Evolutionary stability and resistance to cheating in an indirect reciprocity model based on reputation

Luis A. Martinez-Vaquero[1] and José A. Cuesta[1,2,*]

[1]*Grupo Interdisciplinar de Sistemas Complejos, Departamento de Matemáticas, Universidad Carlos III de Madrid,*
28911 Leganés, Madrid, Spain
[2]*Instituto de Biocomputación y Física de Sistemas Complejos, Universidad de Zaragoza, 50009 Zaragoza, Spain*

Indirect reciprocity is one of the main mechanisms to explain the emergence and sustainment of altruism in societies. The standard approach to indirect reciprocity is reputation models. These are games in which players base their decisions on their opponent's reputation gained in past interactions with other players (moral assessment). The combination of actions and moral assessment leads to a large diversity of strategies; thus determining the stability of any of them against invasions by all the others is a difficul task. We use a variant of a previously introduced reputation-based model that let us systematically analyze all these invasions and determine which ones are successful. Accordingly, we are able to identify the third-order strategies (those which, apart from the action, judge considering both the reputation of the donor and that of the recipient) that are evolutionarily stable. Our results reveal that if a strategy resists the invasion of any other one sharing its same moral assessment, it can resist the invasion of any other strategy. However, if actions are not always witnessed, cheaters (i.e., individuals with a probability of defecting regardless of the opponent's reputation) have a chance to defeat the stable strategies for some choices of the probabilities of cheating and of being witnessed. Remarkably, by analyzing this issue with adaptive dynamics we fin that whether an honest population resists the invasion of cheaters is determined by a Hamilton-like rule, with the probability that the cheat is discovered playing the role of the relatedness parameter.

I. INTRODUCTION

The human being is the social animal par excellence. An individual can help another even if it is the firs time they meet or if they know that they will never meet again. Several mechanisms have been proposed to explain cooperation between unrelated individuals. Among them reciprocity, either direct or indirect, stands as one of the most successful explanations of altruism [1]. In direct reciprocity individuals pay back the help received in repeated encounters with the same partner ("I help you if you help me") [2]. In society, however, many interactions have low chances to be repeated with the same individual. To explain altruism in those interactions, the concept of indirect reciprocity was introduced [3,4]. Through this mechanism, individuals do not receive the consequences of their actions directly from the individuals they interact with but indirectly through society ("I help others to be helped by others"). Indirect reciprocity is an important mechanism for the emergence and sustainment of altruism not only in small-scale human societies [5–9] but in other species as well [10]. And it certainly plays an important role in communication networks [11,12].

There are two types of indirect reciprocity: upstream and downstream. In upstream reciprocity [13,14] an individual opts for a given action taking into account if she was previously helped or not. In this respect upstream reciprocity is more akin to a learning mechanism because individuals adapt their choices based on their past experience. In downstream reciprocity, also called reputation-based indirect reciprocity, an individual assigns a reputation to the others, taking into account how they interact with the rest of the society [6,15–18].

These reputations allow her to decide whether she should help these individuals or not in potential future encounters with them. Accordingly, downstream indirect reciprocity is a cognitively very demanding task: it requires observation, memory, and communication. It is this reputation-based indirect reciprocity that will be the focus of the present work.

Two different kinds of models of reputation-based indirect reciprocity have been considered in the literature. In indirect observation models [16] each action is observed and judged only by one individual, who spreads this information across the population through verbal communication and gossip. Therefore all individuals share the same opinion about each other. On the contrary, in direct observation models [17,19,20] everyone witnesses the action and makes a private judgment of it. Thus individuals' different opinions about the rest of the members of the society can coexist in this kind of model.

Ohtsuki and Iwasa [16] and Brandt and Sigmund [17] have proposed a classificatio of the different strategies in games with indirect reciprocity through their assessment and action modules. Strategies can be classifie either as second-order or as third-order strategies. In both cases, the reputation is assigned taking into account the observed action and the reputation of the individual who received it. But third-order strategies also look at the reputation of the individual who performs the action. The dynamics of second-order assessments has been explored in [21]. Ohtsuki and Iwasa [16] also studied systematically the evolutionary stability of third-order strategies. Their model is an indirect observation model, and therefore the whole society shares the same moral assessment. Stability is studied by confronting strategies with different action rules. They concluded that there are eight strategies, the so-called *leading eight*, which are evolutionarily stable strategies (ESS) under these assumptions. The meaning and

————————
*cuesta@math.uc3m.es

success of these strategies has also been studied by Ohtsuki and Iwasa [22]. On the other hand, Uchida and Sigmund [23] have chosen some of the leading eight strategies that share the same action rules but have different moral assessments and have confronted them in a model with private opinions.

In this work, we extend the systematic study carried out by Ohtsuki and Iwasa confronting strategies with different moral assessments. Unlike their work, we use a direct observation model in which individuals no longer share the same opinion about the rest of the population. We introduce the concept of coherence as a measure of the relation between the moral assessment and the action rules of a strategy and study how it relates to the stability and efficiency of the strategies. We identify which strategies resist the invasion of all the other strategies, i.e., which combinations of moral assessment and action rules emerge under this evolutionary competition. Finally, we explore the effect that an action is witnessed by no one in the population. Individuals can then face the risk to cheat, i.e., defect regardless of the opponent's reputation, at no reputation cost.

The present paper is structured as follows. In Sec. II we introduce the model. In Sec. III we describe its mathematical implementation. We study homogeneous populations and discuss their stability against invasions by other strategies. We also analyze the effect on introducing the probability of cheating when actions have a chance not to be witnessed. Finally, our results are shown in Sec. IV and are discussed in Sec. V.

II. MODEL

Brandt and Sigmund [17] introduced a very stylized model of indirect reciprocity based on reputation, and Ohtsuki and Iwasa [16,22] investigated the stability of its strategies under the assumption that all individuals share the same moral judgment.

The model we will be dealing with in this work is a slight modification of this basic model. It consists of an infinite well-mixed population of interacting and judging individuals. Every time step a pair of individuals are randomly and equiprobably drawn from the population. One of them plays the role of the *donor* and the other one of the *recipient*. The donor then decides whether to pay a cost $c > 0$ to help (C) the recipient or not (D). If the recipient is helped, she receives a benefit $b > c$. This *action* is observed by every individual of the population (including themselves). Observers privately judge the donor for the action taken on the recipient according to their own *moral assessment* and assign her a reputation, either good (G) or bad (B), accordingly. Therefore every individual in the population has a private opinion of every other individual, including herself.

This process is repeated until the population reaches an equilibrium (we will define this equilibrium in more precise terms in the next section). Then the average payoff that every individual receives in this repeated game is computed. Direct reciprocity is excluded from this game because the population is virtually infinite hence the probability that two people meet again is negligible.

We consider third-order indirect reciprocity; i.e., each strategy is described by two moduli: the action rules and the moral assessments. The action rules determine what the donor must do (either help or refuse to help) given the reputation

of both players. Specifically, $a_{i\alpha\beta} = 1$ (C) if strategist i with reputation α helps an individual with reputation β (both according to i's moral judgments) and 0 (D) otherwise.

The moral assessments tell the individual if the action just witnessed should be judged as good or bad, hence revising the donor's reputation. Specifically, $m_{i\alpha\beta}(a) = 1$ (G) if strategist i assigns a good reputation to a donor previously judged α by i, who performs an action a on a recipient previously judged β by i, and is 0 (B) otherwise.

Thus each strategy is defined by 12 numbers: 4 for the action module and 8 for the moral module. This amounts to 4096 different possible strategies. Although a thorough study of mutual invasions and the coexistence of different strategies, such as that performed in Ref. [24] for a direct reciprocity model, would be desirable, the abundance of strategies forbids it, and we content ourselves with a pairwise test of mutual invadability.

We will assume that sometimes players do not act according to their action rules [16,19,25–27]. Thus, with a probability ϵ_A a donor defects regardless of her action rules, and with $1 - \epsilon_A$ she performs the action she planned to. Another source of errors is misjudgment; i.e., an individual can make a mistake in interpreting the action. In this category lies social pressure. This is a kind of error that is especially important if the information on the action performed is spread by gossiping because, then, a misjudgment of the witness will lead to a misjudgment of the entire population. Otherwise, it affects only a small fraction of the individuals. Since keeping track of errors may lead to a proliferation of judgments, even between individuals sharing the same moral assessment, and may render the model computationally unfeasible, we will content ourselves by implementing only errors in the action.

III. MATHEMATICAL IMPLEMENTATION OF THE MODEL

A. Homogeneous populations

Let us start by assuming that there is only one strategy i present in the population. Let x_i be the fraction of individuals considered good by the whole population (there is a unique moral assessment). Then the rate of change of x_i is given by

$$\frac{dx_i}{dt} = \sum_{\alpha\beta} \chi_\alpha(x_i)\chi_\beta(x_i)P_{i,\alpha\beta} - x_i, \tag{1}$$

where $P_{i,\alpha\beta}$ is the probability that a donor of reputation α acting on a recipient with reputation β is considered good by the population. This probability can be obtained as

$$P_{i,\alpha\beta} = (1 - \epsilon_A)m_{i\alpha\beta}(a_{i\alpha\beta}) + \epsilon_A m_{i\alpha\beta}(D) \tag{2}$$

because with probability ϵ_A no help is provided and with probability $1 - \epsilon_A$ the action performed is $a_{i\alpha\beta}$, as prescribed by the action module. We have also introduced the auxiliary function $\chi_\gamma(x_i)$,

$$\chi_\gamma(x_i) = \gamma x_i + (1 - \gamma)(1 - x_i), \tag{3}$$

which in this case represents the fraction of individuals with reputation γ.

The dynamics reaches an equilibrium when $x_i = \sum_{\alpha\beta} \chi_\alpha(x_i)\chi_\beta(x_i)P_{i,\alpha\beta}$. Therefore the fraction of good

individuals in a homogeneous population in equilibrium is the solution $0 \leqslant x_i \leqslant 1$ of the quadratic equation $F(x_i) = 0$, where

$$F(x_i) = x_i^2(P_{i,11} + P_{i,00} - P_{i,10} - P_{i,01})$$
$$+ x_i(P_{i,10} + P_{i,01} - 2P_{i,00} - 1) + P_{i,00}. \quad (4)$$

As $F(0) = P_{i,00} \geqslant 0$ and $F(1) = P_{i,11} - 1 \leqslant 0$, there is always a solution in $[0,1]$, but in some cases there may be two (when $P_{i,00} = 0$ or $P_{i,11} = 1$ or both), one stable and one unstable, and there is a degenerate case [when all coefficient in $F(x_i)$ vanish] in which any x_i is a solution. In this latter case, adding a small error ϵ_m in the moral assessment determines uniquely a stable solution. When the population is homogeneous, this can be done at no computational cost by simply replacing $P_{i,\alpha\beta}$ in Eq. (4) by $(1 - 2\epsilon_m)P_{i,\alpha\beta} + \epsilon_m$. This yields the expression $F(x_i) = \epsilon_m(1 - 2x_i)$, whose only root is $x_i = 1/2$, regardless of ϵ_m. Hence we take this solution, which holds even in the limit $\epsilon_m \to 0$, as the solution of this degenerate case.

Given the equilibrium fraction x_{iH}, the probability that an individual helps another is

$$\theta_{iH} = (1 - \epsilon_a) \sum_{\alpha\beta} \chi_\alpha(x_{iH})\chi_\beta(x_{iH})a_{\alpha\beta}. \quad (5)$$

Therefore the average payoff that any individual in this population obtains is

$$W_{iH} = (b - c)\theta_{iH}. \quad (6)$$

As the whole population shares the same strategy, it can be regarded a measure of "self-efficiency." This provides a mean to classify strategies.

Coherence provides an alternative classification criterion. Given an action a that a donor with reputation α performs on a recipient with reputation β, we call an individual coherent if placed on the donor's feet she performs the same action a when she morally assesses it as good and the opposite action $1 - a$ when she morally assesses it as bad. In other words, an individual is coherent if she performs actions that she judges as good and does the opposite for actions that she judges as bad. Thus we can introduce a coherency index h as

$$h_i = \frac{1}{2}\sum_{\alpha\beta a}[1 - |m_{i\alpha\beta}(a) - \delta(a, a_{i\alpha\beta})|]\chi_\alpha(x_{iH})\chi_\beta(x_{iH}), \quad (7)$$

where $\delta(x, y) = 1$ if $x = y$ and 0 otherwise. This index can range from 0 (no coherence) to 1 (full coherence). Notice that the coherence of a strategy can change when more strategies are present in the population because it depends on the fraction of good and bad individuals. Nevertheless, for the sake of classification we have defined this index for a homogeneous population so that it is uniquely determined by x_{iH} and therefore is an intrinsic feature of each strategy.

B. Stability of strategies

Consider now a homogeneous population where individuals share the same *resident* strategy. From time to time a small fraction of the population can adopt a new *mutant* strategy. This mutant strategy will eventually invade the resident population if mutants obtain a higher payoff than residents.

Calculating these payoffs requires computing the four fractions of individuals that are considered good and bad by the first and the second strategies in equilibrium. In the limit where the fraction of mutants is very small both residents and mutants interact only with residents. The dynamics of these four fractions of individuals is given in this limit by the equations

$$\frac{dx_1^{\Lambda_1\Lambda_2}}{dt} = \sum_{\substack{\alpha_1\alpha_2 \\ \beta_1\beta_2}} x_1^{\alpha_1\alpha_2} x_1^{\beta_1\beta_2} P_{1,\alpha_1\beta_1,\alpha_2\beta_2}^{\Lambda_1\Lambda_2} - x_1^{\Lambda_1\Lambda_2}, \quad (8)$$

$$\frac{dx_2^{\Lambda_1\Lambda_2}}{dt} = \sum_{\substack{\alpha_1\alpha_2 \\ \beta_1\beta_2}} x_2^{\alpha_1\alpha_2} x_1^{\beta_1\beta_2} P_{2,\alpha_1\beta_1,\alpha_2\beta_2}^{\Lambda_1\Lambda_2} - x_2^{\Lambda_1\Lambda_2}, \quad (9)$$

where $x_i^{\Lambda_1\Lambda_2}$ are the fractions of i strategists ($i = 1$ for residents and $i = 2$ for mutants) who are judged Λ_1 by residents and Λ_2 by mutants; $P_{i,\alpha_1\beta_1,\alpha_2\beta_2}^{\Lambda_1\Lambda_2}$ is the probability that an i strategist with reputation α_1 for residents and α_2 for mutants, acting on a recipient of the resident population with reputation β_1 for other residents and β_2 for mutants, is judged Λ_1 by residents and Λ_2 by mutants. The form of this probability is

$$P_{i,\alpha_1\beta_1,\alpha_2\beta_2}^{\Lambda_1\Lambda_2} = (1 - \epsilon_A)\delta(\Lambda_1, m_{1\alpha_1\beta_1}(a_{i\alpha_i\beta_i}))\delta(\Lambda_2, m_{2\alpha_2\beta_2}(a_{i\alpha_i\beta_i}))$$
$$+ \epsilon_A\delta(\Lambda_1, m_{1\alpha_1\beta_1}(D))\delta(\Lambda_2, m_{2\alpha_2\beta_2}(D)). \quad (10)$$

Equations (8) and (9) can be simplified in the equilibrium. Nonetheless, some of the equations need to be numerically solved (see Appendix A). To this purpose we must start from a sensible initial condition. We will assume that just before the invasion begins, all individuals, both mutants and residents, share the same opinion about everybody. The rationale for this choice is that, before the change of strategy undergone by mutants takes place, the population was homogeneous. Therefore $x_i^{GG}(0) = x_{iH}$, $x_i^{BB}(0) = 1 - x_{iH}$ and $x_i^{GB}(0) = x_i^{BG}(0) = 0$.

Once the fractions in equilibrium $x_i^{\Lambda_1\Lambda_2}$ are known, the probabilities $\theta_{i,j}$ that an i strategist helps a j strategist ($i, j = 1, 2$) are obtained as

$$\theta_{1,j} = (1 - \epsilon_A) \sum_{\alpha\beta} \chi_\alpha(x_1^{G*}) \chi_\beta(x_j^{G*}) a_{1\alpha\beta},$$
$$\theta_{2,j} = (1 - \epsilon_A) \sum_{\alpha\beta} \chi_\alpha(x_2^{*G}) \chi_\beta(x_j^{*G}) a_{2\alpha\beta}, \quad (11)$$

where we have introduced the short-hand notation $x_i^{G*} = \sum_{\Lambda_2} x_i^{G\Lambda_2}$ and $x_i^{*G} = \sum_{\Lambda_1} x_i^{\Lambda_1 G}$ to denote the sum over a given reputation. Obviously, x_i^{G*} (x_i^{*G}) is the fraction of i strategists that are judged as good by the resident (mutant) players irrespective of the mutant's (resident's) judgment.

Finally, the average payoff $W(i|j)$ that an i strategist receives from a j strategist can be computed as

$$W(i|j) = \begin{cases} (b - c)\theta_{i,i}, & i = j, \\ b\theta_{j,i} - c\theta_{i,j}, & i \neq j. \end{cases} \quad (12)$$

The resident population cannot be invaded by the mutants if $W(1|1) > W(2|1)$ or if $W(1|1) = W(2|1)$ and $W(1|2) > W(2|2)$. If the resident strategy resists the invasions of all the other mutant strategies, it is considered evolutionarily stable.

IV. RESULTS

A. Stability of strategies

Our aim is to identify strategies that are evolutionarily stable. In principle this requires us, for every strategy, to check whether it can be invaded by every other strategy. However, the number of pairs of strategies is larger than 1.5×10^7, so this becomes too demanding a computational task. Accordingly, we proceed in two steps: (i) we look for all strategies that are stable against invasions by other strategies sharing the same moral assessment, and (ii) we study the stability of these selected strategies against all the remaining ones.

Our Eqs. (8) and (9) reduce to those used in Ref. [16] if we fi the moral assessments and neglect moral errors. We carried out our analysis for different values of the action error ϵ_A (0.1, 0.01, and 0.001) and benefit-to-cos b/c ratio (1.2, 1.5, 2, and 3).

In Fig. 1 we represent the strategies that are stablae against invasions by all strategies sharing the same moral assessment as a function of their normalized average payoff $\widetilde{W}_H = W_H [(b-c)(1-\epsilon_A)]^{-1}$ and their coherence. These strategies always appear in pairs since there is a symmetry in the reputation: if labels "good" and "bad" are exchanged, the results are not affected (see [16] for more details). Notice, though, that there is symmetry only in the moral assessment but not in the action. The reason is that cooperating and defecting are not just labels because they have consequences in the

payoffs obtained. It is easy to show, using Eq. (7), that the sum of the coherences of a strategy and its "mirror" strategy is always 1. Coherence thus provides an external assessment on moral labels, breaking the symmetry and permitting us to differentiate between a strategy and its mirror. In Fig. 1 we only show the results for the coherent strategy ($h \geqslant 0.5$) of the pair and report how many pairs N_p are shown.

Figure 1 shows that the larger the benefit-to-cos ratio is, the higher the number of stable strategies is; in other words, it is difficul to break into a population whose individuals obtain high rewards for help. Moreover, we have counted the number of pairs of strategies in which each strategy can be invaded by the other; i.e., at least one mixed equilibrium is formed. The number of these pairs also appears to be larger the higher the benefit-to-cos ratio is (2500–2600 pairs for $b/c = 3$ vs 1500–1700 for the remaining cases). Therefore, even if a mutant invades a resident strategy, it is less likely that it eventually dominates the population if b/c is high. From Fig. 1 we also conclude that a high ϵ_A allows invaders to spread easier in the resident population.

On the other hand, payoff and coherence seem to be correlated. Specificall , stable strategies with high payoff are highly coherent (incoherent for the their mirror strategies). In Table I we list all coherent stable strategies along with their payoffs for $b/c = 2$ and $\epsilon_A = 0.01$. Most of them coincide with those found by [16]. There are some minor differences, though, because we are using slightly different models. The
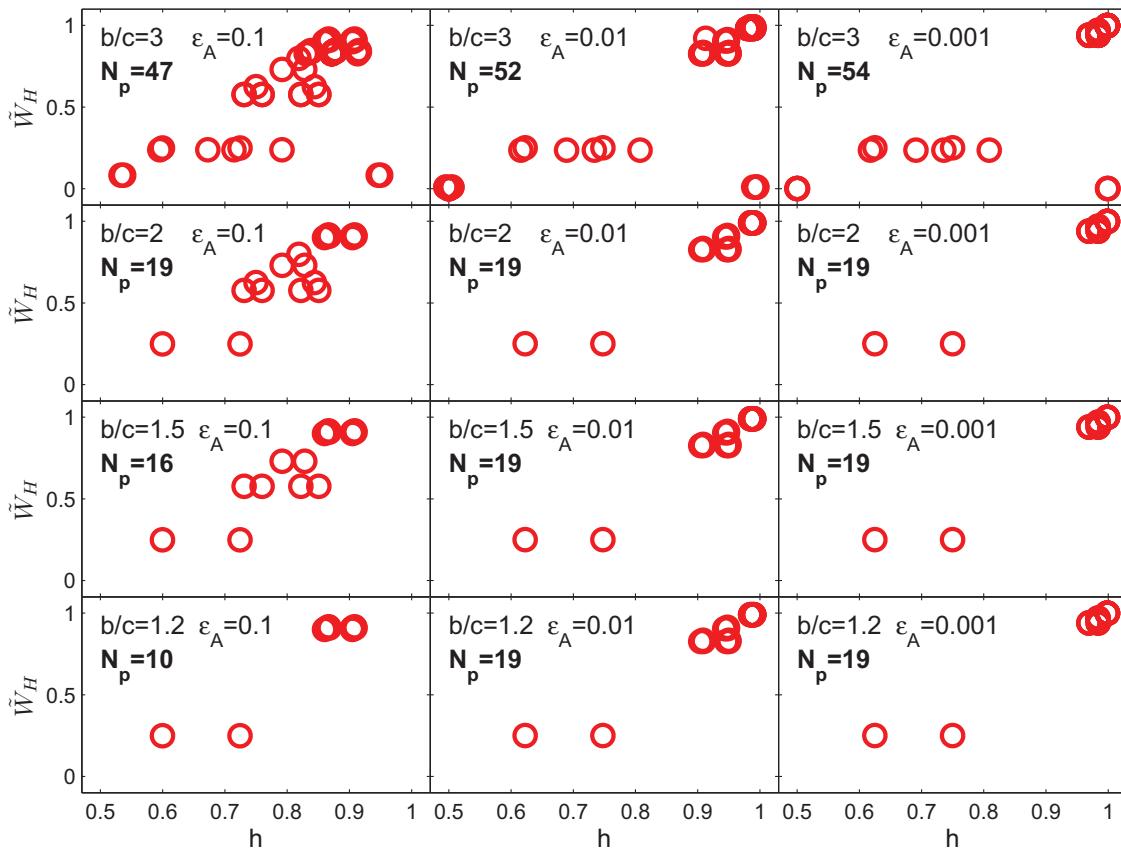

FIG. 1. (Color online) Representation of the normalized average payoff $\widetilde{W}_H$ as a function of the coherence h for the stables strategies. Only the coherent strategy ($h > 0.5$) of each pair is represented. Different panels show results for different ϵ_A and b/c.

TABLE I. Coherent stable strategies and their normalized average payoffs $\widetilde{W}_H$ for the case $b/c = 2$ and $\epsilon_A = 0.01$. The top eight strategies (labeled Ia to IIIb) are the so-called leading eight [16]. They are the ones with the highest payoffs among all the stable strategies obtained for a given benefit-to-cos ratio b/c.

	$m_{GG}(C)$	$m_{GG}(D)$	$m_{GB}(C)$	$m_{GB}(D)$	$m_{BG}(C)$	$m_{BG}(D)$	$m_{BB}(C)$	$m_{BB}(D)$	a_{GG}	a_{GB}	a_{BG}	a_{BB}	$\widetilde{W}_H$
Ia	G	B	G	G	G	B	G	B	C	D	C	C	0.9902
Ib	G	B	B	G	G	B	G	B	C	D	C	C	0.9902
IIa	G	B	G	G	G	B	G	G	C	D	C	D	0.9901
IIb	G	B	G	G	G	B	B	G	C	D	C	D	0.9901
IIc	G	B	B	G	G	B	G	G	C	D	C	D	0.9901
IId	G	B	B	G	G	B	B	G	C	D	C	D	0.9901
IIIa	G	B	G	G	G	B	B	B	C	D	C	D	0.9900
IIIb	G	B	B	G	G	B	B	B	C	D	C	D	0.9900
	G	B	B	B	G	B	G	B	C	D	C	C	0.9135
	G	B	B	B	G	B	G	G	C	D	C	D	0.9049
	G	B	B	B	G	B	B	G	C	D	C	D	0.9049
	G	B	B	G	B	B	G	B	C	D	D	C	0.8340
	G	B	G	G	B	B	G	B	C	D	D	C	0.8340
	G	B	B	G	B	B	B	G	C	D	D	D	0.8264
	G	B	B	G	B	B	G	G	C	D	D	D	0.8264
	G	B	G	G	B	B	B	G	C	D	D	D	0.8264
	G	B	G	G	B	B	G	G	C	D	D	D	0.8264
	B	B	B	G	G	B	B	B	D	D	C	D	0.2500
	B	B	G	G	G	B	B	B	D	D	C	D	0.2500

eight strategies with the highest payoff correspond to the so-called "leading eight" [16]. These strategies are present in all cases shown in Fig. 1. All stable strategies have some common features: (i) not helping good individuals is always considered bad, (ii) good individuals never help bad ones, and (iii) good individuals always help good individuals, except when errors occur, and that is judged as good. (There are two strategies for which the last feature is quite the opposite, but they receive rather low payoffs.)

Notice that the absence of errors in the moral assessments renders all defective strategies (strategies that always defect) vulnerable to invasions. Ohtsuki and Iwasa [16] found that all defective strategies were stable; the reason is that, although these strategies are never rewarding, errors in judgments provide them some payoff. This does not happen in the present model. Thus defective strategies are no longer stable.

Once we identifie the strategies that cannot be invaded by others with the same moral assessments, we studied which of them are actually stable against the invasion by any other strategy. We have found that all those strategies remain stable even if strategies with different moral assessments try to invade them. We have also checked that strategies that can be invaded by other strategies with the same moral assessment can be invaded by some strategies with different moral assessment as well.

B. Robustness against initial misjudgments

We have checked the sensitivity of these results with respect to a different choice of the initial conditions to solve Eqs. (8) and (9). In Sec. III B we made the assumption that, before a mutation occurs, all individuals share the same opinion about everybody because the population is homogeneous. Initial misjudgments can lead a fraction of the population to disagree with the general opinion. This choice for initial conditions may be modeled as

$$x_i^{GG}(0) = \left(1 - \epsilon_r^B\right)x_{iH},$$
$$x_i^{GB}(0) = \epsilon_r^B x_{iH}, \quad x_i^{BB}(0) = \left(1 - \epsilon_r^G\right)(1 - x_{iH}), \quad (13)$$
$$x_i^{BG}(0) = \epsilon_r^G(1 - x_{iH}),$$

where ϵ_r^B (ϵ_r^G) is the fraction of individuals that are misjudged as bad (good) by the mutants. Note that if $\epsilon_r^B = \epsilon_r^G = 0$, the whole population agrees in its judgments, and we recover the former initial conditions.

Depending on the (small) values of ϵ_r^B and ϵ_r^G, we have checked that the initial conditions (13) may lead to three different scenarios. In the firs one $x_i^{GG} = x_{iH}$, $x_i^{BB} = 1 - x_{iH}$, and $x_i^{GB} = x_i^{BG} = 0$, so that misjudgments fade away, and we recover a homogeneous population. In the second scenario initial misjudgments remain or even grow (x_i^{GG} and x_i^{BB} decrease and x_i^{GB} and x_i^{BG} increase), but the payoff obtained by the mutants is lower than that obtained by the residents. Consequently, the mutants are expelled, and a homogeneous population is restored. In the third scenario initial misjudgments also remain, and the mutants obtain higher payoffs than the residents, so that misjudgments eventually spread. We have found that around 850 strategies lie in this last case (considering differences between mutants' and residents' payoffs higher than 10^{-6}) when $b/c = 2$ and $\epsilon_A = 0.01$. Fortunately, none of these strategies belong to the group of stable ones, so this misjudgment spreading does not affect the evolutionary fate of the population.

C. Stability in the presence of cheating

Consider now the situation in which actions are not always witnessed; instead, there is a chance that they pass unnoticed
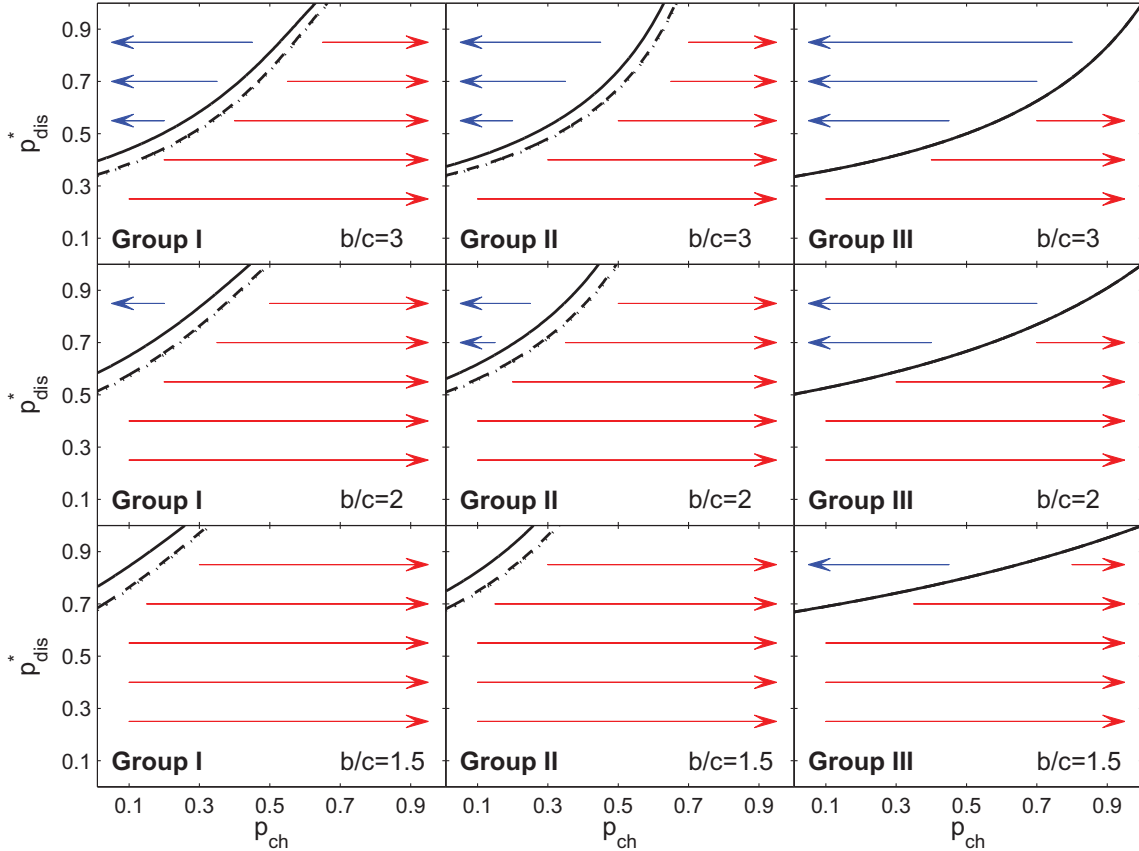
FIG. 2. (Color online) Limit curves of $p^*_{\rm dis}$ as a function of $p_{\rm ch}$ that divide the regions where cheating can be increased (above the curves) and decreased (below the curves) through the invasion of mutants with different p_{ch}. Different types of lines represent different values of ϵ_A: 0.1 (solid), 0.01 (dashed), and 0.001 (doted). Different panels show results for different groups of the leading eight strategies and b/c.

by the rest of the population. In this situation individuals may have the temptation to cheat by defecting regardless of their action rules. The appearance of this kind of mutation introduces a new set of strategies, parameterized by the cheating probability $p_{\rm ch}$, which might render unstable strategies that would otherwise resist invasions. The stability will, of course, be a function of the probability that the action is witnessed $p_{\rm dis}$.

To address this issue let us consider that residents decide to cheat with a probability $p_{\rm ch,1}$ and mutants do so with a probability $p_{\rm ch,2}$ in the hope that they are not discovered. However, their cheating will actually be discovered with a probability $p_{\rm dis}$. Assuming the same moral assessments and action rules for all individuals, $x_1^{G*} = x_1^{*G} = x_{1H}$ and $x_2^{G*} = x_2^{*G} = x_2^{\rm ch}$, where the fractions $x_{1H}^{\rm ch}$ and $x_2^{\rm ch}$ are calculated as from Eqs. (4) and (A3), but incorporating the probability of being discovered if they cheat. Likewise, $P_{i,\alpha\beta}$ in Eq. (2) has to be replaced by

$$P^{\rm ch}_{i,\alpha\beta} = (1 - p_{\rm dis}p_{\rm ch,i})P_{i,\alpha\beta} + p_{\rm dis}p_{\rm ch,i}m_{i\alpha\beta}(D), \quad (14)$$

which expresses the fact that nothing changes if player i either does not cheat or does so without being discovered [probability $1 - p_{\rm ch,i} + p_{\rm ch,i}(1 - p_{\rm dis}) = 1 - p_{\rm dis}p_{\rm ch,i}$]; otherwise (probability $p_{\rm dis}p_{\rm ch,i}$), she is judged good or bad according to $m_{i\alpha\beta}(D)$.

Finally, the probabilities of cooperation [cf. Eq. (11)] are modifie as

$$\theta^{\rm ch}_{1,1} = (1 - p_{\rm ch,1})(1 - \epsilon_A)\sum_{\alpha\beta}\chi_\alpha(x_{1H}^{\rm ch})\chi_\beta(x_{1H}^{\rm ch})a_{1\alpha\beta},$$

$$\theta^{\rm ch}_{1,2} = (1 - p_{\rm ch,1})(1 - \epsilon_A)\sum_{\alpha\beta}\chi_\alpha(x_{1H}^{\rm ch})\chi_\beta(x_2^{\rm ch})a_{1\alpha\beta},$$

$$\theta^{\rm ch}_{2,1} = (1 - p_{\rm ch,2})(1 - \epsilon_A)\sum_{\alpha\beta}\chi_\alpha(x_2^{\rm ch})\chi_\beta(x_{1H}^{\rm ch})a_{2\alpha\beta}, \quad (15)$$

$$\theta^{\rm ch}_{2,2} = (1 - p_{\rm ch,2})(1 - \epsilon_A)\sum_{\alpha\beta}\chi_\alpha(x_2^{\rm ch})\chi_\beta(x_2^{\rm ch})a_{2\alpha\beta}.$$

We have studied the stability of the leading eight strategies against the invasion of cheaters. We divided the leading eight strategies in groups I, II, and III as a function of their different behaviors (as was done in [16]). Figure 2 represents the limiting $p^*_{\rm dis}$ below (above) which mutants cheat with a higher (lower) probability than residents can invade. In Appendix B we calculate analytically the shape of this curve in the limit $\epsilon_A \to 0$. Figure 2 shows that below the curve $p^*_{\rm dis}(p_{\rm ch})$ cheating increases without bound through subsequent invasions until the whole population is dominated by defectors. In other words, if cheating occurs and the probability of being discovered is not high enough, none of the leading eight strategies survives. In particular, if $p_{\rm dis} < c/b$, full defection is the unavoidable fate

of the population. Thus, if only small mutations are allowed in an honest population, we fin the Hamilton-like rule $bp_{\text{dis}} > c$ for the survival of cooperation [28].

Increasing ϵ_A makes it even easier for cheaters to invade, with the exception of the strategies belonging to group III, which seem to be insensitive to the effect of errors in action.

V. DISCUSSION

We have carried out a systematic study of the stability of all possible third-order indirect reciprocity strategies. We extended the work of Ohtsuki and Iwasa [16] confronting all the strategies against the others regardless of whether they have the same moral assessments or not. The main difference with their model is that in ours individuals directly witness all actions. Allowing individuals in the same population to have different moral assessments and action rules makes indirect observation models computationally unfeasible (we must store everybody's opinion of everybody else at every time step). For the same reason, errors in judgments cannot be accounted for in direct observation models. Thus we only consider errors in performing the actions. The only exception to this assumption is the need to introduce errors in judgment to calculate, in some special cases, the stationary fractions of good and bad individuals in homogeneous populations. But this is just a technical issue that allows us to resolve a degeneracy of solutions, and there is no inconsistency because the results do not depend on the value of this error.

The strategies which are stable against invasions by other strategies sharing the same moral assessment turn out to also be stable against invasions by any other strategy. This means that if a strategy can resist the invasion of all the other strategies that share its same moral assessment, it can resist any invasion whatsoever.

We have checked that the higher the benefit-to-cos ratio is and the lower the action errors are, the higher the number of stable strategies obtained is. One possible interpretation of the errors in action is a lack of resources. Under this interpretation our results show that scarcity of resources favors invasions. On the other hand, we have checked that populations whose members receive a greater benefi for a given cost are more resistant to invasions.

As pointed out in Ref. [16], there is a symmetry between the moral assessments of the strategies. Good and bad are just labels with no proper meaning, in contrast to actions, which have a direct influenc in the payoffs. In order to break that symmetry and provide a meaning to those labels we have introduced the concept of coherence. Coherence links moral assessments with action rules. We have shown that stable strategies appear in pairs due to the above-mentioned symmetry, but coherence values are complementary. This allows us to choose only one of the strategies (the most coherent) within each pair for later analysis and interpretation.

The stable strategies we obtain include the leading eight found by Ohtsuki and Iwasa [16]. These are also the most efficien ones (those with the highest payoffs). Both the leading eight and the remaining stable strategies that we have obtained share some features, and except for the two least efficien strategies (with $\widetilde{W}_H = 0.25$), all of them obtain high average payoffs ($\widetilde{W}_H > 0.8$). They identify defectors

$[m_{GG}(D) = m_{BG}(D) = B]$ and, except the two least efficien strategies, maintain cooperation [$a_{GG} = C$ and $m_{GG}(C) = G$]. All of them punish defectors ($a_{GB} = D$), although three of the stable strategies (with $\widetilde{W}_H \sim 0.9$) do not judge this as a good behavior. Finally, the most efficien stable strategies ($\widetilde{W}_H > 0.9$) forgive bad individuals who help good players [$m_{BG}(C) = G$ and $a_{BG} = C$]. The higher the number of these features the strategies follow is, the higher their payoff is. For instance, the three strategies with $\widetilde{W}_H \sim 0.9$ turn good punishers into bad individuals, and they can only restore their reputation by helping good individuals. And in the case of strategies with $\widetilde{W}_H < 0.9$, bad individuals cannot increase their reputation by helping good players, only by interacting with other bad individuals.

We have also found that all these strategies may become unstable if cheaters arise. If the probability of witnessing a cheat is not high enough, cheaters can take over an honest population. Upon increasing the cheating probability $p_{\text{dis}} > c/b$ the population eventually turns into pure defectors. Interestingly, the condition for a population to resist this effect is of the Hamilton type, namely, $bp_{\text{dis}} > c$, where b is the benefi and c is the cost. Errors in action make this condition even more restrictive for the stability of an honest population.

Cheating is always a danger for cooperation based on indirect reciprocity. Even in societies where this mechanism is of utmost importance cheating always threatens honest behavior. For instance, the (now extinct) Patagonian tribes of the Yámana are among the reported societies more strongly based on indirect reciprocity [29]. Sharing food even with nonrelatives appeared to be the default behavior. Not sticking to it brought a bad reputation and severe social punishment (e.g., not participating in further food sharing). Yet cheating among the Yámana was reported to occur when chances of being discovered were low (for instance, because the prey obtained was easy to hide; see Ref. [29], p. 197).

One of the problems that emerges from considering different moral assessments is the possibility that the fractions of good and bad individuals may depend on the initial setup. We sort out this issue by choosing realistic initial conditions for the differential equations describing the evolution of these fractions. Essentially, we assume that mutations do not change the previous judgments that individuals had of each other. This notwithstanding, we have analyzed other initial conditions in which not all individuals have the same opinion. A typical setup where this might happen is when a rumor is spread over a fraction of the population. We have checked that, although misjudgment can survive or even spread over a larger fraction of the population, it eventually disappears because mutants with a wrong judgment get less payoff than residents who use one of the stable strategies.

Admittedly, in order to carry out such a systematic analysis as we have performed here, we have had to sacrific some realism in the model. On the one hand, we have considered that reputation can only have two states: good and bad. This binary reputation has been used in several preceding studies [16,23] and implies that only the actions that happen in the last round are taken into account when assigning reputation. However, Tanabe *et al.* [30] have studied a model with trinary reputations and showed that some strategies (like the so-called *image scoring*) can be stable in a trinary-reputation model

but not in a binary-reputation one. On the other hand, we have considered that every player has complete information of every single interaction in the population (except when we introduced cheating). This is too strong an assumption, and some studies discuss the effect of limited access to the information (see [31] and references therein).

APPENDIX A

The two sets of Eqs. (8) and (9) can be simplifie in the steady state $dx/dt = 0$. Thus, summing over the reputation Λ_2 in Eqs. (8), we obtain

$$x_1^{G*} = \sum_{\Lambda_2} x_1^{G\Lambda_2} = x_{1H}. \quad (A1)$$

Therefore we can reduce Eqs. (8) to just two equations in two unknowns (e.g., x_1^{GG} and x_1^{BB}) by setting

$$x_1^{GB} = x_1^{G*} - x_1^{GG}, \quad x_1^{BG} = 1 - x_1^{G*} - x_1^{BB}. \quad (A2)$$

The two remaining equations from (8) have to be solved numerically using the initial conditions discussed in Sec. III B.

On the other hand, the set of Eqs. (9) is decoupled from the set (8), and so they can be solved analytically after solving the latter. This is easier if x_2^{*G} is calculated first

$$x_2^{*G} = \left[x_1^{*G} P_{2,01} + \left(1 - x_1^{*G}\right) P_{2,00}\right]\left[1 + x_1^{*G}(P_{2,01} - P_{2,11}) + \left(1 - x_1^{*G}\right)(P_{2,00} - P_{2,10})\right]^{-1}. \quad (A3)$$

Hence Eq. (9) reduces to a linear system of two equations in the two unknowns x_2^{GG}, x_2^{BB}.

There are scenarios where the solution of $x_2^{\Lambda_1\Lambda_2}$ turns out to be degenerate. In these situations the set of Eqs. (9) needs to be integrated along with the set of Eqs. (8).

APPENDIX B

Consider a resident population whose individuals play one of the leading eight strategies with probability $1 - p_{ch,1}$ but defect otherwise. Consider mutants who do the same, but with a probability $1 - p_{ch,2}$. For simplicity let us assume the limiting case $\epsilon_A \to 0$. Applying adaptive dynamics [32], the curve separating the regions where the mutant can or cannot invade the population is given by

$$\left.\frac{dW(p_{ch,2}, p_{ch,1})}{dp_{ch,2}}\right|_{p_{ch,2} = p_{ch,1}} = 0, \quad (B1)$$

where the payoff $W(p_{ch,2}, p_{ch,1})$ is equivalent to $W(2|1)$. According to Eq. (12),

$$\frac{dW(p_{ch,2}, p_{ch,1})}{dp_{ch,2}} = b\frac{d\theta_{1,2}^{ch}}{dp_{ch,2}} - c\frac{d\theta_{2,1}^{ch}}{dp_{ch,2}}. \quad (B2)$$

To go further we need to separate the strategies of the three groups.

1. Group I strategies

Using Eqs. (15) for the leading eight strategies, the probabilities of cooperation $\theta_{i,j}^{ch}$ are

$$\theta_{1,2}^{ch} = (1 - p_{ch,1})\left[x_2^{ch} + \left(1 - x_{1,H}^{ch}\right)\left(1 - x_2^{ch}\right)\right],$$
$$\theta_{2,1}^{ch} = (1 - p_{ch,2})\left[x_{1,H}^{ch} + \left(1 - x_{1,H}^{ch}\right)\left(1 - x_2^{ch}\right)\right]. \quad (B3)$$

Thus

$$\frac{d\theta_{1,2}^{ch}}{dp_{ch,2}} = (1 - p_{ch,1})\, x_{1,H}^{ch}\, \frac{dx_2^{ch}}{dp_{ch,2}},$$
$$\frac{d\theta_{2,1}^{ch}}{dp_{ch,2}} = \left(1 - x_{1,H}^{ch}\right)\left[x_2^{ch} - (1 - p_{ch,2})\frac{dx_2^{ch}}{dp_{ch,2}}\right] - 1. \quad (B4)$$

The fractions $x_{1,H}^{ch}$ and x_2^{ch} are obtained from Eqs. (4) and (A3). To that purpose we need to substitute

$$P_{i,11}^{ch} = P_{i,01}^{ch} = 1 - p_{dis}p_{ch,i}, \qquad P_{i,10}^{ch} = 1 \quad (B5)$$

and

$$P_{i,00}^{ch} = 1 - p_{dis}p_{ch,i}. \quad (B6)$$

Thus $x_{1,H}^{ch}$ is the solution of

$$p_{dis}p_{ch,1}\left(x_{1,H}^{ch}\right)^2 = (1 - p_{dis}p_{ch,1})\left(1 - x_{1,H}^{ch}\right), \quad (B7)$$

and once it is obtained,

$$x_2^{ch} = \frac{1 - p_{dis}p_{ch,2}}{1 - (1 - x_{1,H}^{ch})p_{dis}p_{ch,2}},$$
$$\frac{dx_2^{ch}}{dp_{ch,2}} = -\frac{p_{dis}x_{1,H}^{ch}}{\left[1 - \left(1 - x_{1,H}^{ch}\right)p_{dis}p_{ch,2}\right]^2}. \quad (B8)$$

Substituting into (B4) and setting $p_{ch,2} = p_{ch,1} \equiv p_{ch}$ yield

$$\frac{d\theta_{1,2}^{ch}}{dp_{ch}} = -\frac{(1 - p_{ch,1})p_{dis}\left(x_{1,H}^{ch}\right)^2}{[1 - p_{dis}p_{ch}(1 - x_{1,H}^{ch})]^2},$$
$$\frac{d\theta_{2,1}^{ch}}{dp_{ch,2}} = \frac{x_{1,H}^{ch}[p_{dis}(1 - x_{1,H}^{ch}) - 1]}{\left[1 - p_{dis}p_{ch}(1 - x_{1,H}^{ch})\right]^2}. \quad (B9)$$

Therefore p_{dis}^* is the solution of the system

$$p_{dis}^*[b(1 - p_{ch})x^* + c(1 - x^*)] = c,$$
$$p_{dis}^*p_{ch}(x^*)^2 = (1 - p_{dis}^*p_{ch})(1 - x^*). \quad (B10)$$

2. Group II strategies

For the strategies of this group

$$\theta_{1,2}^{ch} = (1 - p_{ch,1})x_2^{ch}, \qquad \theta_{2,1}^{ch} = (1 - p_{ch,2})x_{1,H}^{ch}; \quad (B11)$$

hence their derivatives are

$$\frac{d\theta_{1,2}^{\mathrm{ch}}}{dp_{\mathrm{ch},2}} = (1 - p_{\mathrm{ch},1})\frac{dx_2^{ch}}{dp_{\mathrm{ch},2}}, \qquad \frac{d\theta_{2,1}^{\mathrm{ch}}}{dp_{\mathrm{ch},2}} = -x_1^{\mathrm{ch}}. \quad \text{(B12)}$$

Probabilities $P_{i,\alpha\beta}^{\mathrm{ch}}$ are now given by (B5) as well as $P_{i,00}^{\mathrm{ch}} = 1$. Thus, after Eqs. (4) and (A3),

$$x_{1,H}^{\mathrm{ch}} = \frac{1}{1 + p_{\mathrm{dis}}p_{\mathrm{ch},1}}, \qquad x_2^{\mathrm{ch}} = 1 - x_{1,H}^{\mathrm{ch}}p_{\mathrm{dis}}p_{\mathrm{ch},2}. \quad \text{(B13)}$$

Substituting into (B12) and setting $p_{\mathrm{ch},2} = p_{\mathrm{ch},1} \equiv p_{\mathrm{ch}}$ yield

$$\frac{d\theta_{1,2}^{\mathrm{ch}}}{dp_{\mathrm{ch},2}} = -\frac{(1 - p_{\mathrm{ch}})p_{\mathrm{dis}}}{1 + p_{\mathrm{dis}}p_{\mathrm{ch}}}, \quad \frac{d\theta_{2,1}^{\mathrm{ch}}}{dp_{\mathrm{ch},2}} = -\frac{1}{1 + p_{\mathrm{dis}}p_{\mathrm{ch}}}, \quad \text{(B14)}$$

and therefore

$$p_{\mathrm{dis}}^* = \frac{c}{b\,(1 - p_{\mathrm{ch}})}. \quad \text{(B15)}$$

3. Group III strategies

For the strategies of this group the probabilities of cooperation and their derivatives are also given by Eqs. (B11) and (B12), and the probabilities $P_{i,\alpha\beta}^{\mathrm{ch}}$ are given by (B5) as well as $P_{i,00}^{\mathrm{ch}} = 0$. Thus, after Eqs. (4) and (A3),

$$x_{1,H}^{\mathrm{ch}} = 1 - p_{\mathrm{dis}}p_{\mathrm{ch},1}, \qquad x_2^{\mathrm{ch}} = 1 - p_{\mathrm{dis}}p_{\mathrm{ch},2}. \quad \text{(B16)}$$

Substituting into (B12) and setting $p_{\mathrm{ch},2} = p_{\mathrm{ch},1} \equiv p_{\mathrm{ch}}$ yields

$$\begin{aligned}
\frac{d\theta_{1,2}^{\mathrm{ch}}}{dp_{\mathrm{ch},2}} &= -(1 - p_{\mathrm{ch}})p_{\mathrm{dis}}, \\
\frac{d\theta_{2,1}^{\mathrm{ch}}}{dp_{\mathrm{ch},2}} &= -(1 - p_{\mathrm{dis}}p_{\mathrm{ch}}),
\end{aligned} \quad \text{(B17)}$$

and therefore

$$p_{\mathrm{dis}}^* = \frac{c}{c\,p_{\mathrm{ch}} + b\,(1 - p_{\mathrm{ch}})}. \quad \text{(B18)}$$

[1] E. Fehr and U. Fischbacher, Nature (London) **425**, 785 (2003).

[2] R. L. Trivers, Q. Rev. Biophys. **46**, 35 (1971).

[3] R. Sugden, *The Economics of Rights, Co-operation and Welfare* (Basil Blackwell, Oxford, 1986).

[4] R. D. Alexander, *The Biology of Moral Systems* (Aldine de Gruyter, New York, 1987).

[5] M. Dufenberg, U. Gneezy, W. Güth, and E. Van Demme, Homo Oeconomicus **18**, 19 (2001).

[6] M. Milinski, D. Semmann, and H. J. Krambeck, Nature (London) **415**, 424 (2002).

[7] K. Panchanathan and R. Boyd, Nature (London) **432**, 499 (2004).

[8] D. Semmann, H.-J. Krambeck, and M. Milinski, Behav. Ecol. Sociobiol. **56**, 248 (2004).

[9] S. Suzuki and E. Akiyama, J. Theor. Biol. **245**, 539 (2007).

[10] R. Bshary and A. S. Gutter, Nature (London) **441**, 975 (2006).

[11] G. E. Bolton, E. Katok, and A. Ockenfels, Manage. Sci. **50**, 1587 (2005).

[12] C. Keser, IBM Syst. J. **43**, 498 (2002).

[13] R. Boyd and P. J. Richerson, Social Networks **11**, 213 (1989).

[14] M. A. Nowak and S. Roch, Proc. R. Soc. B **274**, 605 (2007).

[15] M. A. Nowak and K. Sigmund, Nature (London) **393**, 573 (1998).

[16] H. Ohtsuki and Y. Iwasa, J. Theor. Biol. **231**, 107 (2004).

[17] H. Brandt and K. Sigmund, J. Theor. Biol. **231**, 475 (2004).

[18] M. A. Nowak and K. Sigmund, Nature (London) **437**, 1291 (2005).

[19] O. Leimar and P. Hammerstein, Proc. R. Soc. London, Ser. B **268**, 745 (2001).

[20] J. M. Pacheco, A. Traulsen, and M. A. Nowak, Phys. Rev. Lett. **97**, 258103 (2006).

[21] H. Ohtsuki and Y. Iwasa, J. Theor. Biol. **244**, 518 (2007).

[22] H. Ohtsuki and Y. Iwasa, J. Theor. Biol. **239**, 435 (2006).

[23] S. Uchida and K. Sigmund, J. Theor. Biol. **263**, 13 (2010).

[24] L. A. Martinez-Vaquero, J. A. Cuesta, and A. Sánchez, PLoS One **7**, e35135 (2012).

[25] K. Panchanathan and R. Boyd, J. Theor. Biol. **224**, 115 (2003).

[26] M. A. Fishman, J. Theor. Biol. **225**, 285 (2003).

[27] A. Lotem, M. A. Fishman, and L. Stone, Nature (London) **400**, 226 (1999).

[28] W. D. Hamilton, J. Theor. Biol. **7**, 1 (1964).

[29] L. A. Orquera and E. L. Piana, *La vida material y social de los Yámana* (Eudeba-IFIC, Buenos Aires, 1999).

[30] S. Tanabe, H. Suzuki, and N. Masuda, J. Theor. Biol. **317**, 338 (2013).

[31] M. Nakamura and N. Masuda, PLoS Comput. Biol. **7**, e1002113 (2011).

[32] J. Hofbauer and K. Sigmund, *Evolutionary Games and Population Dynamics* (Cambridge University Press, Cambridge, 1998).