Evolution of cooperation and skew under imperfect information

Erol Akçay[a,1], Adam Meirowitz[b], Kristopher W. Ramsay[b], and Simon A. Levin[a,1]

Departments of [a]Ecology and Evolutionary Biology and [b]Politics, Princeton University, Princeton, NJ 08544

Contributed by Simon A. Levin, July 27, 2012 (sent for review March 26, 2012)

The evolution of cooperation in nature and human societies depends crucially on how the benefits from cooperation are divided and whether individuals have complete information about their payoffs. We tackle these questions by adopting a methodology from economics called mechanism design. Focusing on reproductive skew as a case study, we show that full cooperation may not be achievable due to private information over individuals' outside options, regardless of the details of the specific biological or social interaction. Further, we consider how the structure of the interaction can evolve to promote the maximum amount of cooperation in the face of the informational constraints. Our results point to a distinct avenue for investigating how cooperation can evolve when the division of benefits is flexible and individuals have private information.

other-regarding preferences | social evolution | incentive compatibility | reproductive transactions | cheap-talk bargaining

Cooperative interactions drive much of the ecological, evolutionary, and social dynamics of organisms ranging from soil bacteria to primates, including—and especially—humans. Whereas much theory focuses on various mechanisms that promote cooperative behaviors (1–6), some fundamental questions remain unresolved. Among them is how the benefits of cooperation are to be divided among cooperating agents. Most theoretical work conceives of cooperation as a binary affair with payoffs to individuals from each outcome set a priori. However, frequently, the surplus from cooperation, whether it is the kill of a cooperatively hunting group or the reproductive output of a breeding group, can be partitioned among individuals in different ways; and how this division is achieved affects how likely individuals are to cooperate. Further, most research on biological cooperation focuses implicitly or explicitly on situations where individuals make decisions under perfect information of their and others' payoffs. However, private information, where some individuals have access to information and others do not, is a feature of many biological and social interactions. Although private information has been studied in a few specific contexts before, including mate choice (7, 8), parental care (9, 10), and animal conflicts (11), the role of private information in the evolution of cooperation in general remains understudied.

We introduce a distinct approach to biology to study how cooperation can be maintained when the division of benefits is flexible and individuals have private information. This approach, called mechanism design (12) and borrowed from economics, inverts the standard methodology of game-theoretic modeling. Instead of specifying a particular game and analyzing its equilibria, we analyze the properties of equilibrium outcomes in a large class of games and also ask what the consequences of different game structures are for the fitness of different individuals and the group's reproductive output.

As a case study, we use a problem of central importance to behavioral ecology and social evolution: the partitioning of reproduction, or reproductive skew, within a breeding group. A large body of work in behavioral ecology aims to understand the evolution of reproductive skew (13) as a function of demographic, individual, and ecological variables. However, patterns of reproductive skew remain contradictory: A recent review (14) concludes that whereas theory explains between-species patterns with some success, within-species patterns of skew often do not conform to theoretical predictions. We suggest that these failures occur because existing theory (14–16) assumes that reproductive skew evolves under perfect information about all relevant variables (17). In reality, however, individuals might be expected to have private information about themselves or the environment, which as we show dramatically affects both the scope of cooperation and the division of the benefits when cooperating. A related problem is that the proliferation of models in skew theory, driven in part by the empirical difficulties, has resulted in a situation where many contradictory patterns can be predicted, depending on the details of the model (14). Together with a systematic theory of which models apply in different settings, this could be a desirable property, but there is currently no such theory; hence the abundance of models fails to generate the clarity that theory is supposed to provide. Our approach avoids this problem by obtaining results independent of the precise game structure for a large class of games and also provides a first step in asking how the transactions game itself might evolve.

Our basic setup is a twist on the canonical reproductive skew model. Consider two individuals, labeled 1 and 2, who have the option of forming a group and breeding together or breeding alone. Label their expected success when breeding alone—their outside options—as o_1 and o_2, respectively. We depart from the canonical model in assuming that these options are not observed directly by both individuals; individual 1 only "knows," i.e., can condition its behavior on, o_1, but cannot condition on o_2, and vice versa. The outside options are distributed according to some probability distribution, and hence natural selection will lead to optimal strategies according to their expected fitness consequences. This assumption differs from that in previous models, where the outside options are commonly known; hence individual 1's strategy can be conditional on the lowest share 2 will accept, and vice versa. This is not possible in our setting and thus optimal strategies will typically miss some mutually beneficial opportunities for cooperation. If the individuals form the group, they can obtain a joint breeding success of Ω. This joint reproductive success is to be divided between 1 and 2 through some sort of game; our goal is to study which groups and divisions of reproduction are compatible with evolutionary stability of strategies in any kind of game and what game structures can implement the optimal outcome. For most of the following, it is more convenient to work with the potential losses and gains from group formation, defined as $g_1 = o_1$ (the reproduction individual 1 gives up by entering the group) and $g_2 = \Omega - o_2$ (the reproduction individual 2 could potentially gain by entering the group). We sometimes call g_1 and g_2 players' "types." Suppose that g_1 ($\in [a_1, b_1]$) and g_2 ($\in [a_2, b_2]$) are distributed according to $f_1(g_1)$ and $f_2(g_2)$, with cumulative distributions $F_1(\cdot)$ and $F_2(\cdot)$, respectively. We assume that these distributions are attributes of the environmental variation, i.e., do not change with the strategies of individuals or the game structure. Hence,

Author contributions: E.A., A.M., K.W.R., and S.A.L. designed research; E.A., A.M., and K.W.R. performed research; E.A., A.M., K.W.R., and S.A.L. contributed new reagents/analytic tools; and E.A., A.M., K.W.R., and S.A.L. wrote the paper.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. E-mail: eakcay@princeton.edu or slevin@princeton.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1212925109/-/DCSupplemental.

natural selection can optimize individual strategies with regard to their expected fitness under these distributions. (In the rationality-based language of economic game theory, these distributions would have been "common knowledge" in a Bayesian game; see *SI Text* for more on the correspondence between evolutionary stability and Bayesian Nash equilibria.) Finally, individuals are related to each other with a coefficient of relatedness, r, so that they gain indirect fitness benefits from each other's reproduction. The relatedness in a group is a function of demographic parameters (18), including the reproductive skew in the population (19); as a simplifying assumption, we take r as constant and to be the appropriately scaled relatedness that is consistent with the equilibrium level of skew in our model.

Overall, our setting is analogous to the bilateral trade problem with private information in economics (20), where the trade corresponds to formation of groups, and the payment corresponds to the allocation of the reproduction within the group. The important analytical difference is that previous work in economics assumes null relatedness, which as we show has important consequences. As such, our results for $r > 0$ also contribute to the economics literature.

If individuals' actions could be conditioned on both outside options o_1 and o_2, then it would be mutually beneficial for groups to form (with a division that makes both players better off) when $o_1 + o_2 < \Omega$, which translates to $g_2 > g_1$. With private information, however, individuals cannot compute this condition. They could potentially signal their outside options, but such signaling will not, in general, be evolutionarily stable. An individual with higher outside option can demand a higher share of the reproduction, and therefore individuals have an incentive to behave as if their outside options were higher. We show that this incentive problem can preclude cooperation even in cases where it is mutually beneficial, unless individuals are highly related to each other. Furthermore, the possibility of costly signaling cannot always remedy this problem.

We begin by specifying (*i*) an information structure for an interaction (i.e., which variables individual strategies can be conditioned on), (*ii*) a set of feasible games (i.e., mappings from combinations of actions to different payoffs), (*iii*) an equilibrium concept (e.g., evolutionary stability), and (*iv*) an objective function (e.g., maximizing the probability of cooperation). We obtain two kinds of results: First, we find the properties of evolutionarily stable outcomes in any possible game. Second, we describe a family of games that include the most salient transactional models and describe equilibrium play and its fitness consequences in this family of games as a first step to consider how the structure of the social interaction can evolve.

Our first kind of result is made possible by a celebrated theorem from economics, the revelation principle (21), which allows us to focus on a special class of simple games—direct mechanisms—that can represent all possible equilibria in all possible games. A direct mechanism is the simplest possible game structure, in which individuals simply send a message reporting their information to an (imagined) central arbiter and get assigned a payoff on the basis of the messages received by the arbiter. (A "mechanism" simply refers to a game or, more precisely, a mapping from combinations of strategies to the vector of payoffs to the players.) The function that determines the players' payoffs (assumed to be known to the players before sending their messages) determines whether players will find it optimal to report their information truthfully instead of misrepresenting it. If truth telling is optimal, then the mechanism is called incentive compatible. The revelation principle (21) states that all Bayesian Nash equilibria [a necessary condition for evolutionarily stable strategies (*SI Text*)] to any game of imperfect information can be represented by incentive-compatible direct mechanisms. Thus, determining whether an outcome is possible in a direct mechanism tells us whether that outcome could ever be the result of equilibrium behavior in any evolutionary game. (See more on the revelation principle in *SI Text*.)

Denoting the players' reports of their outside options by θ_1 and θ_2, respectively, we characterize a direct mechanism with two functions of these reports. The first function, $p(\theta_1, \theta_2)$ denotes the probability of group formation [$p(\cdot,\cdot)$ could be binary or continuous], and the second function, $x(\theta_1, \theta_2)$ gives the share of reproduction allocated to individual 1 (and taken from individual 2, and hence $x(\cdot,\cdot)$ can be viewed as a "payment"). For this section we are interested in the most general class of games, so we do not necessarily require that the payment is made only when the group forms (in the section *Nondirect Mechanisms and Incomplete Control*, we investigate a class of games that do have this more realistic property). Hence, $x(\theta_1, \theta_2)$ is the expected payment to individual 1 when the reports are θ_1, θ_2, regardless of whether the group forms or not. Given a direct mechanism with functions (p, x), individual 1's expected change in inclusive fitness when it has a loss of g_1 and reports θ_1 is given by

$$W_1(g_1,\theta_1) = \int_{a_2}^{b_2} ((1-r)x(\theta_1,g_2) - (g_1 - rg_2)p(\theta_1,g_2))f_2(g_2)dg_2. \quad [1]$$

[It should be noted that inclusive fitness calculations may in general be dependent on the frequency of different genotypes in the population when there are nonadditive fitness effects (22). However, with weak selection or small-effect mutants, additivity can be approximately restored and inclusive fitness calculations become approximately accurate. Here, we make use of this approximation, as our interest is to compute the optimal conditional strategies given the strategic interaction and not the dynamics of a particular set of genotypes.] Similarly, individual 2's expected change in inclusive fitness is given by

$$W_2(g_2,\theta_2) = \int_{a_1}^{b_1} ((g_2 - rg_1)p(g_1,\theta_2) - (1-r)x(g_1,\theta_2))f_1(g_1)dg_1. \quad [2]$$

A mechanism is said to be incentive compatible (IC) if $W_1(g_1, g_1) \geq W_1(g_1, \theta_1)$ for all $\theta_1 \in [a_1, b_1]$ and $W_2(g_2, g_2) \geq W_2(g_2, \theta_2)$ for all $\theta_2 \in [a_2, b_2]$. Furthermore, we require that participation in the game (i.e., sending the reports and accepting the outcome of the arbiter) is voluntary: Individuals can opt out of the interaction altogether and breed alone (thus obtaining their outside option) if their expected gains from the interaction are negative. Hence, in addition to IC, we require from our mechanism that $W_1(g_1, g_1) \geq 0$ and $W_2(g_2, g_2) \geq 0$ for all g_1 and g_2. Note that this condition applies after individuals know their own outside option, but before they have learned their partner's, so we call it the interim participation constraint (IPC). In *SI Text*, we provide the necessary and sufficient conditions for a mechanism to be both IC and IPC.

Our first major result concerns whether any game exists that ensures cooperation in all cases that are mutually beneficial (i.e., whenever $g_2 > g_1$). The Myerson–Satterthwaite theorem answers this question in the negative: If there are any pairs of individuals for whom cooperation is not mutually beneficial (i.e., the distributions of g_1 and g_2 overlap), there are no mechanisms that guarantee that all groups that *are* mutually beneficial will form. However, as we show below, this result is potentially changed when individuals are related and maximize their inclusive fitness (alternatively, one can think of the game being played by two agents with other-regarding preferences where the level of other regard is parameterized by r). In particular, assuming that $b_1 \geq b_2$ (see *SI Text* for $b_2 > b_1$), full cooperation becomes possible when

$$r\int_{a_1}^{a_1+b_2-a_2} F_1(t)(1-F_2(t))\,dt \geq \int_{a_2}^{b_2} F_1(t)(1-F_2(t))\,dt. \quad [3]$$

Thus, with high enough relatedness between the individuals, a game exists that ensures the individuals will form a group whenever it is mutually beneficial, a condition we term full cooperation,

or efficiency. If relatedness is below that threshold, some cooperation will still take place, but some pairs of individuals that would be better off in a group will nevertheless not cooperate at equilibrium, because of the incentives to misrepresent their private information. Note that the level of relatedness needed to ensure full efficiency can be very high, even exceeding 1. In the special case where both 1 and 2's outside options come from the same distribution, the integrals on both sides of Eq. **3** are identical to each other, and full cooperation becomes possible only when $r = 1$, i.e., with clonal groups.

In biology the study of costly signaling has become a common approach to dealing with informational inefficiencies (7–9). In the context of mate choice and parent–offspring relations, costly signaling can indeed restore efficiency [e.g., choosing the optimal mate (8) or optimally provisioning for offspring (23–25)], so one might be tempted to claim that allowing individuals to signal their outside options before the interaction might allow them to circumvent the inefficiency due to private information. This intuition turns out to be only partially correct: When signal costs depend only on the message (θ_i) and not on individuals' true type (g_i) (this is the case in models of parent–offspring signaling), we can show (*SI Text*) that all outcomes that can be implemented in any game with potential for costly signaling can also be implemented in a game without costly signaling, along with some that cannot be implemented with costly signals. Hence, costly signaling where costs depend only on the signal value cannot remedy the inefficiencies identified above. The reason is that the impossibility of sustaining inefficiency does not result from an inability to incentivize truth telling per se. Rather, the reason is that the payments required for truth telling become so high that they make some individuals better off not participating in the game at all, violating the IPC. Signaling with costs only exacerbates this problem, because the signal costs (unlike payments) are wasted and do not supply an inclusive fitness benefit. The same applies to games where parties can invest in a costly competitive behavior to grab a portion of reproduction for themselves (e.g., compromise models) (16, 26).

When signals can depend on the true type of the individual (e.g., individuals with high type pay a relatively lower cost for the same signal) (8, 27), the effect of costly signaling depends on the exact nature of the cost function. It is easy to show functions always exist that allow truthful revelation at little or no cost in equilibrium (28), effectively removing the private information problem and restoring full efficiency. However, if the signaling mechanism inevitably leads to costs at the equilibrium path (as is commonly presumed), the participation constraint might again preclude full efficiency. More generally, mechanism design can be used to learn about the types of cost function that can implement efficient outcomes. However, any sharper prediction about whether efficient cooperation can be maintained would require careful accounting of biological constraints specific to the interaction. Likewise, if the costly actions of one individual actually provide a benefit to the other, it is easy to show that full efficiency can be restored provided that the benefits are high enough. However, such a solution arguably redefines the problem, because it implies that cooperative benefits are not all accounted for in the original problem statement.

Given that full cooperation may not be possible for some values of relatedness between individuals, we next consider the properties of the interaction that maximizes the aggregate expected gains to the individuals, i.e., the expected value of $W_1(g_1, g_1) + W_2(g_2, g_2)$. This question can be answered by maximizing the expected probability that groups form,

$$\bar{p} = \int_{a_2}^{b_2} \int_{a_1}^{b_1} p(g_2, g_1) f_2(g_2) f_1(g_1)\, dg_1 dg_2, \qquad [4]$$

subject to the constraints $g_2 \geq g_1$ and inequality Eq. **S19** in *SI Text*. We can show that the mechanism maximizing the gains from cooperation prescribes that groups should form whenever

$$g_2 - g_1 \geq \max\left(\frac{1}{1+r}\left[\frac{1-F_2(g_2)}{f_2(g_2)} + \frac{F_1(g_1)}{f_1(g_1)} - \frac{1}{\lambda}\right], 0\right), \qquad [5]$$

where $\lambda > 0$ is a Lagrange multiplier, which is determined by substituting the function Eq. **5** into the constraint Eq. **S19** and satisfying that constraint with equality. The solution Eq. **5** can be interpreted as follows: In the space of gains (Fig. 1), the *Upper Left* region (shaded) consists of cases where the group is efficient and forms in equilibrium, and the *Lower Right* region (below the dashed diagonal) consists of cases where the group is not efficient and does not form. The slice of g_1-g_2 pairs in between the shaded region and the dashed diagonal corresponds to cases where the group does not form even though it is mutually beneficial. The intuition for why groups do not form in this region is that each individual has relatively high outside options and therefore "demands" too high a share of the group production to consent to cooperation. The size of the slice between the shaded region and the diagonal in Fig. 1 captures the extent to which the incentive problem causes inefficiencies and is determined by the distributions and r.

An Example

To apply the general results shown above, consider the scenario where a dominant and a subordinate individual interact to determine their share of the group reproduction. Normalize the group's productivity to $\Omega = 1$ and assume that the outside options o_1 and o_2 are uniformly distributed over $[w_d, 1]$ and $[0, w_s]$, respectively. In other words, w_d is the lowest possible reproduction that a dominant can expect to achieve on its own, and w_s is the highest possible reproduction for a subordinate. To focus on a particular interesting set of cases we assume $w_s \leq 0.5 \leq w_d$ and $1 - w_s \leq w_d$. With these assumptions, the optimal mechanism maximizing the total benefit while being IC and IPC (i.e., the solution Eq. **5**) prescribes that groups will form when

$$g_2 - g_1 \geq \frac{(1-r)(1-w_d)}{2(2+r)}. \qquad [6]$$

Thus, unless the two individuals are clones of each other with $r = 1$, some combinations of dominants and subordinates will not be able to form a group even though there is a degree of skew that allows both individuals to benefit from the group. This result illustrates the essence of the inefficiency that is created by private information.

If groups do form, the share of the group reproduction going to the dominant (given by the payment function) is predicted to be $x(g_1, g_2) = 1/3g_1 + 1/3g_2 + (1 + w_s)/6$. Thus, the dominant's share of the reproduction (and hence, the reproductive skew) increases with the outside option of the dominant and decreases with the outside option of the subordinate (as $g_2 \equiv \Omega - o_2$). Among all of the groups that form, the mean payment (i.e., skew) is predicted to be $\bar{x} = (1 + w_d)/2$, which interestingly does not depend on the relatedness between individuals. Similarly, we can compute the variance of skew among all of the groups that form and find $\mathrm{var}(x) = (1 - w_d)^2(1 + r)^2/(24(2 + r)^2)$, which is increasing in r, but decreasing in w_d.

Nondirect Mechanisms and Incomplete Control

The above analysis tells us what outcomes can be evolutionarily stable in any possible game and characterizes the equilibrium properties of the game structures that maximize the total benefit (a similar analysis can be done for any other objective function). It does not, however, answer the questions of how such an outcome might be implemented in an actual game that does not involve the imaginary arbiter or whether natural selection (or cultural evolution or learning) will actually lead to such a game structure. The latter problem is very much understudied; to our knowledge, very few studies have dealt with how the payoff

structure of a social interaction might evolve (29, 30). This section focuses on a family of simple negotiation games and shows that the optimal mechanism with uniform distributions can be implemented by a member of this family. Although our analysis does not provide a complete answer to the question of how the game might evolve, it suggests possible ways this question can be addressed.

In the context of reproductive transactions, a focal and contentious question about game structures has been who controls the division of the reproduction (16). In the so-called concession models, the dominant concedes to the subordinate the minimum reproduction that is required for the subordinate to prefer being in the group, whereas in the restraint model, the roles are reversed. Between these two extremes, both individuals would have a say, with the final division somewhere in between the two individuals' offers, termed incomplete control by both individuals (31).

Here, we present a model that combines the concession and restraint models and extends them to situations with private information. The basic informational environment remains as above with uniform distributions, but instead of focusing on direct mechanisms, we now consider the following class of games: Individuals simultaneously declare offers $\theta_i(g_i)$; θ_1 is the minimum amount of reproduction that individual 1 (the dominant) requests to assent to group formation, whereas θ_2 is the maximum amount of reproduction individual 2 (the subordinate) is willing to "pay" to the dominant. If $\theta_2 \geq \theta_1$, the group forms, and the dominant gets a share of reproduction $k\theta_2 + (1-k)\theta_1$, where k is between 0 and 1. Each value of k defines a particular game: With $k = 0$, the payment to 1 is its own offer, which corresponds to a restraint model (because the dominant is getting the minimum it requires). Similarly, $k = 1$ corresponds to a concession model, and $0 < k < 1$ to cases where neither side has complete control over the division. This setup is similar to two-person bargaining under incomplete information (32), again with the difference of nonzero relatedness that creates interdependent preferences. Note that the offer signals θ_i are not costly; hence this is a model of "cheap-talk" bargaining where individuals are free to "bluff" if they choose to.

First, we consider the equilibrium of the game for a given k. In particular, we are interested in separating equilibria, where individuals' offers θ_i are continuous and increasing functions of their private information g_i. Using the first-order conditions for the optimal offer strategies, we arrive at a set of coupled differential equations,

$$\begin{aligned}&\left(\theta_1^{-1}(\theta_2(g_2)) - rg_2 - (1-r)\,\theta_2(g_2)\right) f_2(g_2)\\ &\quad = -(1-r)(1-k)(1-F_2(g_2)\,\theta_2'(g_2))\end{aligned} \qquad [7]$$

$$\left(\theta_2^{-1}(\theta_1(g_1)) - rg_1 - (1-r)\,\theta_1(g_1)\right) f_1(g_1) = (1-r)\,kF_1(g_1)\,\theta_1'(g_1), \qquad [8]$$

where $\theta_i^{-1(\cdot)}$ denotes the inverse of the offer function of individual i. To illustrate what these equations entail, we assume g_i are uniformly distributed as in the previous example. In that case, the equilibrium offer strategies θ_i are linear in g_i, with the slope and intercept being functions of w_d, k, and r (see *SI Text* for full expressions). Regardless of k and w_d, for $r < 1$, the slopes of both offer functions are less than 1, and the intercepts are nonnegative. Fig. 2 illustrates the general nature of the offer functions. Importantly, each individual i will "shade" or "mark up" its offer (i.e., bid below or above g_i, respectively), with the level and direction of this shading dependent on the relatedness.

The expected total gain in group output from cooperation is then given by

$$\frac{(1-w_d)^2(2-k(1-r)+r)(1+k+r(2-k))}{2w_s(2+r)^3}. \qquad [9]$$

Likewise, the expected inclusive fitness gains for a dominant and a subordinate are given by

$$\overline{W_d} = \frac{(1-w_d)^2(2-k(1-r)+r)^2(1+k+r(2-k))}{6w_s(2+r)^3} \qquad [10]$$

$$\overline{W_s} = \frac{(1-w_d)^2(2-k(1-r)+r)(1+k+r(2-k))^2}{6w_s(2+r)^3}, \qquad [11]$$

respectively. Eqs. **10** and **11** imply that the expected inclusive fitness of the dominant is decreasing in k, whereas the expected inclusive fitness of the subordinate is increasing. On the other hand, Eq. **9** implies that $k = ½$; i.e., a "split-the-difference" rule maximizes the total expected gain from cooperation. Such a rule corresponds to an incomplete control model where neither individual is able to impose his/her preferred division (θ_i) on the other. Moreover, equilibrium behavior under the splitting-the-difference rule implies that groups will form when

$$g_2 - g_1 > \frac{(1-r)(1+w_d)}{2(2+r)}, \qquad [12]$$

which is the exact same condition as in Eq. **6**. In other words, the evolutionarily stable strategies in the split-the-difference game yield the maximum gains from cooperation among all possible games. Fig. 1 depicts the region of g_1 and g_2 where groups are predicted to form and not form and the predicted share of the dominant in the groups that form. Dominants and subordinates with high outside options (higher g_1 and lower g_2, respectively) are

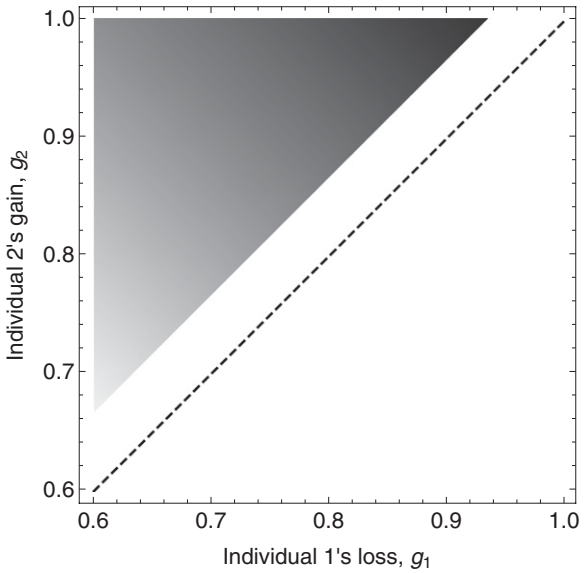

Fig. 1. The predicted outcome in the bargaining mechanism with k = 1/2 see as a function of g1 and g2. The shaded area denotes the region in the $g1$-$g2$ space where groups are predicted to form. The dashed diagonal separates the regions where groups are mutually beneficial (above the diagonal) and not (below the diagonal); the region between the diagonal and the shaded region represents groups that are mutually beneficial but cannot form at equilibrium. In the region where the group is predicted to form, darker shading indicates a higher share of reproduction for the dominant, i.e., higher skew. Parameters are w_d = 0.6, w_s = 0.5, r = 0.25.

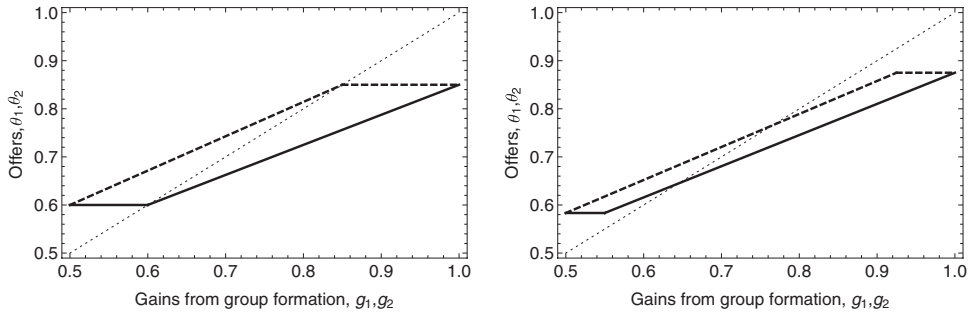

Fig. 2. The offer functions $\theta_1(g_1)$ (dashed lines) and $\theta_2(g_2)$ (solid lines). The dotted line is the 45° diagonal, corresponding to the offers being equal to the gains g_i (i.e., truthful revelation). (*Left*) $r = 0$; (*Right*) $r = 0.4$. For low values of g_2, the subordinate cannot make an offer that the dominant will accept, and hence starts out with a constant offer strategy that is given by the offer that is accepted by the dominant with the lowest g_1. Likewise, for high values of g_1, no subordinate can afford to pay enough to the dominant, and hence dominants with high enough g_1 also follow a constant offer strategy. Note that when $r = 0$, no individual ever offers below its gain or loss from group formation, whereas with positive r, such offers can be part of the equilibrium strategy when either individual has high outside options (corresponding to high g_1 and low g_2), due to the inclusive fitness effect. Parameters for both panels: $w_s = w_d = 0.5$, $k = 0.6$.

predicted to get relatively higher shares of reproduction compared with the same role individuals with lower outside options. When one looks at the mean personal fitness of dominants and subordinates, however, one can see that for a large region of the parameter space, solitary individuals of either role do better than their counterparts within groups (Fig. 3). This result is due to a self-selection effect: Only individuals with relatively high outside options are predicted to remain solitary at equilibrium. Hence, even though group formation confers a net benefit to those individuals that form groups, those who choose to remain solitary are still better off on the average. This simple selection effect might explain findings where the solitary fitness of individuals is higher than the share of reproduction they are getting in the group (14, 33).

Although the split-the-difference rule maximizes the total gains from group formation, this fact does not automatically mean that this bargaining game will be the end result of evolution, even if the social system is constrained to the family of games parameterized by k. A full treatment of how evolution can change the mechanisms is beyond the scope of the current paper; however, we can note a few basic predictions from our results. In particular, Eqs. **10** and **11** imply that alleles that alter the social game to favor the subordinate's offer (i.e., higher k) will be selected for in subordinates (individuals 2) and alleles that effectively lower k will be favored in dominants (individuals 1). Where the exact balance will be depends on the demography of the species, including the relatedness between individuals, which itself is a function of the skew resulting from the social game (19). These factors can lead to complex feedbacks between the evolution of the social game and the demographic properties of a species; elucidating these feedbacks remains an open question. Another important caveat here is that Eqs. **10** and **11** assume that each individual is playing an optimal strategy given the bargaining rule k. Hence, any statement based on these fitness functions is predicated on the bargaining rule evolving much more slowly than individuals' strategies in a given social structure (or alternatively, that optimal strategies are learned behaviorally) (30). Relaxing this assumption and allowing individual's strategies to be "mismatched" to the game they are playing is likely to change evolutionary dynamics substantially.

Conclusions

For reproductive skew theory, our analyses show that the addition of private information to reproductive transactions theory can change the predictions from the models significantly. Our main result highlights a previously unrecognized constraint to the evolution of cooperation: When the distributions of individuals' gains overlap, so that there is uncertainty over whether cooperation would be mutually beneficial, it may not be possible to guarantee cooperation in all cases where it is in fact mutually beneficial. Rather than being an idiosyncrasy of a particular game setup, this result holds in any evolutionarily stable equilibrium of any game. In a sense, the possibility that cooperation might not be beneficial to both "poisons the well" and as a result, some mutually beneficial cases are forsaken. Moreover, the nature of the inefficiency is one-sided. Groups that should form do not, but groups that should not form will not mistakenly form in the optimal equilibria.

Empirically, our model draws the distinction between patterns at the between-species (represented by species means) and within-species levels (represented by distributions within populations).

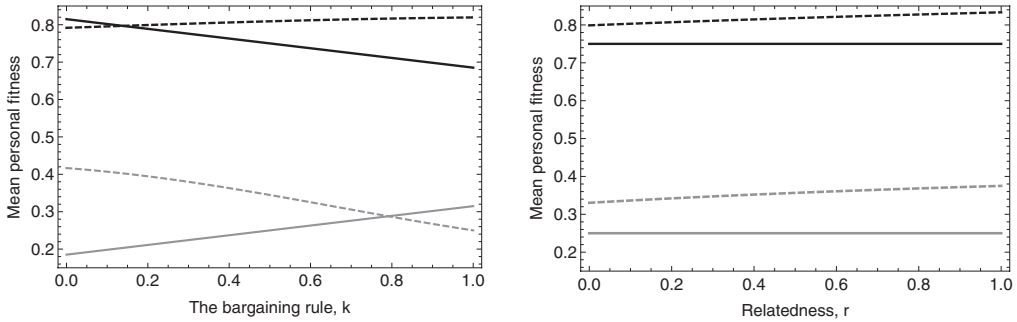

Fig. 3. The mean fitness of individuals who form a group vs. stay solitary under the cheap-talk bargaining game. The solid curves are for mean fitness within group, and dashed curves are for solitary individuals; black represents the dominants and gray the subordinates. Parameters are $w_d = w_s = 0.5$, $r = 0.25$ (*Left*), and $k = 0.5$ (*Right*).

An appreciation of this distinction might help explain why extant theory does well at explaining the former, but not the latter. However, we also see that predictions at both levels can depend on how individuals' outside options are distributed. A cautionary note here is that the estimation of these distributions can be tricky unless important selection effects are addressed. The naturally observed distribution of outside options will not represent the underlying distribution, but will be biased toward those with high outside options. This fact might resolve the apparent paradox that in some species, the within-group fitnesses of individuals appear too low to explain group stability (33). A recent resolution to this paradox has been to account for benefits later in life of individuals (34, 35); our model suggests that such benefits might not be needed to explain this pattern.

One of the prominent criticisms of reproductive skew theory is that the multitude of models of reproductive skew makes it possible to make dramatically different predictions depending on the details of the model, without an overarching theory of which model should be used in each case. In particular, pure transactional models (e.g., ref. 15) allow a range of outcomes that are consistent with group stability, with the precise level of skew being dependent on who is assumed to be in control. In contrast, compromise models (26) do predict a unique skew, but those predictions are critically sensitive to the details of the model (such as the functional shape of the reproductive share resulting from competitive effort) (14). Our approach avoids both of these issues and allows us to derive general results that are independent of the structure of the game and—given the distribution of outside options—to provide sharp predictions for the expected level of skew. Moreover, considering a specific class of games that include the concession and restraint models, we show that for uniform distribution of the outside options, the highest amount of cooperation is maintained by a mechanism where both the dominant and the subordinate have partial control over the division of reproduction. Our results thus represent a first step in elucidating the larger question of how natural selection shapes the structure of the social interaction leading to the division of reproduction.

For the evolution of social behavior and cooperation in general, our study illustrates how the methods of mechanism design can be used to study evolution of both individual behaviors and the social interaction under uncertainty and private information. Here, we studied the effects of private information about outside options; different kinds of uncertainty (such as over fighting ability) will be important in different contexts, such as parental care (9) and agonistic interactions (11). We believe that mechanism design will be a powerful tool in the evolutionary biologists' toolkit because it can be used to obtain results about evolutionary stability that are independent of particular assumptions about game structures. The general approach of mechanism design also provides a framework for studying how the social structure of a species—in addition to individuals' strategies in a given social game—evolves. Addressing this question fully will require extending existing mechanism design theory to a dynamic evolutionary setting.

Finally, our results also have some significance for economics, where our analysis can be interpreted as applying to bargaining over trades between individuals with other-regarding preferences (36, 37). Ever since Myerson and Satterthwaite (20), economists have accepted that private information leads to unavoidable inefficiencies. Our results show that other regard mitigates and, in some cases, can entirely counteract the inefficiencies in trade. One implication of this result is that in human history, groups with other-regarding agents would have an easier time taking advantage of the opportunities offered by trade. This effect would produce another route for the evolution of other regard, as such groups would have an advantage in (cultural or genetic) between-group selection. Hence, our results suggest that other-regarding preferences might have facilitated the emergence of, and coevolved with, trade and economic activity in human history.

ACKNOWLEDGMENTS. We thank seminar participants at Princeton University and University of California at Irvine for discussion and comments on the work and Christina Riehl, Jeremy Van Cleve, Dustin Rubenstein, and two reviewers for valuable feedback on the manuscript. This study was supported by National Science Foundation Grant EF-1137894.

1. Trivers RL (1971) The evolution of reciprocal altruism. *Q Rev Biol* 46:35–57.
2. Axelrod R, Hamilton WD (1981) The evolution of cooperation. *Science* 211: 1390–1396.
3. Sachs JL, Mueller UG, Wilcox TP, Bull JJ (2004) The evolution of cooperation. *Q Rev Biol* 79:135–160.
4. 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.
5. Levin SA (2009) *Games, Groups, and the Global Good (Springer Series in Game Theory)*, ed Levin SA (Springer, Berlin), pp 143–153.
6. Levin SA (2010) Crossing scales, crossing disciplines: Collective motion and collective action in the Global Commons. *Philos Trans R Soc B* 365:13–18.
7. Zahavi A (1975) Mate selection: A selection for a handicap. *J Theor Biol* 53: 205–214.
8. Grafen A (1990) Biological signals as handicaps. *J Theor Biol* 144:517–546.
9. Godfray HCJ (1991) Signalling of need by offspring to their parents. *Nature* 352: 328–330.
10. McNamara JM, Gasson C, Houston AI (1999) Incorporating rules for responding into evolutionary games. *Nature* 401:368–371.
11. Parker GA, Rubenstein DI (1981) Role assessment, reserve strategy and acquisition of information in asymmetric animal conflicts. *Anim Behav* 29:221–240.
12. McCarty NA, Meirowitz A (2007) *Political Game Theory: An Introduction* (Cambridge Univ Press, Cambridge, UK).
13. Vehrencamp SL (1983) Optimal degree of skew in cooperative societies. *Am Zool* 23: 327–335.
14. Nonacs P, Hager R (2011) The past, present and future of reproductive skew theory and experiments. *Biol Rev Camb Philos Soc* 86:271–298.
15. Reeve HK, Ratnieks FLW (1993) *Queen Number and Sociality in Insects*, ed Keller L (Oxford Univ Press, Oxford), pp 45–85.
16. Johnstone RA (2000) Models of reproductive skew: A review and synthesis. *Ethology* 106:5–26.
17. Kokko H (2003) Are reproductive skew models evolutionarily stable? *Proc R Soc Lond B Biol Sci* 270:265–270.
18. Frank S (1998) *Foundations of Social Evolution* (Princeton Univ Press, Princeton).
19. Johnstone RA (2008) Kin selection, local competition, and reproductive skew. *Evolution* 62:2592–2599.
20. Myerson R, Satterthwaite M (1983) Efficient mechanisms for bilateral trading. *J Econ Theory* 29:265–281.
21. Myerson R (1979) Incentive compatibility and the bargaining problem. *Econometrica* 47:61–74.
22. Queller DC (1992) Quantitative genetics, inclusive fitness, and group selection. *Am Nat* 139:540–558.
23. Nöldeke G, Samuelson L (1999) How costly is the honest signaling of need? *J Theor Biol* 197:527–539.
24. Roughgarden J, Song Z (2013) *Human Nature, Early Experience and the Environment of Evolutionary Adaptedness*, eds Narvaez D, Panksepp J, Schore A, Gleason T (Oxford Univ Press, New York).
25. Akçay E (2012) Incentives in the family ii: Behavioral dynamics and the evolution of non-costly signaling. *J Theor Biol* 294:9–18.
26. Reeve HK, Emlen ST, Keller L (1998) Reproductive sharing in animal societies: Reproductive incentives or incomplete control by dominant breeders? *Behav Ecol* 9: 267–278.
27. Spence M (1973) Job market signaling. *Q J Econ* 87:355–374.
28. Bergstrom CT, Lachmann M (1998) Signaling among relatives. III. Talk is cheap. *Proc Natl Acad Sci USA* 95:5100–5105.
29. Worden L, Levin SA (2007) Evolutionary escape from the prisoner's dilemma. *J Theor Biol* 245:411–422.
30. Akçay E, Roughgarden J (2011) The evolution of payoff matrices: Providing incentives to cooperate. *Proc Biol Sci* 278:2198–2206.
31. Clutton-Brock TH (1998) Reproductive skew, concessions and limited control. *Trends Ecol Evol* 13:288–292.
32. Chatterjee K, Samuelson W (1983) Bargaining under incomplete information. *Oper Res* 31:835–851.
33. Nonacs P, Liebert A, Starks P (2006) Transactional skew and assured fitness return models fail to predict patterns of cooperation in wasps. *Am Nat* 167:467–480.
34. Field J, Cronin A, Bridge C (2006) Future fitness and helping in social queues. *Nature* 441:214–217.
35. Leadbeater E, Carruthers J, Green J, Rosser N, Field J (2011) Nest inheritance is the missing source of direct fitness in a primitively eusocial insect. *Science* 333: 874–876.
36. Fehr E, Gächter S (2000) Fairness and retaliation: The economics of reciprocity. *J Econ Perspect* 14(3):159–181.
37. Sobel J (2005) Interdependent preferences and reciprocity. *J Econ Lit* 43:392–436.

Supporting Information

Akçay et al. 10.1073/pnas.1212925109

SI Text

Correspondence Between Bayesian Nash Equilibria and Evolutionary Stability. We first clarify the relationship between evolutionarily stable strategies and the appropriate version of Nash equilibria for games in which agents possess private information. Whereas a normal-form game is characterized by a set of players N, a strategy set S_i for each player $i \in N$, and a profile of utility functions $u_i : S \to R^1$, where $S = \prod_{i \in N} S_i$ is the space of strategy profiles, a Bayesian normal-form game includes a few extras. Each agent is assumed to possess a type, γ_i drawn from the set Γ_i. The types are assumed to have a joint distribution, $F(\cdot)$ on support $\Gamma = \prod_{i \in N} \Gamma_i$. A player now can condition their strategy choice on their type, so we think of strategies as mappings, $\phi_i : \Gamma_i \to S_i$. We let $\phi_i(\cdot)$, or even ϕ_i when it is clear, denote such a strategy and apply the same conventions to write ϕ_{-i} for the mapping describing the functions for all players other than i. In addition, the utility functions are now also dependent on the types so that player i obtains payoff, $u_i(s, \gamma)$ with $s \in S$ and $\gamma \in \Gamma$. Accordingly, optimality of a strategy $\phi_i(\gamma_i)$ for player i with type γ_i when the other players play strategies ϕ_{-i} is assessed by evaluating the expected utility $\int u_i(s_i, \phi_{-i}(\gamma_{-i}); \gamma_i, \gamma_{-i}) dF(\gamma_{-i} \mid \gamma_i)$. In a Bayesian game, each player selecta a function ϕ_i, and the appropriate equilibrium concept, Bayesian Nash equilibrium requires that these functions are mutual best responses (alternatively put, that ϕ calls for mutual best responses for each player possessing each possible type in their support). Our subsequent analysis will draw on results about the existence of Bayesian Nash equilibria satisfying certain conditions. These results are directly applicable to biological contexts because, as in the case of standard normal-form games, ESS is stronger than Bayesian Nash. The extension of ESS to Bayesian games is natural. Here instead of an evolutionary dynamic operating on strategy choices, $s_i \in S_i$ and producing an equilibrium (or stable) joint distribution over $s \in S$, we think of an evolutionary dynamic operating on the space of functions mapping from Γ_i into S_i. Let $\Gamma_i^{S_i}$ denote this set. Let s_{-i}, γ_{-i}, and ϕ_{-i} denote the vectors for players $N \backslash i$, so that $\gamma_{-i} = (\gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_n)$, for example. Then an ESS is simply a profile of functions, $\phi_i \in \Gamma_i^{S_i}$ that is ESS in a normal-form game in which the strategy space is $\Gamma_i^{S_i}$ and the payoff functions are given by $\int u_i(\phi(\gamma_i), \phi_{-i}(\gamma_{-i}); \gamma_i, \gamma_{-i}) dF(\gamma_{-i} \mid \gamma_i)$. Armed with these definitions we can state the relevant equivalence between ESS in evolutionary games with private information and Bayesian Nash equilibria in static Bayesian games (1).

Theorem 1. *Every ESS in an evolutionary game with private information is a Bayesian Nash equilibrium in the corresponding static game. Every strict Bayesian Nash in the static game corresponds to an ESS in the evolutionary game.*

Revelation Principle. The revelation principle relies on a deceptively simple idea. As above, note that in a Bayesian game, the strategy of an individual i is a mapping from the type space to the strategy space $\phi_i(\gamma_i)$: $\Gamma_i \to S_i$. Suppose that the N-tuple of strategies $\phi^* = (\phi_1^*, \phi_2^*, \cdots, \phi_N^*)$ constitutes a Bayesian Nash equilibrium of the game, and the outcome (i.e., the vector of payoffs to players with this Bayesian Nash equilibrium) is given by $u^*(\phi^*)$.

Now we show that any Bayesian Nash equilibrium in the original game can be implemented by a direct mechanism. A direct mechanism is simply a game where $S = \Gamma$; i.e., the strategy space for the players is to declare their types and the payoff is assigned (say, by an imagined arbiter) to the players as a function of the declarations. Consider the direct mechanism where the payoffs to individuals as a function of the players' declarations θ_i are given by $u^*(\phi_1^*(\theta_1), \phi_2^*(\theta_2), \cdots, \phi_N^*(\theta_N))$. In other words, the direct mechanism calculates the payoffs as if all players i played the equilibrium actions of a player of type θ_i in the original Bayesian game. In such a direct mechanism, truthful revelation, i.e., $\theta_i = \gamma_i$ is a Bayesian Nash equilibrium. In other words, such a direct mechanism is incentive compatible. To show this, suppose that all players except i reveal truthfully and that player i sends a different message $\theta_i \neq \gamma_i$. Then the expected payoff to individual i is $\bar{u}_i^* = u_i(\phi_i^*(\theta_i), \phi_{-i}^*(\gamma_{-i}))$. However, due to the strategy profile ϕ^* being a Bayesian Nash equilibrium, we know that $u_i(\phi_i^*(\gamma_i), \phi_{-i}^*(\gamma_{-i})) \geq u_i(s_i, \phi_{-i}^*(\gamma_{-i}))$, for any strategy s_i. Hence, individual i cannot gain by deviating from truthful revelation under this direct mechanism. Therefore, any Bayesian Nash equilibrium to any game of imperfect information can be implemented with an incentive-compatible direct mechanism.

The intuition behind this result can be illustrated by imagining the following thought experiment. Suppose that instead of playing the original Bayesian game themselves, each player is given a "robot" that is programmed to play the equilibrium strategy for that player for any contingency of the game, conditional on what the robot thinks the type of the player is. The robots then play the original game against each other and, at the end of the game, return to their respective "owner" with the payoffs they earned. The only information the robots need to carry out their mission is the type of their owner. As far as the players are concerned, this is a direct mechanism because players simply input their types to their robot and receive the payoffs collected by their mechanical servants. However, in such a scenario, no player would have any incentive to lie to their robot, as that would only make the robot play a suboptimal strategy and bring home a lower payoff (because we assumed that the robots are playing the equilibrium strategies conditional on type). Thus, for any Bayesian Nash equilibrium of the original game, an incentive-compatible direct mechanism exists that results in the exact same payoff distribution.

Obviously, this robot story is science fiction, and direct mechanisms are theoretical constructs that—for many reasons—we might not expect to find in real life. However, they provide a powerful tool to analyze properties of equilibria in a great multitude of games. This is because of a corollary of the revelation principle: If we can prove that no direct mechanism exists that implements a given payoff distribution, then we have also proved that there cannot be any game (no matter how complicated) that implements said distribution. This corollary is what we use here to prove our impossibility result.

Group Formation (Trade) Between Related Individuals. In this section, we derive the necessary and sufficient conditions for a direct mechanism to be incentive compatible and satisfy the participation constraint when the individuals are related to each other. Our results thus extend the Myerson–Satterthwaite theorem (2) to other-regarding agents. Suppose that individual 1's outside option (what it loses when the group forms) is given g_1 and individual 2's gain is given by g_2. Then, 1's (potential, before any transfer of fitness) inclusive fitness loss from group formation is $\Delta w_1 = g_1 - rg_2$, whereas 2's gain is $\Delta w_2 = g_2 - rg_1$. The condition for efficient group formation is still $\Delta w_2 > \Delta w_1$, which reduces to $g_2 > g_1$. However, because g_1 is known only to individual 1 and g_2 to individual 2, this means that neither individual has complete knowledge of their loss or gain. Hence, when deciding whether to reveal their information or not, each individual has to take into account the expected values of their opponent's private information. Although

this complicates the picture in some interesting ways, the key insight of Myerson and Satterthwaite still obtains.

First, we prove a slightly modified version of theorem 1 of Myerson and Satterthwaite that imposes constraints on the payoffs (inclusive fitness) of the individuals that are of the extreme types in any incentive-compatible mechanism. The expected gain of the individuals given their private information under a mechanism with the allocation function $p(g_1, g_2)$ and transfer function $x(g_1, g_2)$ is defined as

$$W_1(g_1) = \int_{a_2}^{b_2} ((1-r)x(g_1,g_2) - \Delta w_1 p(g_1,g_2)) f(g_2) dg_2 = \int_{a_2}^{b_2} ((1-r)x(g_1,g_2) - (g_1 - rg_2)p(g_1,g_2)) f_2(g_2) dg_2 \quad \textbf{[S1]}$$

$$W_2(g_1) = \int_{a_1}^{b_1} ((g_2 - rg_1)p(g_1,g_2) - (1-r)x(g_1,g_2)) f_1(g_1) dg_1. \quad \textbf{[S2]}$$

Here, a_i and b_i denote the upper and lower limits of the support of the gains g_i, respectively. Using the definitions of $\bar{x}_i(g_i)$ and $\bar{p}_i(g_i)$ given in the main text, we can rewrite W_1 and W_2 as

$$W_1(g_1) = (1-r)\bar{x}_1(g_1) - g_1\bar{p}_1(g_1) + r\int_{a_2}^{b_2} g_2 p(g_1,g_2) f_2(g_2) dg_2 \quad \textbf{[S3]}$$

$$W_2(g_2) = g_2\bar{p}_2(g_2) - (1-r)\bar{x}_2(g_2) - r\int_{a_1}^{b_1} g_1 p(g_1,g_2) f_1(g_1) dg_1 \quad \textbf{[S4]}$$

We define $\bar{g}_2(g_1) = \int_{a_2}^{b_2} g_2 p(g_1,g_2) f_2(g_2) dg_2$ and $\bar{g}_1(g_2) = \int_{a_1}^{b_1} g_1 p(g_1,g_2) f_1(g_1) dg_1$. Now, the incentive-compatible (IC) constraint in a direct mechanism requires that for each individual, reporting their true information is always optimal. For individual 1, this means the following inequalities both must hold for any g_1', g_1:

$$W_1(g_1) \geq (1-r)\bar{x}_1(g_1') - g_1\bar{p}_1(g_1') + r\bar{g}_2(g_1') \quad \textbf{[S5]}$$

$$W_1(g_1') \geq (1-r)\bar{x}_1(g_1) - g_1'\bar{p}_1(g_1) + r\bar{g}_2(g_1). \quad \textbf{[S6]}$$

This in turn implies

$$\bar{p}_1(g_1)(g_1' - g_1) \geq W_1(g_1) - W_1(g_1') \geq \bar{p}_1(g_1')(g_1' - g_1). \quad \textbf{[S7]}$$

Dividing by $g_1' - g_1 > 0$, and taking the limit as $g_1' \to g_1$, we have

$$\frac{dW_1}{dg_1} = -\bar{p}_1(g_1), \quad \textbf{[S8]}$$

and

$$W_1(g_1) = W_1(b_1) + \int_{g_1}^{b_1} \bar{p}_1(t_1) dt_1. \quad \textbf{[S9]}$$

Similarly, for individual 2, we have

$$W_2(g_2) = W_2(a_2) + \int_{a_2}^{g_2} \bar{p}_2(t_2) dt_2. \quad \textbf{[S10]}$$

Given this, we can write

$$\int_{a_2}^{b_2}\int_{a_1}^{b_1} (1+r)(g_2 - g_1) p(g_1,g_2) f_1(g_1) f_2(g_2) dg_1 dg_2, \quad \textbf{[S11]}$$

which can be expanded to

$$\int_{a_2}^{b_2}\int_{a_1}^{b_1} [(g_2 - rg_1)p(g_1,g_2) - (1-r)x(g_1,g_2)] f_1(g_1) f_2(g_2) dg_1 dg_2 + \int_{a_2}^{b_2}\int_{a_1}^{b_1} [(-g_1 + rg_2)p(g_1,g_2) + (1-r)x(g_1,g_2)] f_1(g_1) f_2(g_2) dg_1 dg_2. \quad \textbf{[S12]}$$

If we evaluate the integrals over g_1 and g_2 in the first and second terms, respectively, we obtain

$$\begin{aligned}\int_{a_2}^{b_2} W_2(g_2) f_2(g_2) dg_2 + \int_{a_1}^{b_1} W_1(g_1) f_1(g_1) dg_1 &= W_1(b_1) + \int_{a_1}^{b_1}\int_{g_1}^{b_1} \bar{p}_1(t_1) dt_1 f_1(g_1) dg_1 \\ &\quad + W_2(a_2) + \int_{a_2}^{b_2}\int_{g_2}^{b_2} \bar{p}_2(t_2) dt_2 f_2(g_2) dg_2 \\ &= W_1(b_1) + W_2(a_2) + \int_{a_1}^{b_1} \bar{p}_1(t_1) F_1(t_1) dt_1 \\ &\quad + \int_{a_2}^{b_2} \bar{p}_2(t_2)(1 - F_2(t_2)) dt_2, \end{aligned} \quad \textbf{[S13]}$$

where the last step follows from a change of the order of integration. Because the integration variables t_1 and t_2 are dummy variables, we can exchange them with g_1 and g_2, respectively, and equate the last line of Eq. **S13** with Eq. **S11** and obtain

$$\begin{aligned} W_1(b_1) + W_2(a_2) &= \int_{a_2}^{b_2}\int_{a_1}^{b_1} \left[(1+r)g_2 - \frac{(1-F_2(g_2))}{f_2(g_2)} - \left((1+r)g_1 + \frac{F_1(g_1)}{f_1(g_1)}\right)\right] \\ &\quad \times p(g_1,g_2) f_1(g_1) f_2(g_2) dg_1 dg_2. \end{aligned} \quad \textbf{[S14]}$$

Furthermore, the ex-interim participation constraint implies

$$W_1(b_1) + W_2(a_2) \geq 0. \quad \textbf{[S15]}$$

To finish the proof of our version of theorem 1 of Myerson and Sattertwaite, we need to find at least one transfer function $x(g_1, g_2)$ that satisfies incentive compatibility, assuming that the allocation function $p(g_1, g_2)$ satisfies Eq. **S15**. One such function can be found by modifying the one given by Myerson and Satterthwaite (their equation 6),

$$x(g_1,g_2) = \frac{1}{1-r}\left[\int_{t_2=a_2}^{g_2} t_2 d[\bar{p}_2(t_2)] - \int_{t_1=a_1}^{g_1} t_1 d[-\bar{p}_1(t_1)] + r(\bar{g}_1(g_2) - \bar{g}_2(g_1)) + \text{constant}\right], \quad \textbf{[S16]}$$

where the constant will be chosen to ensure that individual 1 with the highest loss g_1 or individual 2 with the lowest gain g_2 will receive zero payoff. To see that this transfer function makes the mechanism incentive compatible, we evaluate inequality Eq. **S5**:

$$\begin{aligned} W(g_1) - (1-r)\bar{x}_1(g_1') + g_1\bar{p}_1(g_1') - r\bar{g}_2(g_1') &= -\int_{t_1=a_1}^{g_1} t_1 d[-\bar{p}_1(t_1)] + \int_{t_1=a_1}^{g_1'} t_1 d[-\bar{p}_1(t_1)] - r\bar{g}_2(g_1) \\ &\quad + r\bar{g}_2(g_1') - g_1\bar{p}_1(g_1) + r\bar{g}_2(g_1) + g_1\bar{p}_1(g_1') - r\bar{g}_2(g_1'). \end{aligned} \quad \textbf{[S17]}$$

Making the appropriate cancellations on the right-hand side, we end up with

$$\int_{t_1=g_1}^{g_1'} t_1 d\left[-\bar{p}_1(t_1)\right] - \int_{t_1=g_1}^{g_1'} g_1 d\left[-\bar{p}_1(t_1)\right] = \int_{t_1=g_1}^{g_1'} (t_1 - g_1) d\left[-\bar{p}_1(t_1)\right] \geq 0, \quad \textbf{[S18]}$$

because $\bar{p}_1(\cdot)$ is a decreasing function and either $g_1' \geq g_1$ or the direction of integration is reversed. Hence the function in Eq. **S16** satisfies incentive compatibility for individual 1. The proof with individual 2 would proceed analogously.

Hence, we conclude that a mechanism (p, x) is IC and interim participation constrained (IPC) if and only if

$$0 \leq \int_{a_2}^{b_2}\int_{a_1}^{b_1}\left[(1+r)g_2 - \frac{(1-F_2(g_2))}{f_2(g_2)} - \left((1+r)g_1 + \frac{F_1(g_1)}{f_1(g_1)}\right)\right] \times p(g_1,g_2) f_1(g_1) f_2(g_2)\, dg_1\, dg_2, \quad \textbf{[S19]}$$

and the integrals $\bar{p}_1(\theta_1) = \int_{a_2}^{b_2} p(\theta_1, g_2) f_2(g_2) dg_2$, $\bar{p}_2(\theta_2) = \int_{a_1}^{b_1} p(g_1, \theta_2) f_1(g_1) dg_1$ are weakly decreasing and increasing, respectively.

Efficient group formation becomes possible with high enough r. Now, we can check whether the main result of Myerson and Satterthwaite also holds with other-regarding agents, i.e., whether there can be incentive-compatible, balanced-budget mechanisms that satisfy the participation constraint. It turns out that in contrast to the case of $r = 0$, which they consider, such mechanisms can exist, provided that r is large enough. To show this, we note that the ex-post–efficient mechanism prescribes group formation whenever $g_2 > g_1$ and not otherwise. Hence, the participation constraint reads

$$0 \leq \int_{a_2}^{b_2}\int_{a_1}^{\min(g_2,b_1)}\left[(1+r)g_2 - \frac{(1-F_2(g_2))}{f_2(g_2)} - \left((1+r)g_1 + \frac{F_1(g_1)}{f_1(g_1)}\right)\right] \times f_1(g_1) f_2(g_2)\, dg_1\, dg_2. \quad \textbf{[S20]}$$

Evaluating the inner integral first (and using integration by parts for the second term), we get

$$\int_{a_2}^{b_2} F_1(\min(g_2,b_1))\left[(1+r)g_2 - \frac{1-F_2(g_2)}{f_2(g_2)}\right] f_2(g_2) dg_2 - \int_{a_2}^{b_2}\left[(1+r)\min(g_2 F_2(g_2), b_1) + r\int_{a_1}^{\min(g_2,b_1)} F_1(g_1) dg_1\right] f_2(g_2) dg_2. \quad \textbf{[S21]}$$

Assuming that $b_2 > b_1$, and changing the order of integration on the remaining double integral, we have

$$-\int_{a_2}^{b_1} F_1(g_2)(1-F_2(g_2)) dg_2 - r\left[\int_{b_1}^{b_2}(F_2(v_2)-1) dv_2 + \int_{a_1}^{b_1+a_1-a_2} F_1(v_1)(F_2(b_1)-F_2(v_1)) dv_1 + \int_{a_1}^{b_1} F_1(g_1)(1-F_2(b_1)) dv_1\right]. \quad \textbf{[S22]}$$

Further simplifying and changing the integration variables, we arrive at

$$-\int_{a_2}^{b_1} F_1(t)(1-F_2(t)) dt + r\left[\int_{a_1}^{b_2} F_1(t)(1-F_2(t)) dt + \int_{b_1}^{b_1+a_1-a_2} F_1(t)[F_2(b_1)-F_2(t)] dt\right]. \quad \textbf{[S23]}$$

This expression reduces to the one derived by Myerson and Satterthwaite when $r = 0$, which provides a check for our result. For the two terms in the brackets, the first one is always positive, and the second one is nonpositive. (If $a_2 > a_1$, the lower limit of the integral is higher than the upper limit, and the integrand is positive; if $a_2 < a_1$, the integrand is negative.) Hence, the first term of expression **S23** is negative, whereas the second term can be positive or negative, and no general claim can be made here.

When $b_1 > b_2$, however, we have a simpler expression and a clearer result:

$$-\int_{a_2}^{b_2} F_1(t)(1-F_2(t)) dt + r\int_{a_1}^{a_1+b_2-a_2} F_1(t)(1-F_2(t)) dt. \quad \textbf{[S24]}$$

Again, the first term is negative, and the second term is positive. This means that with high enough r, the individual rationality constraint can be satisfied by an efficient mechanism. Note that in the special case where the individual's gains are distributed identically, the integrals in the first and second terms will be equal to each other, and hence efficient trade will be possible only when $r = 1$, as shown for a special case below.

Calculating the optimal mechanism. Now we ask which incentive-compatible mechanism satisfying the participation constraint maximizes the expected frequency of group formation, i.e.,

$$\bar{p} = \int_{a_1}^{b_1}\int_{a_2}^{b_2} p(g_1,g_2) f_1(g_1) f_2(g_2) dg_1 dg_2. \quad \textbf{[S25]}$$

Using the constraint from Eqs. **S19** and **S15** to construct the following Lagrangian,

$$\mathcal{L} = \int_{a_2}^{b_2}\int_{a_1}^{b_1}\left\{1 + \lambda\left[(1+r)g_2 - \frac{(1-F_2(g_2))}{f_2(g_2)} - \left((1+r)g_1 + \frac{F_1(g_1)}{f_1(g_1)}\right)\right]\right\} \times p(g_1,g_2) f_1(g_1) f_2(g_2)\, dg_1 dg_2, \quad \textbf{[S26]}$$

where λ is the Lagrange multiplier. Now, this integral is maximized when $p(g_1, g_2) = 1$ if the expression in the braces is positive and 0 otherwise. Thus, the optimal mechanism is given by

$$p(g_2,g_1) = \begin{cases} 1 & \text{when} \quad g_2 - g_1 \geq \dfrac{1}{1+r}\left[\dfrac{1-F_2(g_2)}{f_2(g_2)} + \dfrac{F_1(g_1)}{f_1(g_1)} - \dfrac{1}{\lambda}\right] \\ 0 & \text{otherwise,} \end{cases} \quad \textbf{[S27]}$$

where λ is determined by considering when the participation constraint becomes binding. To illustrate what the optimal mechanism prescribes, we consider the special case considered above, where the outside options O_i are both distributed uniformly on [0, Ω]. Hence, we have $a_1 = a_2 = 0$ and $b_1 = b_2 = \Omega$. In which case, the group will form only if $g_2 - g_1 \geq \frac{\lambda\Omega - 1}{(2+r)\lambda}$. To calculate λ, we evaluate the integral Eq. **S19** and find that $\lambda = \frac{2}{(1+r)\Omega}$, which yields the condition for trade as

$$g_2 - g_1 \geq \frac{\Omega(1-r)}{2(2+r)}. \quad \textbf{[S28]}$$

Note that the "gap" between g_2 and g_1, which represents the instances where trade would be efficient, but does not happen due to the incentive-compatibility constraint, becomes smaller as r grows and vanishes as $r \to 1$. Hence, the inefficiency in the system is a decreasing function of how highly related (or other regarding) the agents are.

More generally, for g_1 and g_2 distributed uniformly on $[a_1, b_1]$ and $[a_2, b_2]$, respectively, we have the following condition for trade in the optimal mechanism:

$$g_2 - g_1 > \max\left\{\frac{\lambda(b_2 - a_1) - 1}{\lambda(2 + r)}, 0\right\}. \quad \textbf{[S29]}$$

Costly signaling is never better than no signaling. Now we give both individuals the option of engaging in some costly behavior that might signal their private information. We are interested in whether costly signaling can improve the scope for efficient mechanisms, through eliminating the private information problem. However, it turns out that this is not the case. To begin, we write down the expected inclusive fitness $w_i^c(\cdot,\cdot)$ of each party that includes their signal costs under a mechanism, which now consists of the probability that the group forms $p(g_1, g_2)$, the transfer from individual 2 to individual 1, $x(g_1, g_2)$, and the signal costs to the two individuals, denoted by $y_1(g_1)$ and $y_2(g_2)$, respectively:

$$\begin{aligned} w_1^c(g_1, g_2) = &-p(g_1, g_2)g_1 + x(g_1, g_2) - y_1(g_1) \\ &+ r[p(g_1, g_2)g_2 - x(g_1, g_2) - y_2(g_2)] \end{aligned}$$

$$\begin{aligned} w_2^c(g_1, g_2) = &\, p(g_1, g_2)g_2 - x(g_1, g_2) - y_2(g_2) \\ &+ r[-p(g_1, g_2)g_1 + x(g_1, g_2) - y_1(g_1)]. \end{aligned}$$

Taking the expectation of $w_1^c(g_1, g_2)$ and $w_2^c(g_1, g_2)$ over g_2 and g_1, respectively, we write the expected inclusive fitness conditional on private information,

$$W_1^c(g_1) = -\bar{p}_1(g_1)g_1 + r\bar{g}_2(g_1) + (1 - r)\bar{x}_1(g_1) - y_1(g_1) - r\bar{y}_2$$

$$W_2^c(g_2) = \bar{p}_2(g_2)g_2 - r\bar{g}_1(g_2) - (1 - r)\bar{x}_2(g_2) - y_2(g_2) - r\bar{y}_1,$$

where $\bar{y}_i$ denotes the average signal costs paid by player i, which is not dependent on the other player's type. Incentive compatibility means that for all g_1, g_1', g_2, g_2' in the support of the individuals' types, the following must hold:

$$W_1^c(g_1) \geq -\bar{p}_1(g_1')g_1 + r\bar{g}_2(g_1') + (1 - r)\bar{x}_1(g_1') - y_1(g_1') - r\bar{y}_2$$

$$W_2^c(g_2) \geq \bar{p}_2(g_2')g_2 - r\bar{g}_1(g_2') - (1 - r)\bar{x}_2(g_2') - y_2(g_2') - r\bar{y}_1.$$

This implies, for $W_1^c(\cdot)$,

$$\bar{p}_1(g_1)(g_1' - g_1) \geq W_1^c(g_1) - W_1^c(g_1') \geq \bar{p}_1(g_1)(g_1' - g_1),$$

from which we obtain again

$$W_1^c(g_1) = W_1^c(b_1) + \int_{g_1}^{b_1} \bar{p}_1(t_1)dt_1. \quad \textbf{[S30]}$$

Similarly, we have

$$W_2^c(g_2) = W_2^c(a_2) + \int_{a_2}^{g_2} \bar{p}_2(t_2)dt_2. \quad \textbf{[S31]}$$

To obtain the individual rationality constraint under incentive compatibility, we start with the following integral:

$$\begin{aligned} \int_{a_2}^{b_2}\int_{a_1}^{b_1} &(1 + r)[(g_2 - g_1)p(g_1, g_2) - (y_1(g_1) + y_2(g_2))] \\ &\times f_1(g_1)f_2(g_2)dg_1dg_2. \end{aligned} \quad \textbf{[S32]}$$

Adding and subtracting $(1 - r)x(g_1, g_2)$, and using the definitions of $W_1(g_1)$ and $W_2(g_2)$, we can see that this double integral is equal to

$$\int_{a_2}^{b_2} W_2^c(g_2)f_2(g_2)dg_2 + \int_{a_1}^{b_1} W_1^c(g_1)f_1(g_1)dg_1. \quad \textbf{[S33]}$$

From this point we proceed in exactly the same way as above (the steps leading to Eq. **S13**) and obtain

$$\begin{aligned} W_1^c(b_1) + W_2^c(a_2) = \int_{a_2}^{b_2}\int_{a_1}^{b_1} \Bigg[\Bigg[&(1 + r)g_2 - \frac{(1 - F_2(g_2))}{f_2(g_2)} \\ &- \left((1 + r)g_1 + \frac{F_1(g_1)}{f_1(g_1)}\right)\Bigg]p(g_1, g_2) \\ &- (1 + r)(y_1(g_1) + y_2(g_2))\Bigg] f_1(g_1)f_2(g_2)dg_1dg_2. \end{aligned} \quad \textbf{[S34]}$$

Given that the $y_1(\cdot)$ and $y_2(\cdot)$ are both positive by assumption, this means that $W_1^c(b_1) + W_2^c(a_2)$ for any mechanism with costly signaling is strictly smaller than $W_1(b_1) + W_2(a_2)$ for a mechanism without costly signaling. Hence, for any allocation function $p(\cdot, \cdot)$, if the latter cannot satisfy the participation constraint, the former cannot satisfy it either. This result shows that costly signaling before negotiating cannot improve efficiency. An alternative approach is to allow for cost functions that depend on both the action and the type. This approach can lead to efficiency but leaves open many questions about how costs can depend on the agent's private information in precisely the required way.

Nondirect, Cheap-Talk Bargaining Mechanism. The bargaining game closely follows that of Chatterjee and Samuelson (3) and has the following structure: Each individual declares an offer $\theta_i(g_i)$; $\theta_1 g_1$ is the reproductive output (in absolute terms) individual 1 (the dominant) is willing to accept to forsake its outside option and form the group. Likewise, $\theta_2(g_2)$ is the reproductive output that individual 2 (the subordinate) is willing to concede to the dominant and still be willing to forsake its outside option to join the group. If $\theta_2(g_2) > \theta_1(g_1)$, the demands of the individuals are compatible, and the group forms. The actual division of reproduction depends on the bargaining rule, denoted by k: Individual 1's reproduction (the payment to it) is given by

$$x(g_1, g_2) = k\theta_2(g_2) + (1 - k)\theta_1(g_1). \quad \textbf{[S35]}$$

The bargaining rule k ($\in [0, 1]$) thus denotes whose offer the actual payment is closer to: $k = 0$ means that the payment is individual 1's declared "reservation price" (i.e., the minimum it is willing to accept to stay in the group), whereas $k = 1$ means that the payment is the maximum individual 2 is willing to pay. In the terminology of reproductive skew theory, these rules correspond to restraint and concession models, respectively. One interpretation of the bargaining rule k is that it quantifies the control each party has on the final outcome; the party with more control can in principle move k in the direction it prefers. We are then interested in how selection would act on k.

First, though, we need to calculate the equilibria of this game. As is common in games with uncertainty, there are many possible equilibria that can be supported by mutually compatible beliefs on both sides. However, we are interested in a class of equilibria where the offer functions $\theta_i(g_i)$ are continuous and increasing. We can find these equilibria by writing down the expected fitness of both parties conditional on their private information and their offers and finding the offer that maximizes this expected fitness. In particular, for an individual 1 with private information g_1 who sends a message θ_1, we have

$$\begin{aligned} W_1(g_1, \theta_1) = \int_{\theta_2^{-1}(\theta_1)}^{b_2} &[-g_1 + rg_2 + (1 - r)((1 - k)\theta_1 + k\theta_2(g_2))] \\ &\times f_2(g_2)dg_2, \end{aligned} \quad \textbf{[S36]}$$

where $\theta_2^{-1}(\cdot)$ is the inverse of the offer function of individual 2 (that is to be calculated from individual 2's optimization). Taking

the derivative of Eq. **S36** with respect to θ_1 and setting it equal to zero, we obtain

$$\frac{\partial W_1}{\partial \theta_1} = -\left(\theta_2^{-1}\right)'(\theta_1)\left[-g_1 + r\theta_2^{-1}(\theta_1) + (1-r)\theta_1\right] f_2\left(\theta_2^{-1}(\theta_1)\right) + (1-r)(1-k)\left(1 - F_2\left(\theta_2^{-1}(\theta_1)\right)\right) = 0. \quad \textbf{[S37]}$$

Now, we define a dummy variable $z_2 = \theta_2^{-1}(\theta_1)$ and note that $(\theta_2^{-1})' = \frac{1}{\theta_2'(\theta_2^{-1}(\theta_1))} = \frac{1}{\theta_2'(z_2)}$ and $g_1 = \theta_1^{-1}(\theta_2(x_2))$ to arrive at

$$\frac{\partial W_1}{\partial \theta_1} = -\frac{1}{\theta_2'(z_2)}\left[-\theta_1^{-1}(\theta_2(x_2)) + rz_2 + (1-r)\theta_2(z_2)\right] f_2(z_2) \quad \textbf{[S38]}$$

$$+(1-r)(1-k)(1-F_2(z_2))\theta_2'(z_2) = 0.$$

Multiplying the last expression with $\theta_2'(z_2)$ and exchanging the dummy variable z_2 for g_2 gives us Eq. **8** in the main text. The derivation of the second first-order (Eq. **9** of the main text) condition proceeds analogously. These conditions are identical to those of Chatterjee and Samuelson (3), except for the extra terms introduced by the fact that our agents are relatedness (or other regarding). They supply us with two coupled differential equations that can be integrated to find the offer functions. To find the boundary conditions for integrating these equations, consider a candidate pair of solutions $\tilde{\theta}_1(g_1)$ and $\tilde{\theta}_2(g_2)$ to Eqs. **8** and **9**, where $\tilde{\theta}_2(b_2) > \tilde{\theta}_1(g_1)$ for some $g_1 < b_1$. In other words, some individual 2s would be offering more than the maximum possible demand individual 1 will make. For any $k > 0$, such a strategy is strictly dominated by individuals 2 with high g_2 offering $\tilde{\theta}_1(b_1)$. Hence, in such a case, the boundary condition for individual 2's offer would be $\theta_2(g_2^t) = \tilde{\theta}_1(b_1)$, where $g_2^t = \tilde{\theta}_2^{-1}(\tilde{\theta}_1(b_1))$. Boundary conditions in the other cases can be found by analogous reasoning.

Bargaining game with uniform distributions. We return to the example case in the main text, where the distributions of g_1 and g_2 are uniform on $[w_d, 1]$ and $[1 - w_s, 1]$, respectively, and $w_s \leq 0.5 \leq w_d$. It can be shown that linear offer functions $\tilde{\theta}_1(g_1) = \alpha_1 g_1 + \epsilon_1$ and $\tilde{\theta}_2(g_2) = \alpha_2 g_2 + \epsilon_2$ satisfy the first-order conditions Eqs. **8** and **9**, when

$$\alpha_1 = \frac{1+r}{2+r-k(1-r)}$$
$$\epsilon_1 = \frac{2+r+k^2(1-r)(1-w_d)+k\left(\left(1+r+r^2\right)w_d-3\right)}{(2+r)(2+r-k(1-r))} \quad \textbf{[S39]}$$

$$\alpha_2 = \frac{1+r}{1+2r+k(1-r)}$$
$$\epsilon_2 = \frac{(1-k)(k(1-r)+r(2+r))+k(1+k(1-r)+2r)w_d}{(1+k(1-r)+2r)(2+r)}. \quad \textbf{[S40]}$$

It can be seen from these expressions that $\alpha_i < 1$ and $\varepsilon_i > 0$. To fix the boundary conditions, we assume that $w_d \geq 1 - w_s$, so that $\tilde{\theta}_1(w_d) > \tilde{\theta}_2(1-w_s)$ and $\tilde{\theta}_1(1) > \tilde{\theta}_2(1)$ [except for $r = 1$, when we have $\tilde{\theta}_1(w_d) = \tilde{\theta}_2(1-w_s)$, but this nongeneric case can be subsumed by assuming groups do not form at the boundary cases]. In other words, using the above offer functions, both the dominants (individual 1) and the subordinates (individual 2) with the highest outside options (highest g_1 and lowest g_2, respectively) do not form groups with anyone. Then, the expected fitness of a dominant (before it learns its private information) is given by

$$\bar{W}_1 = \int_{w_d}^{\tilde{\theta}_1^{-1}\left(\tilde{\theta}_2(1)\right)} W_1\left(g_1, \tilde{\theta}_1(g_1)\right) f_1(g_1) dg_1, \quad \textbf{[S41]}$$

where $W_1(g_1, \tilde{\theta}_1(g_1))$ is given by Eq. **S36**. Similarly, the expected fitness of a subordinate is given by

$$\bar{W}_2 = \int_{\tilde{\theta}_2^{-1}\left(\tilde{\theta}_1(w_d)\right)}^{1} W_2\left(g_2, \tilde{\theta}_2(g_2)\right) f_2(g_2) dg_2. \quad \textbf{[S42]}$$

Substituting the functions $\tilde{\theta}_i(\cdot)$ and their inverses, Eqs. **S41** and **S42** evaluate to Eqs. **10** and **11** in the main text.

1. Ely J, Sandholm WH (2005) Evolution in Bayesian games I: Theory. *Games Econ Behav* 53:83–109.
2. Myerson R, Satterthwaite M (1983) Efficient mechanisms for bilateral trading. *J Econ Theory* 29:265–281.
3. Chatterjee K, Samuelson W (1983) Bargaining under incomplete information. *Oper Res* 31:835–851.