detailed analysis of this model, focusing on the slope of the linear-response rules. In a local analysis centered on the behavioral equilibrium, this slope, or "responsiveness" of individuals, corresponds to our response coefficient, ρ, and our ESS analysis is analogous to theirs. However, André and Day (28) also consider responsiveness as a global description of the behavioral dynamics. In contrast, both the local and global dynamics in our model are driven by payoff-based motivations, and ρ is an index summarizing the feedback brought about by the individuals' motivations in the vicinity of the behavioral equilibrium. In general, ρ can describe a wide variety of evolved behavioral feedbacks; thus, our behavioral model represents a less-restrictive approach to how animals make decisions.

On the other hand, we too restrict our strategy set through the use of a multiplicative objective function modulated by a single parameter β. In a larger space of possible objective functions, our result regarding the evolutionary stability of other-regard might not hold true. This caveat applies to both our model and the

linear-response rule models (5, 6, 28) and highlights a structural feature of two-tiered models that calculate fitness at the behavioral equilibrium: Any strict ESS found in a restricted strategy space will only be neutrally stable against mutants from a larger strategy space that lead to the same behavioral equilibrium. Furthermore, a different parametrization of the objective function would lead to different ES objectives and different behavioral outcomes. Therefore, the value of an ESS analysis in a restricted strategy space is not that it identifies universally robust ESS outcomes but that it reveals the selection pressures acting in those dimensions that we focus on. Animals' objectives are bound to vary in many dimensions, and therefore more types of objectives, motivated by empirical evidence, and their interrelationships need to be modeled in order to arrive at a complete picture of how the proximate mechanisms of behavior evolve. We believe that such analyses in restricted strategy spaces, combined with the type of analysis presented in *Synergism and Other-Regarding Objectives*, are powerful tools to elucidate selection pressures acting on objectives and to generate testable comparative hypotheses.

Our model is also closely related to the "indirect-evolution approach" in economics (30, 31). This approach ascribes preference functions to agents and models the evolution of the preferences according to the Nash equilibria they produce in a noncooperative game. The main difference between our approach and the indirect-evolution models is that the latter assume strategic agents that can calculate best responses and reason their way to an equilibrium, whereas we model animals that "myopically" follow the gradient of their objectives. If individuals are strategic agents, then their actions depend on the type of this partner, so information about the partner's type becomes important for determining actions. Accordingly, results from the indirect-evolution literature show that for nonindividualistic preferences to be stable, sufficiently reliable information about the preferences of interacting individuals must be available (29, 32) because it is individually advantageous to increase one's partner's payoff only when one is sufficiently certain that the partner will also do the same. In contrast, we assume that in a social interaction individuals act and react to each other in close physical proximity rather than make independent decisions by individual reasoning. In such dynamics, even though individuals at any given instant act with regard only to the local shape of their own objectives, the behavioral feedbacks over the course of the interaction allow them, in effect, to learn the global shape of each other's objectives. Thus, the outcome is the same as that produced by fully rational agents with full information about each others' preferences. We believe that our behavioral model is applicable to a wider range of species because strategic reasoning requires considerable cognitive machinery and imposes time costs on the actors, and therefore is likely to be reserved for situations where myopic responses are inadequate and sufficient information about partners is available. Additionally, because of the mathematical correspondence between the approaches, our results on the complementarity in payoff functions and conditional regard represent contributions to the study of preferences in the indirect-evolution literature.

Finally, our behavioral model based on other-regarding motivations and objectives yields important empirical predictions for proximate mechanisms of behavior. Most directly, field observations of an other-regarding individual would see it spontaneously undertaking actions that could be costly to itself but benefit another individual with whom it interacts. This spontaneous helping should occur even among unrelated individuals and in situations in which immediate reciprocity is unlikely (13). If a species has evolved other-regarding objectives, directly assaying the motivation of an individual animal [e.g. by choice trials (13, 14) or functional MRI of reward pathways (34)] would reveal genuine caring for the payoff of the other. Other mechanisms, such as reciprocal altruism (2, 33) and strong reciprocity (12), can also produce a high degree of helping between unrelated individuals. Reciprocity might be achieved by individuals explicitly attempting to maximize their own payoffs across multiple rounds of an interaction, but it can also be generated by other-regarding objectives that change in response to previous interactions.

The behavior that results from other-regarding objectives has been termed empathy-based altruism by de Waal (35) [distinct from altruism in the evolutionary sense (1, 4)]. Our model suggests that empathy-based altruism should be much more common in nature than conventionally recognized, as it does not require kin or group selection. This suggestion is supported by the empirical evidence for other-regarding preferences that has been accumulating at an increasing rate (e.g. 13–15).

ACKNOWLEDGMENTS. We thank Laurent Lehmann for extensive discussions, Drew Fudenberg for providing helpful references, Andy Gardner and three anonymous referees for constructive feedback and input that improved this manuscript. J.V. was supported by National Library of Medicine Training Grant LM-07033. This research was supported in part by National Institutes of Health grant GM-28016 (to M.W.F.).

1. Hamilton WD (1964) The genetical evolution of social behaviour i. *J Theor Biol* 7:1–16.
2. Axelrod R, Hamilton WD (1981) The evolution of cooperation. *Science* 211:1390–1396.
3. Lehmann L, Keller L (2006) The evolution of cooperation and altruism—a general framework and a classification of models. *J Evol Biol* 19:1365–1376.
4. West SA, Griffin AS, Gardner A (2007) Social semantics: Altruism, cooperation, mutualism, strong reciprocity and group selection. *J Evol Biol* 20:415–432.
5. McNamara JM, Gasson CE, Houston AI (1999) Incorporating rules for responding into evolutionary games. *Nature* 401:368–371.
6. Wahl LM, Nowak MA (1999) The continuous prisoner's dilemma: I. linear reactive strategies. *J Theor Biol* 200:307–321.
7. Roughgarden J, Oishi M, Akçay E (2006) Reproductive social behavior: Cooperative games to replace sexual selection. *Science* 311:965–969.
8. McFarland DJ, Sibly RM (1975) The behavioural final common path. *Philos Trans R Soc London Ser B* 270:265–293.
9. McFarland D, Houston A (1981) *Quantitative Ethology: The State Space Approach* (Pitman Advanced Pub Program, Boston, MA).
10. Trivers RL (2002) *Natural Selection and Social Theory: Selected Papers of Robert L. Trivers* (Oxford Univ Press, New York).
11. Schultz W (2006) Behavioral theories and the neurophysiology of reward. *Annu Rev Psychol* 57:87–115.
12. Gintis H, Bowles S, Boyd R, Fehr E (2003) Explaining altruistic behavior in humans. *Evol Hum Behav* 24:153–172.
13. Burkart JM, Fehr E, Efferson C, van Schaik CP (2007) Other-regarding preferences in a non-human primate: Common marmosets provision food altruistically. *Proc Natl Acad Sci USA* 104:19762–19766.
14. de Waal FBM, Leimgruber K, Greenberg A (2008) Giving is self-rewarding for monkeys. *Proc Natl Acad Sci USA* 105:13685–13689.
15. Lakshminarayanan VR, Santos LR (2008) Capuchin monkeys are sensitive to others' welfare. *Curr Biol* 18:R999–R1000.
16. Gintis H (2007) A framework for the unification of the behavioral sciences. *Behav Brain Sci* 30:1–16.
17. Silk JB, et al. (2005) Chimpanzees are indifferent to the welfare of unrelated group members. *Nature* 437:1357–1359.
18. de Waal FBM (2000) Attitudinal reciprocity in food sharing among brown capuchin monkeys. *Anim Behav* 60:253–261.
19. Nash J (1950) The bargaining problem. *Econometrica* 18:155–162.
20. Maynard Smith J (1974) The theory of games and the evolution of animal conflicts. *J Theor Biol* 47:209–221.
21. Bishop DT, Cannings C (1976) Models of animal conflict. *Adv Appl Prob* 8:616–621.
22. Eshel I, Motro U (1981) Kin selection and strong evolutionary stability of mutual help. *Theor Popul Biol* 19:420–433.
23. Eshel I (1996) On the changing concept of evolutionary population stability as a reflection of a changing point of view in the quantitative theory of evolution. *J Math Biol* 34:485–510.
24. Christiansen FB, Loeschcke V (1980) Evolution and intraspecific exploitative competition 1. One-locus theory for small additive gene effects. *Theor Popul Biol* 18:297–313.
25. Queller DC (1985) Kinship, reciprocity and synergism in the evolution of social behavior. *Nature* 318:366–367.
26. Hauert C, Holmes M, Doebeli M (2006) Evolutionary games and population dynamics: Maintenance of cooperation in public goods games. *Proc R Soc London Ser B* 273:2565–2570.
27. Doebeli M, Hauert C, Killingback T (2004) The evolutionary origin of cooperators and defectors. *Science* 306:859–862.
28. 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.
29. Dekel E, Ely JC, Yilankaya O (2007) Evolution of preferences. *Rev Econ Stud* 74:685–704.
30. Güth W (1995) An evolutionary approach to explaining cooperative behavior by reciprocal incentives. *Int J Game Theory* 24:323–344.
31. Heifetz A, Shannon C, Spiegel Y (2007) The dynamic evolution of preferences. *Econ Theory* 32:251–286.
32. Heifetz A, Shannon C, Spiegel Y (2007) What to maximize if you must. *J Econ Theory* 133:31–57.
33. Trivers RL (1971) The evolution of reciprocal altruism. *Q Rev Biol* 46:35–57.
34. Fehr E, Camerer CF (2007) Social neuroeconomics: The neural circuitry of social preferences. *Trends Cog Sci* 11:419–427.
35. de Waal FBM (2008) Putting the altruism back into altruism: The evolution of empathy. *Annu Rev Psychol* 59:279–300.