Promises and Endogenous Reneging Costs*

Yuval Heller† David Sturrock‡

September 19, 2017

Abstract

We present a novel theoretical mechanism that explains the capacity for non-enforceable communication about future actions to improve efficiency. We explore a two-player partnership game where, before choosing a level of effort to exert on a joint project, each player makes a cheap talk promise to their partner about their own future effort. We allow agents to incur a psychological cost of reneging on their promises. We demonstrate a strong tendency for evolutionary processes to select agents who incur intermediate costs of reneging, and show that these intermediate costs induce second-best optimal outcomes.

Keywords: Promises, lying costs, joint projects, input games, partnerships.

JEL Classification: C73, D03, D83.

1 Introduction

Communication about future actions in joint projects is pervasive in the household, within and between firms, in political processes, and in casual day-to-day interactions. Often, agents can make statements about their intentions, both as a means of coordination and as a promise. Frequently, they are not contractually bound by these statements and have an incentive to make false promises and renege upon them when choosing how to act. Nevertheless, agents in such circumstances commonly use communication to carry out courses of action that yield a higher payoff to each than would be expected if agents could make and break promises at no direct cost

*We thank Vincent Crawford, Peyton Young, Nils Rochowicz, Rachel Griffith, and seminar audiences at the University of Oxford, Bar Ilan University, and the Institute for Fiscal Studies for valuable discussions and suggestions. This paper is based on David Sturrock's M.Phil. thesis, which was submitted at the University of Oxford under the supervision of Yuval Heller and Peyton Young.

†Department of Economics, Bar Ilan University. yuval.heller@biu.ac.il. URL: https://sites.google.com/site/yuval26/. The author is grateful to the European Research Council for its financial support (starting grant #677057).

‡Institute for Fiscal Studies, david_s@ifs.org.uk. The author is grateful for funding from the Economic and Social Research Council (award #ES/J500112/1) and Lincoln College, University of Oxford.

(cheap talk). Consider, for example, two coauthors initiating a project and making promises about the number of hours they will separately work on it in the following year, or countries making commitments to reduce regional levels of pollution.

Our two key contributions are as follows. First, we present a novel theoretical foundation of the prevalence of intermediate psychological costs of breaking promises (reneging). Second, we demonstrate that these endogenously determined intermediate psychological costs yield second-best optimal outcomes in an important class of strategic interactions. Taken together, these contributions present a novel explanation for the way in which pre-play communication can foster cooperation in one-shot strategic interactions when agents' interests are only partially aligned.

Much of the existing literature on signaling intentions through cheap talk explores the potential for pre-play communication to select among multiple equilibria by breaking symmetries, offering assurance, and creating a focal point for play (for a theoretical discussion, see Farrell, 1988; Farrell & Rabin, 1996; for experimental evidence see Crawford, 1998; Charness, 2000). However, extensive experimental evidence shows that communication can also lead players to coordinate on mutually beneficial but non-equilibrium outcomes (Kerr & Kaufman-Gilliland, 1994; Sally, 1995; Ellingsen & Johannesson, 2004; Bicchieri & Lev-On, 2011). In particular, Charness & Dufwenberg (2006) find that players make and keep promises to cooperate in two-player partnership games where the unique subgame perfect equilibrium involves no such cooperation. Vanberg (2008) presents evidence that this behaviour is driven by an aversion to going back on one's word.

The possibility of repeated interaction with a partner means that reputational concerns could motivate agents to keep their promises, even when this does not maximise their payoff in the present encounter. However, the aforementioned experiments, and indeed much of daily experience, demonstrate that agents are motivated to some extent to keep their word even in one-off encounters and suggests a direct concern for keeping promises. In this paper, we put reputational concerns to one side and consider this second, direct motivation for promise-keeping.

Model and Main Results

We study a class of partnership games (also known as input games; see, e.g., Holmstrom, 1982; Cooper & John, 1988) with cheap talk pre-play communication. In the setting we examine, agents simultaneously communicate promised levels of effort, and, following this, they simultaneously choose their levels of effort. Agents experience a direct convex cost of their effort, and a benefit which is increasing in both their own effort and that of their partner, such that effort choices are strategic complements. Agents always have an incentive to slightly "undercut" (exert less effort than) their partner so that when talk is cheap, the only subgame perfect equilibrium of

the game involves both agents choosing zero effort. However, this outcome is Pareto-dominated by outcomes in which players exert effort.

We explore the impact of introducing into this setting a direct convex cost of reneging on promises. Specifically, we assume that each agent experiences a convex psychological cost of the distance between their promised and actual effort. This aversion to reneging can be seen as representing the guilt or bad feeling that agents experience when going back on promises they have made. Reneging aversion transforms what is ordinarily modeled as a cheap talk promise into a *partially* self-enforcing commitment. We show that positive effort can be sustained in equilibrium either when one player has a high level of reneging aversion and her partner a low level of reneging aversion, or when both have an intermediate level of reneging aversion. In the asymmetric case, the player with the high level of reneging aversion effectively plays like a Stackelberg leader, and the partner effectively responds like a Stackelberg follower. In the symmetric case in which both players have an intermediate level of reneging aversion, both players promise to exert maximal effort in the first stage, and then shirk slightly by under-delivering in the second stage.

A novel insight of the model is to show that two players with high levels of aversion to reneging on promises may fail to maintain positive effort in equilibrium. The intuition for this result is that a high level of reneging aversion renders a promise to make a positive effort credible, which in turn implies an incentive for each player to slightly undercut her partner when making promises in the first stage, which in turn makes a player unwilling to promise a positive level of effort in the first place. By contrast, players with intermediate levels of reneging aversion sustain positive effort as each is sufficiently committed to her promise that her partner can reciprocate a promise of high effort without fear of being severely undercut in the second stage; but, at the same time, each partner knows that if she tries to "shirk" and promise low effort, her partner has sufficient flexibility to respond with a lower choice of effort in the second stage, making this "shirking" unprofitable.

We then study the endogenous determination of this reneging aversion in an evolutionary framework. Our main result demonstrates the strong tendency for evolutionary processes to select for agents who incur intermediate psychological reneging costs. Specifically, we show that when players can observe their partner's level of reneging aversion with a sufficiently high probability, there is a unique stable population state in which all players have the same intermediate level of reneging aversion, and the induced equilibrium effort is a second-best outcome (i.e., it is optimal under the constraint of being consistent with equilibrium behaviour). Finally, we show that a weaker version of this result holds also when players can observe their partner's level of reneging aversion with a low, yet positive, probability. In this latter case, we show that in any stable state players must have positive reneging aversion and exert positive effort in equilibrium.

Related Literature and Contributions

This paper contributes to several strands of literature. The first is the theoretical work incorporating exogenously given (and, typically, small) psychological lying costs into strategic models. Kartik *et al.* (2007) and Kartik (2009) study sender-receiver games in which the informed agent has an incentive to distort the receiver's belief, and incurs a convex cost of sending a false message. Matsushima (2008) and Kartik *et al.* (2014) introduce arbitrarily small lying costs into settings of mechanism design and implementation. The present paper moves beyond the existing literature in three key dimensions. Firstly, we explore bilateral communication. Secondly, we interpret players' messages as a report about their own future actions rather than some exogenously given state of the world. These two aspects add further strategic dimensions to the partnership game. Thirdly, we endogenise the reneging costs, and allow them to be determined as part of a stable population state.[1]

With this focus on commitment to future action, we bring together the literature on lying costs and that on partnership games with strategic complimentarities. Games in which n players experience a common outcome, which is increasing in a privately costly action, are examined from a mechanism design perspective in Holmstrom (1982). Radner *et al.* (1986) analyse a two-player partnership game in which a project succeeds with a probability equal to the minimum of the players' effort choices, which are made at quadratic cost, and show the capacity for repeated interaction to sustain effort when such an outcome is efficient but is not an equilibrium of the one-shot game (see also related models of partnership games in Cooper & John, 1988; Admati & Perry, 1991; Cahuc & Kempf, 1997; Marx & Matthews, 2000). We demonstrate that reneging costs is a new means by which cooperation can be sustained in partnerships in one-off encounters with non-enforceable effort choices.

The role of commitment in strategic situations has been extensively investigated since the seminal work of Schelling (1980) (see, e.g., Caruana & Einav, 2008; Ellingsen & Miettinen, 2008; Heller & Winter (2016); and the references in them for recent papers in this vast literature). One of the main stylised insights of this literature is that the ability to commit is advantageous to a player and that, typically, a better ability to commit yields higher payoffs. Our model yields the insight that too great a capacity for commitment (i.e., too high a level of reneging aversion) might be detrimental. Specifically, we show that there is an intermediate level of commitment that is optimal for an agent, as it balances their interest in making a strong commitment in order to induce high effort from their partner, against their conflicting desire to retain some flexibility

[1] Demichelis & Weibull (2008) study the influence of the introduction of lexicographic reneging costs into a setup in which players communicate before playing a coordination game. They show that the introduction of these lying costs implies that the unique evolutionarily stable outcome is Pareto efficient. Heller (2014) shows that this sharp equilibrium selection result is implied by the discontinuity of preferences, rather than by small lying costs *per se*.

to exert less effort.

We explore not only the *consequences* of an aversion to reneging but also give a theoretical exploration of its possible evolutionary determinants. In doing so, we build on the "indirect" evolutionary approach, which studies the evolution of non-material preferences, that was pioneered by Güth & Yaari (1992), and developed by, among others, Ok & Vega-Redondo, 2001; Guttman, 2003; Dekel *et al.*, 2007; Herold & Kuzmics, 2009; Alger & Weibull, 2010, 2012. We make two main contributions to this literature. First, to the best of our knowledge, we are the first to apply the indirect evolutionary approach to study reneging costs. Second, our main result is qualitatively different from the stylised result in the existing literature, according to which if preferences are observed with high probability, then the Pareto efficient outcome is played in any stable population state. We show that in the setup in which the set of feasible preferences is the set of levels of reneging aversion, evolutionary forces take the population into stable states in which agents have intermediate reneging aversion and the agents achieve partial, rather than full, efficiency.

Heifetz *et al.* (2007b) study payoff-monotonic selection dynamics in normal-form games in which the set of strategies of each player is an open subset of $\mathbb{R}^n$ and preference "distortions" (divergences between the subjective utility function and the material payoff function) are perfectly observable. They show that in almost every such game and for almost every family of distortions of a player's actual payoffs, some degree of distortion is beneficial to the player, and will not be driven out by any evolutionary process in the sense that there will not be a convergence to a population in which everyone has zero distortion. Heifetz *et al.* (2007a) make additional assumptions: (1) the set of actions of each player is an interval in $\mathbb{R}$, (2) the underlying game has a unique pure equilibrium for each pair of distortions, and (3) the type game is dominance solvable. Under these assumptions the authors show that the selection dynamics converge to every player having the same distorted type, and that this result can be extended to a setup with partial observability. The game studied in this paper does not satisfy these additional assumptions (in particular, the set of strategies of the normal-form game is infinite-dimensional). Nevertheless, we are able to show results that are consistent with the results of Heifetz *et al.* (2007a) and, in addition, to explicitly characterize the unique stable level of reneging aversion.

Finally, by demonstrating the significance and evolutionary stability of an aversion to reneging in partnership contexts, we provide a theoretical grounding and explanation of the mass of experimental evidence suggesting that "non-standard" preferences play an important role in communication contexts, and that most people incur some psychological costs of lying, and that these costs are increasing in the size of the lie (see Abeler *et al.*, 2016, for a recent meta-study of a large number of lying experiments). For example, Shalvi *et al.* (2011) and Fischbacher & Föllmi-Heusi (2013) find that subjects do not always lie to gain money, even when their doing so cannot be detected. Significantly less than "full" lying is also found in sender-receiver contexts

(Gneezy, 2005; Hurkens & Kartik, 2009), bargaining games (Lundquist *et al.*, 2009), and hold-up games (Ellingsen & Johannesson, 2004), with some studies finding evidence of lying aversion *per se* (Sánchez-Pagés & Vorsatz, 2007).

The paper is organised as follows. Section 2 sets out the partnership game and analyses its equilibria. Section 3 formally defines the evolutionary model and presents our main result about the stability of intermediate levels of reneging aversion. Section 4 demonstrates the robustness of this result to partial observability. In Section 5 we discuss the significance and interpretation of our results and indicate directions for further research. In general, we confine formal proofs to appendices, with exceptions where the proof is brief and aids intuition.

2 The Partnership Game

In this section, we formally describe the partnership game and analyse the subgame perfect equilibria of encounters between any two players with weakly positive aversion to reneging on promises.

2.1 The Model

There are two players (i and j) and two stages of the game. In the first stage, both players simultaneously send a message $s_k \in [0,1] \equiv S$ to their opponent (where $k = i, j$). The interpretation is that players' messages take the form of a promise about effort in the second stage. In the second stage, players simultaneously choose their level of effort, $x_k \in S$.

Remark 1. For simplicity, we define the maximum message (and level of effort) to be one. All of our results remain qualitatively the same, for any other upper limit $M > 0$ to the set of messages.

For a given outcome of the game, we define the "material payoff" to player i as follows (player j's material payoff is defined analogously):

$$V_i(x_i, x_j, c) = x_i x_j - \frac{c x_i^2}{2} \quad : \quad c > 1 \tag{1}$$

The interpretation of the material payoff is as follows. Both players receive the same gross return from the partnership, equal to the product of their two effort choices. They each incur a cost proportional to the square of their own effort. The parameter c governs the cost of effort. We focus in our evolutionary analysis on low values of c in the interval of $(1, 1.24)$, as these prove most illuminating.

Player i's subjective utility is defined as follows (player j's subjective utility is defined anal-

ogously):

$$U_i(x_i, x_j, s_i, c) = x_i x_j - \frac{cx_i^2}{2} - \frac{\lambda_i}{2}(s_i - x_i)^2 \quad : \quad c > 1 \tag{2}$$

Subjective utility is the sum of a player's material payoff and a term representing the psychological cost of breaking a promise (reneging). Here, reneging is defined as exerting a level of effort not equal to the message sent (i.e., the effort promised) in the first stage. The "size" of player i's reneging is defined as $|s_i - x_i|$. The utility loss from reneging is proportional to the square of its size, multiplied by λ_i, a parameter that we call i's *level of reneging aversion*. In the following two sections we assume that all players perfectly observe their partner's level of reneging aversion, i.e., that the parameters λ_i, λ_j are common knowledge. In Section 4 we deal with the case of partial observability.

A mixed strategy of player i in the second stage is a distribution $\chi_i \in \Delta(S)$. Let μ_{χ_i} denote the expectation of the distribution. We assume that players are expected utility maximisers. The fact that the utility function U_i depends linearly on the effort of the opponent (x_j) implies that player i's expected utility depends only on the expected effort of the partner (μ_{χ_j}), which replaces x_j in Eq. (2) to yield an expected utility function, i.e.,

$$U_i(x_i, \chi_j, s_i, c) = x_i \cdot \mu_{\chi_j} - \frac{cx_i^2}{2} - \frac{\lambda_i}{2}(s_i - x_i)^2 \quad : \quad c > 1 \tag{3}$$

2.2 Unique Second-Stage Equilibrium

In the second stage of the game, player i's first-order condition for her choice of x_i is given by[2]

$$\mu_{\chi_j} - cx_i + \lambda_i(s_i - x_i) = 0 \tag{4}$$

The concavity of the utility function in x_i implies that the second-stage best response is a unique pure strategy, given by the function

$$x_i^*(\chi_j, s_i, s_j, \lambda_i, \lambda_j, c) = \frac{\mu_{\chi_j} + \lambda_i s_i}{c + \lambda_i} \tag{5}$$

This equation embodies a player's (possibly conflicting) desires to undercut (exert less effort than) their opponent and to minimise their reneging.

Fact 1. *We first observe that when $\lambda_i = \lambda_j = 0$ (i.e., both players' messages are cheap talk) the best response of player i reduces to $\frac{\mu_{\chi_j}}{c}$. This implies that when talk is cheap, both players wish to undercut their opponent in the second stage, effort choices are independent of messages sent,*

[2]The second derivative of the utility function with respect to x_i is $-c-\lambda_i$. The fact that it is always negative guarantees that the solution to the first-order condition is a global maximum of the utility function and that the optimal choice in the second stage is a unique pure strategy.

and in all subgame perfect equilibria, neither player exerts effort and communication plays no committing role.

To consider the general case of positive reneging costs, we solve the best-response functions simultaneously and obtain the unique Nash equilibrium strategy for player i in the subgame induced by an arbitrary pair of messages s_i and s_j:

$$x_i^e(s_i, s_j, \lambda_i, \lambda_j, c) = \frac{(c+\lambda_j)\lambda_i s_i + \lambda_j s_j}{(c+\lambda_i)(c+\lambda_j) - 1} \tag{6}$$

To gain some intuition, we can consider the subgame after $s_i = s_j = s$ is played. In this case, $x_i < x_j \iff \lambda_i < \lambda_j$. Both players have an incentive to undercut one another (and by implication renege on their own first-stage promises), but they also do not want to incur too great a cost from reneging. Due to the convex cost of reneging and the diminishing material gains from reducing effort towards $\frac{x_j}{c}$, the optimal choice of x_i balances these two aims. In the general case where $s_i \neq s_j$, the Nash equilibrium choice of x_i is some convex combination of s_i, s_j,[3][4] and 0. As a player's level of reneging aversion increases, she will exert effort closer to her own promise.

2.3 First-Stage Best-Response Functions

The subgame perfect equilibrium of the game is easily obtained using backwards induction. Given the unique Nash equilibrium strategies in each subgame, we can find a player's optimal choice of message given her opponent's choice. Taking the choices of effort in each subgame as given, a player i's utility as a function of first-stage messages is given by

$$\begin{aligned} U_i(s_i, s_j, c) =& \frac{[(c+\lambda_j)\lambda_i s_i + \lambda_j s_j][(c+\lambda_i)\lambda_j s_j + \lambda_i s_i]}{[(c+\lambda_i)(c+\lambda_j) - 1]^2} \\ & - \frac{c[(c+\lambda_j)\lambda_i s_i + \lambda_j s_j]^2}{2[(c+\lambda_i)(c+\lambda_j) - 1]^2} - \frac{\lambda_i}{2}\left[s_i - \frac{(c+\lambda_j)\lambda_i s_i + \lambda_j s_j}{(c+\lambda_i)(c+\lambda_j) - 1}\right]^2 \end{aligned} \tag{7}$$

A mixed strategy of a player at the first stage is a distribution $\sigma_i \in \Delta(S)$. Let μ_{σ_i} denote the expectation of the distribution. Observe that the utility function U_i can be presented as a sum of two functions: (1) a linear function of s_j and (2) an expression that is independent of s_i. This implies that the best-reply function of player i against a partner who plays a mixed strategy σ_j depends only on the partner's expected message μ_{σ_j} (for reasons analogous to those in the argument for μ_{χ_j}above).

[3] To see this, observe that the denominator of the fraction is strictly positive and strictly greater than the sum of the coefficients on s_i and s_{-i} in the numerator.

[4] This guarantees that $x_i^e, x_j^e \in [0, 1)$ and therefore the first-order condition always characterises optimal choice.

When $\lambda_i = 0$, player i's choice of message has no bearing on her optimal effort choice or that of her opponent, and therefore does not impact her utility. Therefore any message is a best response to any message sent by her opponent. When $\lambda_i > 0$, the first derivative of player i's utility function with respect to s_i, taking μ_{σ_j} as given, is a linear function of s_i and μ_{σ_j}:

$$\frac{\partial U_i(s_i, \mu_{\sigma_j}, c)}{\partial s_i} = \left[2 - c(c + \lambda_j) - \frac{1}{(c + \lambda_i)(c + \lambda_j)}\right] s_i + \lambda_j \cdot \mu_{\sigma_j} \tag{8}$$

For ease of exposition, we define Θ_i to be the negative of the coefficient on s_i in Eq. (8):

$$c(c + \lambda_j) + \frac{1}{(c + \lambda_i)(c + \lambda_j)} - 2 \equiv \Theta_i$$

Given that λ_j and μ_{σ_j} are constrained to be (weakly) positive, the second term in Eq. (8) is also (weakly) positive. Therefore, when $\Theta_i > 0$ (and hence the term multiplying s_i in Eq. (8) is strictly negative), the utility function is everywhere strictly concave in s_i, and the following level of s_i, which is positive and satisfies the first-order condition $\frac{\partial U_i(s_i, \mu_{\sigma_j}, c)}{\partial s_i} = 0$, is a necessary and sufficient condition for a global maximum of the utility function:

$$s_i(\mu_{\sigma_j}, \lambda_i, \lambda_j, c) = \frac{\lambda_j}{\Theta_i} \cdot \mu_{\sigma_j} \tag{9}$$

Further, the strict concavity of the utility function in s_i means that when $\frac{\lambda_j}{\Theta_i} \cdot \mu_{\sigma_j} > 1$, the optimal choice of s_i is 1.

When $\Theta_i < 0$ (and hence the term in s_i in Eq. (8) is strictly positive), the utility function is everywhere strictly increasing and convex in s_i. In this case, the optimal choice of s_i is 1, for all $\mu_{\sigma_j} \in S$. When $\Theta_i = 0$, if $\lambda_j > 0$ and $\mu_{\sigma_j} > 0$, then again the utility function is everywhere strictly increasing and convex in s_i and the optimal choice of s_i is 1. If $\Theta_i = 0$ and either $\lambda_j = 0$ or $\mu_{\sigma_j} = 0$, then the utility function is flat in s_i and any message is a best response to the opponent's message. The best-response correspondence in the first stage can therefore be written as

$$s_i^*(\mu_{\sigma_j}, \lambda_i, \lambda_j, c) = \begin{cases} \frac{\lambda_j}{\Theta_i} \cdot \mu_{\sigma_j} & 0 \leq \frac{\lambda_j}{\Theta_i} s_j \leq 1 \textit{ and } \Theta_i > 0 \\ 1 & \frac{\lambda_j}{\Theta_i} s_j > 1 \textit{ or } \Theta_i < 0 \textit{ or } (\Theta_i = 0 \textit{ and } \lambda_j \cdot \mu_{\sigma_j} > 0) \\ [0, 1] & \Theta_i = 0 \textit{ and } \lambda_j \cdot \mu_{\sigma_j} = 0 \end{cases} \tag{10}$$

The choice of the best reply in the latter "knife-edge" case, in which $\Theta_i = \lambda_j \cdot \mu_{\sigma_j} = 0$ does not play any role in our results. In all other cases, the unique best-reply function of both players always induces them to choose a pure message and, as a result, both players choose pure

messages in all equilibria.

Remark 2. Observe that a player can always guarantee a utility level of zero by playing $s_i = x_i = 0$. Further, observe that if $\Theta_i > 0$ ($\Theta_i < 0$ or $[\Theta_i = 0$ *and* $\lambda_j \cdot \mu_{\sigma_j} > 0]$), then the utility function is strictly concave (strictly increasing) in s_i. This implies that if the best response s_i^* is positive (i.e., $s_i^* > 0$) and unique, then it must yield strictly positive utility for player i.

Players wish to minimise their reneging, undercut their opponent (play close to $\frac{x_j}{c}$), and have their opponent put in as much effort as possible. Their optimal choice will therefore balance these three aims. It is straightforward to see that if a player's choice of message has no impact upon her opponent's choice of effort, she will promise, and deliver, effort that undercuts her opponent. However, while a player knows that she in some sense "ties her hands" if she promises to put in higher effort in the presence of a reneging cost, and restricts her ability to undercut in the second stage, such a promise has a second, strategic effect: because the player's opponent knows that he will not be severely undercut, he is willing to put in more effort in the second stage. This strategic effect is a consequence of the strategic complementarity of effort with respect to the material payoffs.

In a set of games with measure zero, all of these considerations cancel out such that any message is a best response.[5] Otherwise, a player's best response to her opponent's message can be classified as one of three kinds. When the incentive to undercut dominates, a player wants to send a message that is some fraction (less than 1) of her opponent's message. When the incentive to strategically commit to high effort dominates, a player wants either to send a message that is some multiple (greater than 1) of her opponent's message or to send the maximum possible message in all cases. Whether a player optimally chooses to undercut her opponent or to strategically commit to high effort depends only on the level of c and the players' reneging costs and is invariant to the commitment made by her opponent.

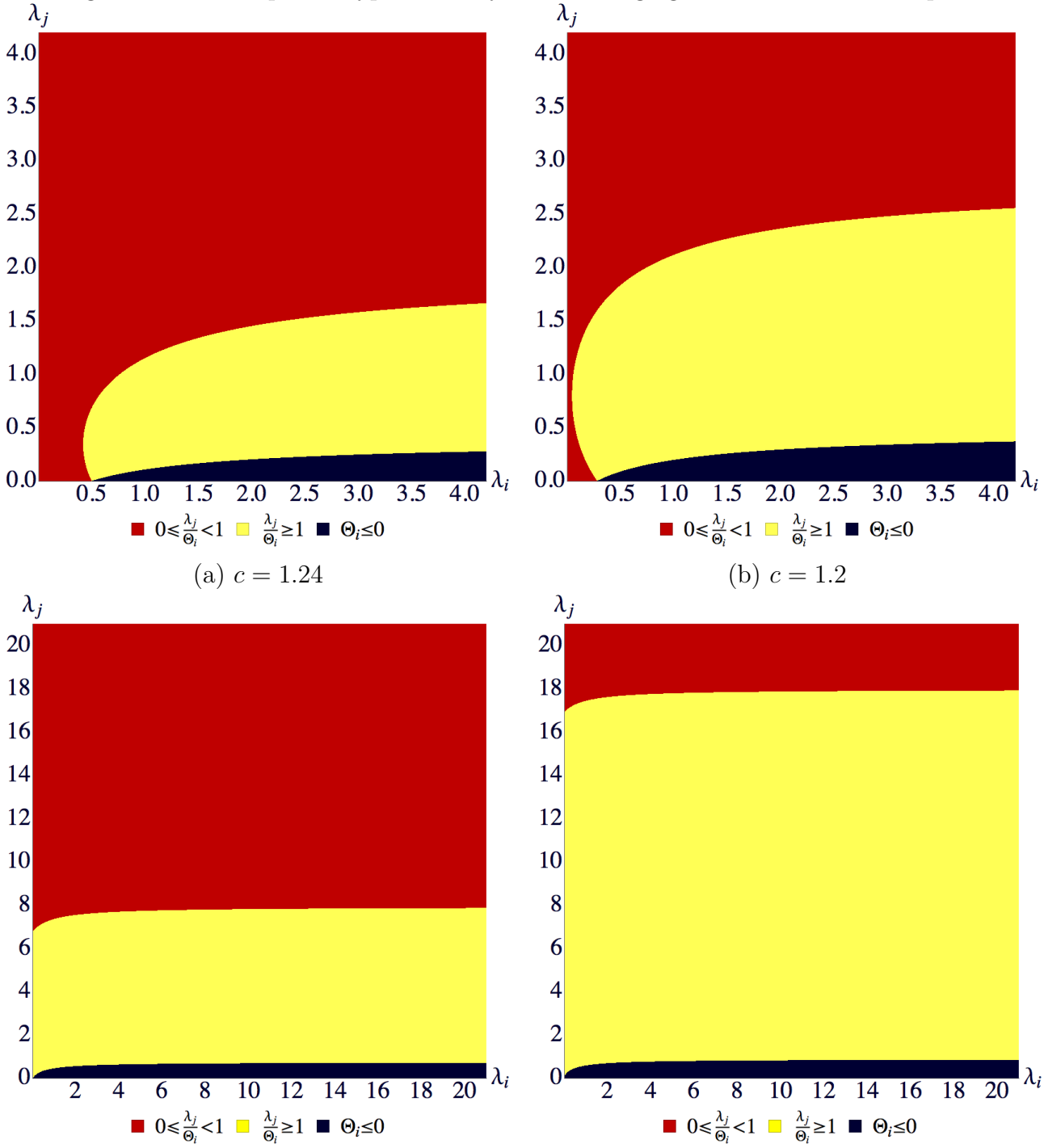
The division of the parameter space into these best-response types is illustrated in Figure 1. The higher a player's reneging aversion is, the more able she is to make a credible, strategic commitment. Such a strategic commitment is more worthwhile when facing a lower cost of effort (lower c) and a player with a lower reneging aversion (who is consequently more responsive to commitments made). Therefore a player will strategically promise high effort when λ_i is sufficiently high and λ_j and c sufficiently low. Lemmas 1, 2, and 3 in Appendix C.1 constitute a full set of necessary and sufficient conditions for a player's best response to be of each type.

2.4 Unique Perfect Equilibrium

We now present the subgame perfect equilibria of the partnership game. All subgame perfect equilibria of the partnership game can be classified as one of three types, set out below. In some

[5] This is the set of cases covered by the third row of the best-response correspondence; see Eq. (10).

Figure 1: Best-Response Types for Player i in Reneging Aversion Parameter Space



(a) $c = 1.24$

(b) $c = 1.2$

(c) $c = 1.1$

(d) $c = 1.05$

(Note that the upper figures focus on the interval $\lambda_i\lambda_j \in [0, 4]$, while the lower figures show the wider interval $\lambda_i\lambda_j \in [0, 20]$.)

parameterisations of the game, the subgame perfect equilibrium is unique. In the remaining set of cases, the game admits two subgame perfect equilibria but one is not a "trembling-hand" perfect equilibrium (see the formal definition in Appendix A), and, thus, we do not consider it as a plausible prediction of play. The unique equilibrium that satisfies the trembling-hand perfection refinement can be classified as one of three types, and its classification depends only on a single pair of parameters (one for each player). For each player i, we define the variable R_i as follows:

$$R_i \equiv \begin{cases} \dfrac{\lambda_j}{\Theta_i} & \Theta_i > 0 \\ \infty & \Theta_i \leq 0 \end{cases}$$

R_j is defined analogously. Each partnership game maps to a pair (R_i, R_j) that corresponds to one type of unique perfect equilibrium with first-stage play as follows:

1. $\min(R_i, R_j) > 1 \Leftrightarrow s_i = s_j = 1$: "Maximum message" equilibrium. This equilibrium arises when both players wish either to send a higher message than their opponent or to send the maximum possible message.

2. $R_i \cdot R_j > 1 > R_j \Leftrightarrow s_i = 1 > s_j > 0$: "Two-message" equilibrium. This equilibrium arises when one player wishes to send a message lower than that of her opponent (undercut) and her opponent wishes to strategically induce higher effort in her. When the player trying to induce higher effort wants to do so more than his opponent wishes to undercut, then a perfect equilibrium exists in which the former sends the maximum message and the latter best responds to it.

3. $R_i \cdot R_j < 1 \Leftrightarrow s_j = s_i = 0$: "No-effort" equilibrium. This is the case in which either (1) both players wish to undercut their opponent or (2) one player wants to undercut his opponent to a greater extent than the opponent wishes to send a higher message than that of the player. This implies that the only subgame perfect equilibrium involves both players promising, and exerting, no effort.

In each of these cases we specify only the messages s_i, s_j, as, given the players' messages, their effort choices are uniquely determined by Eq. (6) in all cases except the "knife-edge" case of $\Theta_i = \lambda_j = 0$, discussed below. Formally, we present the unique perfect equilibria that exist in three exhaustive classes of games (the definition of trembling-hand perfection is presented in Appendix A).

Theorem 1. *Assume that* $\lambda_i, \lambda_j > 0$.

1. *If* $\min(R_i, R_j) > 1$, *then there exists a subgame perfect equilibrium in which* $s_i = s_j = 1$. *The set of pairs* (λ_i, λ_j) *for which* $\min(R_i, R_j) > 1$ *is non-empty if and only if* $c < 1.25$.

Moreover, this set is symmetric, convex, and bounded away from the origin and from infinity in the sense that for all $c < 1.25$ *there exist* $\underline{\lambda}_c, \lambda_c^+ \in (0, \infty)$ *such that* $\min(R_i, R_j) > 1$ *implies that* $\underline{\lambda}_c \leq max(\lambda_i, \lambda_j) \leq \lambda_c^+$.

2. *If* $R_i \cdot R_j > 1 > R_j$*, then there exists a subgame perfect equilibrium in which* $s_i = 1 > s_j > 0$*. The set of pairs* (λ_i, λ_j) *for which* $R_i \cdot R_j > 1 > R_j$ *is non-empty if and only if* $c < \sqrt{2}$*.*

3. *If* $R_i \cdot R_j < 1$*, then there is a unique subgame perfect equilibrium in which* $s_j = s_i = 0$*. The set of pairs* (λ_i, λ_j) *for which* $R_i \cdot R_j < 1$ *is non-empty for all* $c > 1$*.*

Moreover, in cases 1 and 2, the game admits at most one additional subgame perfect equilibrium in which $s_j = s_i = 0$*, and this latter equilibrium fails to satisfy trembling-hand perfection.*

Henceforth, we use the term the *unique perfect equilibrium* to refer to the unique subgame perfect equilibrium that satisfies the trembling-hand perfection refinement when $\lambda_i, \lambda_j > 0$.

In the case where one player has a level of reneging aversion of zero, equilibria exist that are analogous to the three types set out above. Rather than there being a unique perfect equilibrium for each pair of reneging costs, there exists in each case a unique continuum of essentially equivalent subgame perfect equilibria (that are trembling-hand perfect) that differ only in the cheap talk message sent by the player with zero reneging aversion. Within each continuum, effort levels (which are described by Eq. (6)), subjective utilities, and material payoffs are the same in all equilibria. The formal characterisation of equilibria in the case where one player has zero reneging aversion is given in Appendix B. Recall that the case in which $\lambda_i = \lambda_j = 0$ is dealt with in Fact 1.

Multiple perfect equilibria occur only on a "measure zero" of pairs of λ_i, λ_j that satisfy the equality $R_i \cdot R_j = 1$. Assume without loss of generality that $R_j \leq 1 \leq R_i$. In such cases, for any $s_i \in [0, 1]$, there exists a perfect equilibrium in which the messages are $\left(s_i, s_j = \frac{\lambda_i}{\Theta_j} \cdot s_i\right)$ (the argument is analogous to those in Theorem 1 above, and is omitted for brevity).

We formalise one corollary of Theorem 1, which says that if players reneging costs are identical and positive, they send the same message in the unique perfect equilibrium.

Corollary 1. *Let* $\lambda_i = \lambda_j > 0$*. Then the equality* $s_i = s_j$ *holds in the unique perfect equilibrium of the partnership game*

Proof. For $\lambda_i, \lambda_j > 0$, Theorem 1 shows that the only cases (those covered by point 2) where $s_i \neq s_j$ are such that $R_i \cdot R_j > 1 > R_j$. This implies that $R_i \neq R_j$. From the definition of Θ_i, we see that $\lambda_i = \lambda_j \Rightarrow \Theta_i = \Theta_j$. From the definition of R_i, we see that $\lambda_i = \lambda_j$ and $\Theta_i = \Theta_j$ together imply that $R_i = R_j$. Therefore $\lambda_i = \lambda_j \Rightarrow R_i = R_j$, which implies that $s_i = s_j$. □

Figure 2 illustrates, for a range of values of c, the division of reneging aversion parameter space into the three classes of unique equilibria. When both λ_i and λ_j are high or when both are low, the unique equilibrium is a no-effort equilibrium. When one player has a high level of reneging aversion and the other a low level, the unique equilibrium is a two-message equilibrium. Finally, if both players' levels of aversion to reneging are intermediate (and sufficiently similar) then we have the maximum message equilibrium. Here, both players are sufficiently bound by their message so that they will be able to strategically induce high effort in their partner, but are also flexible enough to respond to their partner's promise. A full set of necessary and sufficient conditions for the existence of each type of unique equilibrium, in terms of the parameters of the game, can be derived by combining the conditions presented in the Lemmas 1, 2, and 3 in Appendix C.1.

3 Evolution of Observable Reneging Costs

In this section we endogenise reneging costs, and present a static model to study the evolution of these costs in a setup in which both players observe their partner's level of reneging aversion.

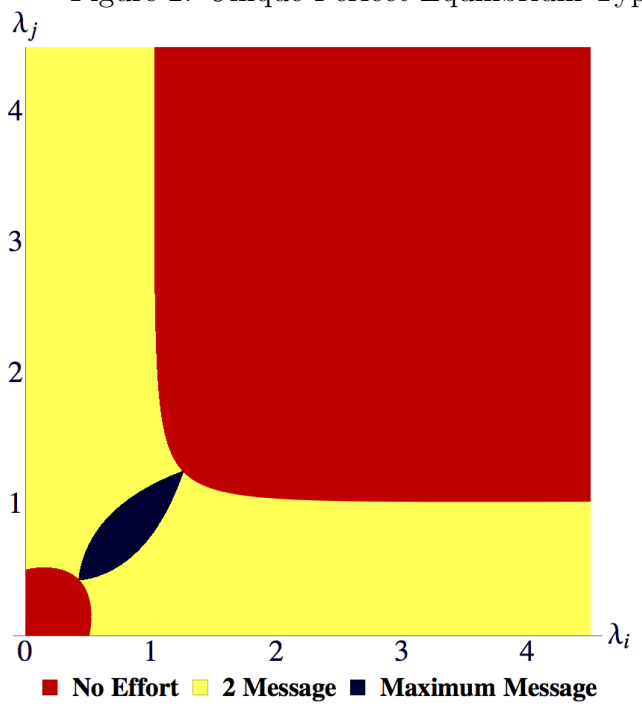
3.1 Population Game

We consider a large population of players (technically, a continuum) in which each player is endowed with a level of reneging aversion. Players are uniformly randomly matched into pairs, and both observe their partner's reneging aversion before starting the two-stage partnership game described above. We assume that in each such partnership game, the players play the unique perfect equilibrium.

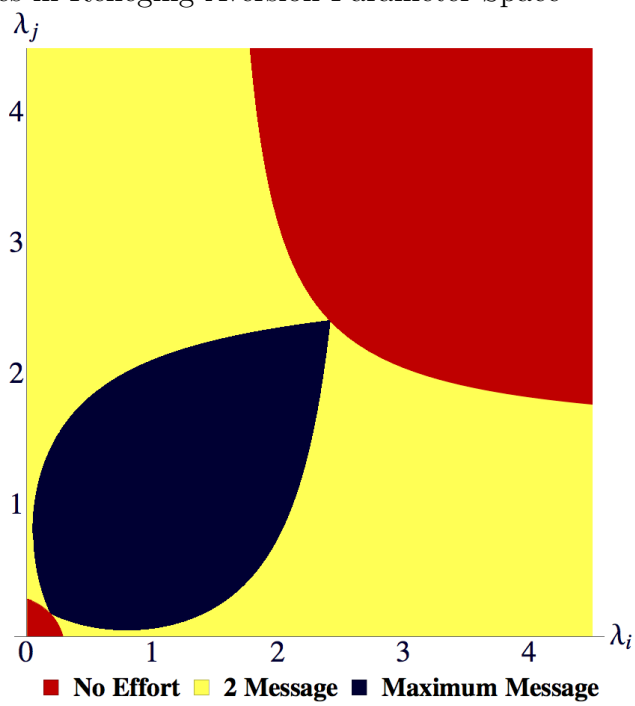
Given levels of reneging aversion $\lambda, \lambda' \in \mathbb{R}^+$ with $R_i \cdot R_j \neq 1$ and cost of effort c, let $G_c(\lambda, \lambda')$ denote the partnership game where players with reneging aversion λ and λ' meet and let $\pi_c(\lambda, \lambda')$ denote the *material* payoff of the player with reneging aversion λ, given that they play the unique perfect equilibrium of the partnership game (or any equilibrium in the unique continuum of perfect equilibria when $\lambda_i = 0$ or $\lambda_j = 0$). To simplify notation, where this does not lead to ambiguity, we omit the subscript c and write $G(\lambda, \lambda')$ and $\pi(\lambda, \lambda')$.

Given symmetric levels of reneging aversion $\lambda = \lambda' \in \mathbb{R}^+$ inducing $R_i = R_j = 1$ and hence multiple perfect equilibria, we assume that the players play the most efficient equilibria, in which they both send the maximum message, i.e., $s_i = s_j = 1$, and we let $\pi(\lambda, \lambda)$ be the material payoffs in these maximum message equilibria. Given levels of reneging aversion $\lambda \neq \lambda' \in \mathbb{R}^+$ inducing $R_i \cdot R_j = 1$ and multiple perfect equilibria, we can assume any arbitrary equilibrium selection function (without affecting our results), and we let $\pi(\lambda, \lambda')$ be the material payoff in these arbitrarily selected equilibria.

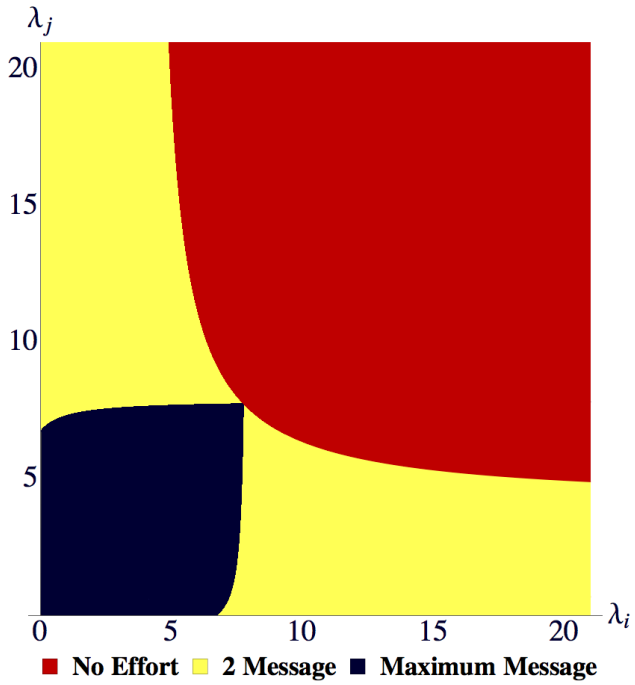
Figure 2: Unique Perfect Equilibrium Types in Reneging Aversion Parameter Space



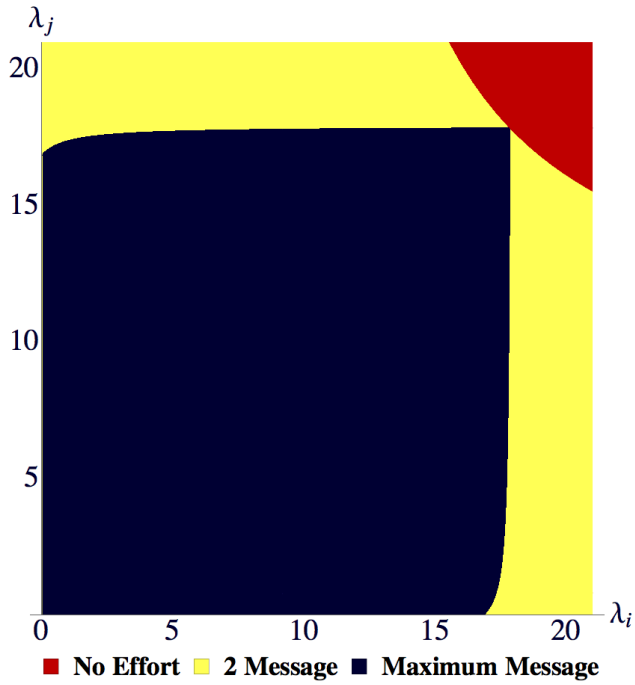
(a) $c = 1.24$



(b) $c = 1.2$



(c) $c = 1.1$



(d) $c = 1.05$

(Note that the upper figures focus on the interval $\lambda_i\lambda_j \in [0, 4]$, while the lower figures show the wider interval $\lambda_i\lambda_j \in [0, 20]$.)

Remark 3. The assumption that the most efficient equilibrium is chosen in $G(\lambda, \lambda)$ when $R_i = R_j = 1$ is motivated as follows. In the model, the set of feasible levels of reneging aversion is a continuum. We consider this to be an approximation for dynamic environments in which the set of feasible levels of reneging aversion is discrete due to either: (1) a finite, albeit very large, set of feasible genotypes in a biological evolutionary process, or (2) some constraints in social evolutionary processes that imply that only a finite number of levels of reneging aversion may be selected; for example, the reneging aversion could represent some rule of thumb that induces a trade-off between keeping promises and making opportunistic gains, where the set of feasible simple rules that agents may adopt is finite. With relatively simple adaptations to the arguments in the main result below (Theorem 2), it can be shown that in such discrete environments, the evolutionary forces will take the population into a homogeneous state in which all agents have the highest level of reneging aversion, λ, that is below λ_c^+. In the game $G(\lambda, \lambda)$, where players have such a level of reneging aversion, $R_i = R_j$ will be slightly above one, and the unique perfect equilibrium will be a maximum message equilibrium. In the model, we wish to abstract away from formally defining the discreteness of the set of feasible types, and thus in Theorem 2 we obtain convergence to λ_c^+, inducing $R_i = R_j = 1$ and multiple equilibria in the game $G(\lambda_c^+, \lambda_c^+)$. We interpret the selection of the maximum message equilibrium as corresponding to the equilibrium of a more elaborate discrete model in which the slightly lower level of λ is chosen.

The payoff function $\pi : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ defines a symmetric two-player population game $\Gamma = (\mathbb{R}^+, \pi)$. A pure (mixed) strategy in this game corresponds to a level of reneging aversion (a distribution over levels of reneging aversion).

It is well known that stable population states correspond to symmetric equilibria of the population game, given a smooth and payoff-monotone dynamic process by which the levels of reneging aversion evolve, such as the replicator dynamics (Taylor & Jonker, 1978; see Weibull, 1995; Sandholm, 2010 for a textbook introduction). Specifically:

1. Any symmetric strict equilibrium corresponds to a stable population state in which all the incumbents have the same reneging aversion. Any agent who is endowed with a different level of reneging aversion (due to random error or experimentation) is strictly outperformed, and is assumed to be eliminated from the population. The same holds for any sufficiently small group of "mutant" agents who are endowed with a different level of reneging aversion. In particular, it is well known that any strict equilibrium is an evolutionarily stable state à la Maynard Smith & Price (1973).

2. Any stable population state must be a symmetric Nash equilibrium (see, e.g., Nachbar, 1990). Otherwise, there is a level of reneging aversion that allows a deviator to strictly outperform the incumbents; we assume that other agents will start to mimic such a successful

deviator, and that the population will move away from the initial state.

Remark 4. As argued by Eshel (1983) and Oechssler & Riedel (2001), strict equilibrium might not be a sufficient condition for dynamic stability in setups in which a small perturbation can slightly change the reneging aversion of all agents in the population. In Section 5 we discuss the notions of stability proposed by these authors, and explain why imposing these more restrictive solution concepts does not affect our results.

3.2 Stability of Intermediate Reneging Aversion

The following result shows that when the cost of effort is sufficiently low (c is low), the population game admits a unique pure strategy Nash equilibrium $(\lambda_c^+, \lambda_c^+)$, which is also a strict equilibrium in which both players promise to exert maximal effort. Moreover, the equilibrium $(\lambda_c^+, \lambda_c^+)$ induces the second-best outcome; i.e., it maximises the sum of material payoffs among all pure strategy equilibria of the partnership game. Finally, λ_c^+ is decreasing in c and converges to ∞ as c converges to one, which implies that the equilibrium material payoffs converge in the limit $c \to 1$ to the first-best outcome in which both players commit to, and exert, maximal effort.

Theorem 2. *Fix $c \in (1, 1.24)$. Then, the population game admits a unique pure symmetric Nash equilibrium $(\lambda_c^+, \lambda_c^+)$. Moreover, (1) the equilibrium $(\lambda_c^+, \lambda_c^+)$ is strict, (2) both players promise to exert maximal effort in the partnership game $G(\lambda_c^+, \lambda_c^+)$, (3) $\pi(\lambda_c^+, \lambda_c^+) > \pi\left(\lambda', \lambda'\right)$ for any $\lambda' \neq \lambda_c^+$, (4) λ_c^+ and $\pi(\lambda_c^+, \lambda_c^+)$ are decreasing in c, and (5) $\lim_{c\to 1} \lambda_c^+ = \infty$, and $\lim_{c\to 1} \pi(\lambda_c^+, \lambda_c^+) = \frac{1}{2}$, which is the first-best payoff.*

Our main result implies that the evolutionary forces move the population into a unique stable state in which all agents have the same intermediate level of reneging aversion λ_c^+, they promise to exert maximal effort, and all interactions yield the second-best outcome. This suggests a strong tendency for evolutionary processes to select this intermediate level of reneging aversion when players each observe their opponent's type.

The stable level of reneging aversion λ_c^+ is the highest level of reneging aversion that induces a maximum message equilibrium when two players with this level of reneging aversion meet and play the partnership game together. Intuitively, this level of reneging aversion is stable because if any mutant with a lower level of reneging aversion were to enter the population, while they would have greater flexibility to undercut a λ_c^+-type partner at the second stage of any encounter, this type of partner would anticipate the undercutting and reduce his effort to such a degree that the mutant would achieve a lower payoff than if she were a λ_c^+-type player. Any alternative mutant with a higher level of reneging aversion would induce a no-effort or two-message equilibrium when meeting the λ_c^+-type players, and so achieve a lower payoff. No

other homogeneous population is stable because either all partnership interactions result in a no-effort equilibrium and yield both players a payoff of zero – in which case there is always some alternative type that could enter the population and achieve a positive payoff (this fact is proven in Lemma 4) – or all partnership interactions are maximum message equilibria – in which case there exists a type of player with a higher level of reneging aversion whose interactions with incumbent-type players result in a maximum message equilibrium (this fact is proven in Lemma 5), and yield them a higher payoff than the incumbent-type players achieve when playing against themselves.

4 Partial Observability

In this section we extend the model endogenising reneging costs to allow for cases in which players sometimes do not observe their partner's level of reneging aversion.

4.1 Population Game with Partial Observability

In what follows, we describe the adaptations to the model in Section 3.1 required to accommodate partial observability. Let $q \in [0,1]$ denote the fraction of matches in which both players observe their partner's level of reneging aversion. That is, we assume that when the agents are randomly matched into pairs, in a share of q of the pairs, both agents observe their partner's reneging aversion, while in the remaining $1-q$ of the pairs the partners are "strangers," and neither of them observes any information about their partner's reneging aversion. One may interpret the observation of reneging aversion to be the result of obtaining information about a partner's past behaviour (either through direct observation or by communicating with agents who interacted with the partner in the past), and with this interpretation q may represent how likely it is that agents who are matched together have prior information about each other.

For tractability, we make the simplifying assumption that the observations of the two matched agents are perfectly correlated, i.e., that an agent observes her partner's reneging aversion if and only if the partner observes the agent's reneging aversion (similar to the model of partial observability in Heifetz *et al.*, 2007a), while leaving the extension to more general observation structures for future research.

Consider a setup in which the incumbent agents have reneging aversion $\lambda \in \mathbb{R}^+$, while occasionally one of the agents is endowed with a different level of reneging aversion (henceforth, a *mutant*). Let $\pi_{no}(\lambda', \lambda|\lambda)$ be the material payoff of a mutant (she) with a reneging aversion of λ' who faces an incumbent partner (he) with a reneging aversion of λ who believes with probability one that his partner has a reneging aversion of λ. Note that this belief is consistent with a situation in which a single mutant experiments with a different level of reneging aversion within

an infinite population of agents. The partner plays his part of the unique perfect equilibrium of the game $G(\lambda, \lambda)$, while the mutant plays her best reply to his strategy.

Given $\lambda, \lambda' \in \mathbb{R}^+$, let $G_q(\lambda, \lambda'|\lambda)$ denote a partnership game between an incumbent with reneging aversion λ and a mutant with reneging aversion λ' in which both players observe their partner's reneging aversion with a probability of q, and neither of them observes their partner's reneging aversion with the remaining probability of $1-q$. In this latter case, both players believe with probability one that the partner has the incumbents' reneging aversion of λ. Let $\pi_q(\lambda', \lambda|\lambda)$ be the mutant's material payoff in $G_q(\lambda, \lambda'|\lambda)$:

$$\pi_q(\lambda', \lambda|\lambda) = q \cdot \pi(\lambda', \lambda) + (1-q) \cdot \pi_{no}(\lambda', \lambda|\lambda)$$

Observe that when $q = 1$, this coincides with the model of perfect observability in Section 3, while the case of $q = 0$ corresponds to the non-observability of the partner's reneging aversion.

We say that the level of reneging aversion $\lambda \in \mathbb{R}^+$ is a *symmetric pure (strict) Nash equilibrium* in the population game with partial observability level q if for each $\lambda' \in \mathbb{R}^+$, $\pi_q(\lambda', \lambda|\lambda) \leq \pi(\lambda, \lambda)$ ($\pi_q(\lambda', \lambda|\lambda) < \pi(\lambda, \lambda)$).

As in the case of perfect observability discussed above, stable homogeneous population states correspond to symmetric pure equilibria of the population game. Specifically:

1. Any symmetric strict equilibrium corresponds to a stable homogeneous population state in which all the incumbents have the same level of reneging aversion.

2. Any homogeneous stable population state must be a symmetric Nash equilibrium.

4.2 Robustness of Theorem 2

The following result demonstrates the robustness of Theorem 2 to almost perfect observability. We show that λ_c^+ is a symmetric strict equilibrium for any $q < 1$ that is sufficiently close to one. Formally:

Theorem 3. *Fix $c \in (1, 1.24)$. Then, there exists $\bar{q} \in (0, 1)$ such that $(\lambda_c^+, \lambda_c^+)$ is a strict equilibrium of the population game with observability level q for each $q \in [\bar{q}, 1]$. Moreover, the equilibrium satisfies all properties (1–5) in the statement of Theorem 2.*

4.3 Non-robustness of No-Effort Equilibrium

We recover a central result from the evolutionary literature on the stability of payoff-maximising preferences under anonymity but show that it is not robust to *any* positive probability of correlated observation of types in our model. The following simple result shows that when there

is no observability (i.e., $q = 0$) no effort is exerted in any equilibrium of the population game. Formally:

Proposition 1. *Fix $c \in (1, 1.24)$, and let $q = 0$. In any symmetric pure Nash equilibrium, all agents exert an effort of zero on the equilibrium path, and any agent i with $\lambda_i > 0$ sends a message of zero.*

This result is similar to those in the existing literature that show that when agents are matched uniformly and anonymously (i.e., no observability or assortativity) and the selection dynamics are payoff monotone, then players maximise their material payoffs in any stable population state (see, e.g., Ok & Vega-Redondo, 2001; Dekel *et al.*, 2007).[6]

Next we show that the no-effort equilibrium is not robust to the presence of any arbitrarily low level of observability. In particular, we show that for any arbitrarily small $q > 0$, the agents must exert positive effort on the equilibrium path, which implies that they make positive promises and have a positive level of reneging aversion.

Proposition 2. *Fix $c \in (1, 1.24)$ and $q > 0$. Then, in any symmetric pure Nash equilibrium, players exert positive levels of effort on the equilibrium path.*

This result demonstrates that even with low levels of observability of reneging aversion, evolutionary dynamics will take the population away from any cheap talk state in which players are unable to make and keep promises.

5 Discussion

5.1 Application to Quality Choice in Supply Chains

Our analysis has potential applications in a wide number of fields. We give one example from industrial organisation. Consider two firms in a supply chain where the first firm produces an intermediate good and the second firm the final good. Imagine that the firms sign a contract where firm 1 supplies a quantity of intermediate goods to firm 2 on condition that firm 2 will then produce an amount of final goods, sell these, and split the revenues with firm 1. The firms are contractually obliged to produce a certain quantity of goods but the quality of production is not contractible (consider an industry such as food production where quality is hard to measure objectively). In this case, we can think of the partnership game explored above as representing a game where the firms choose levels of production quality after the contract to

[6] A notable exception is Frenkel *et al.* (2017) who present a plausible model of evolutionary dynamics that are not payoff-monotone due to sexual inheritance in a biological process, or due to combining traits from more than one mentor in a social learning process. They show that in such processes, stable population states do not correspond to Nash equilibria of the underlying material payoff game.

produce is signed. Plausibly, an increase in the quality of one firm's production will increase the marginal revenue gained by increasing the quality of the production of the other firm, but a firm's quality of production will come at an increasing marginal cost to that firm. When the managers of firms can communicate about their planned production quality, this may facilitate successful supply chains so long as the managers do not simply renege on any agreement. Our analysis suggests that competition in which management styles become more prevalent when they are relatively profitable will select firms run by managers with some tendency to fulfill non-contractual agreements, even when this does not maximise profits.

5.2 Mixed and asymmetric equilibria in the population game.

Our formal results above focused primarily on symmetric pure equilibria. In what follows we comment on the extension of our results to mixed and asymmetric equilibria.

Theorem 2 shows that $(\lambda_c^+, \lambda_c^+)$ is the unique symmetric and pure equilibrium of the population game. Numeric analysis suggests the following stronger result also holds. The population game does not admit any other Nash equilibrium (i.e., $(\lambda_c^+, \lambda_c^+)$ is uniquely stable when allowing also for mixed equilibria and asymmetric equilibria).[7] We leave the analytic verification of this conjecture (which, we believe, holds also for partial observability with a sufficiently high q) for future research.

It is relatively straightforward to extend Propositions 1 and 2 to mixed equilibria and to asymmetric equilibria. We refrain from doing so in order to simplify the notation of Section 4 (the formal definition of symmetric equilibria requires a somewhat more complicated notation). The arguments presented in the proofs of both propositions hold with minor changes also for mixed and asymmetric equilibria, and it can be shown that: (1) if $q = 0$, then all incumbents exert zero effort in any equilibrium of the population game, and (2) for any $q > 0$ in any equilibrium of the population game, a positive share of incumbent agents exert positive effort with positive probability (and, thus, also make positive promises, and are endowed with positive reneging aversion). Thus, the endowment of players with positive levels of reneging aversion in stable population states is a robust property that holds for any positive level of partial observability (at least with the simplifying assumption of perfect correlation between the observations of the two matched agents).

[7] The extension to asymmetric equilibria is especially interesting in setups in which the partnership game is played between agents from two different populations of complementary skills, and a stable state of the two populations corresponds to a possibly asymmetric Nash equilibrium of the two-population game (see the related setup studied in Ritzberger & Weibull, 1995).

5.3 Refinements of Continuous Stability

By using strict equilibrium and Nash equilibrium as our solution concepts describing stable population states, we implicitly assume that a stable population state has to be resistant only to perturbations in which a few agents change their reneging aversion. Eshel (1983) argues that in some setups one should also require stability against perturbation in which many (or all) agents slightly change their reneging aversion. Eshel presents the notion of *continuous stable strategy* to capture stability also against the latter kind of perturbations, and Oechssler & Riedel (2001) further refine it by presenting the notion of evolutionary robustness, which requires stability against all small perturbations consistent with the weak topology (see also the related notions of stability in Milchtaich, 2016). Population state λ^* is *evolutionarily robust* if an agent with cost λ^* outperforms other agents (on average) in any sufficiently close perturbed population state $\mu \in \Delta(\mathbb{R}^+)$, i.e.,

$$\sum_{\lambda \in \Delta(\mu)} \mu(\lambda) \cdot \pi(\lambda^*, \lambda) > \sum_{\lambda, \lambda' \in \Delta(\mu)} \mu(\lambda) \cdot \mu(\lambda') \cdot \pi(\lambda, \lambda') \tag{11}$$

One can show that the population state $(\lambda_c^+, \lambda_c^+)$ satisfies a slightly weaker version of the evolutionary robustness refinement of (11). Specifically, it satisfies the weak inequality counterpart of Eq. (11) for any sufficiently close $\mu \in \Delta(\mathbb{R}^+)$, and it satisfies the strict inequality whenever μ assigns positive mass to agents having a reneging aversion of at most λ_c^+. The intuition is that agents with a slightly higher reneging aversion (i.e., strictly above λ_c^+) play a no-effort equilibrium against all agents in the perturbed state μ. Thus, they are trivially weakly outperformed by a level of aversion λ_c^+, and strictly outperformed as long as μ includes some agents with a reneging aversion of at most λ_c^+ (against whom an agent with reneging aversion λ_c^+ achieves strictly positive payoffs). Finally, minor modifications to the arguments presented in the proof of Theorem 2 show that agents with a reneging aversion strictly below λ_c^+ are strictly outperformed by agents with a reneging aversion of λ_c^+ .

5.4 Conclusions and Directions for Future Research

We have demonstrated the evolutionary stability of an intermediate level of reneging aversion in a class of partnership games. This result is robust to players observing their partner's reneging aversion with only some sufficiently high probability less than 1. Notably, with any positive probability of correlated observation of levels of reneging aversion, positive effort is sustained in the partnership game in any stable state. These results demonstrate a strong tendency for evolution to select preferences for the partial keeping of promises. In stable populations, we see players making slightly "overoptimistic" promises and, while these are not fully realised, the outcome is welfare maximising among symmetric equilibria of the game. This outcome stands in

sharp contrast to the cheap talk prediction of no effort ever being exerted in these partnerships.

We have here developed the first evolutionary analysis of a direct concern for keeping one's word. In doing so, we give an evolutionary explanation of several key observations in the related empirical literature. In our model, a population of players with the stable level of reneging aversion will exert no effort if they are not allowed to communicate before choosing their actions, but the opportunity to send messages will lead to promises being made and higher levels of effort being exerted. This replicates the experimental finding of Charness & Dufwenberg (2006) that players are significantly more likely to make "cooperative" choices in a partnership setting when they have the ability to communicate before playing.[8] Secondly, in the presence of communication, the degree of cooperation in our model is both incomplete (some reneging always takes place) and sensitive to the returns from the partnership. Charness & Dufwenberg (2006) find that: (1) not all pairs make choices that achieve the cooperative outcome, (2) most players keep promises to play the cooperative action but some players break their promise, and (3) players are less likely to promise and achieve cooperation when the return from not cooperating is high.

Finally, we model a cost of promise-breaking *per se* rather than a cost of disappointing others' payoff expectations (so-called *guilt aversion*). While Charness & Dufwenberg (2006) explain their experimental findings with a model of guilt aversion, Vanberg (2008) demonstrates that both a direct cost of promise-breaking (i.e., reneging aversion) and guilt aversion can rationalise the findings of Charness & Dufwenberg (2006), but introduces variants of the partnership game experiments where only reneging costs are able to explain observed behaviour. In these variants, some players who have made promises are randomly re-matched with an alternative partner before choices are made in the subsequent subgame. Both the old and new partners of these players are unaware of the switch but the players find out the promise that had already been made to their new partner by a different player. Players' propensity to keep their promises is sensitive to whether they are re-matched in this way, suggesting that it is a concern with keeping a promise *they* have made rather than a concern with their partner's payoff expectations that motivates them to keep their word.

This research brings support to the focus of experimental and theoretical research on direct costs of lying or reneging on one's word in communication settings. Future research could explore the robustness of the stability of intermediate reneging aversion in alternative types of games and with more general information structures about preferences. Finally, following Alger & Weibull (2013), we conjecture that evolution under positive assortative matching could support the stability of non-cheap talk preferences even when types are unobserved.

[8]The appendix of Charness & Dufwenberg (2006) provides the text of the messages sent by players that demonstrates that they were indeed often used to make explicit promises about their own future action.

A Trembling-Hand Perfection

In this section we define the refinement of trembling-hand perfection in our setup in which each player has a continuous set of pure actions at each stage of the game. There are various ways in which one can apply the notion of trembling-hand perfection to a game with a continuous set of actions. In what follows, we choose one approach, mainly for its simplicity. All of our results hold for any plausible way in which one can apply trembling-hand perfection to this setup.

We begin by defining a perturbed partnership game in which each player may tremble with a small probability in the first stage and randomly choose a message from an arbitrary distribution with full support. Formally:

Definition 1. A perturbed partnership game is described by a tuple $\zeta = (\epsilon, \tilde{\sigma})$, where $\epsilon \in (0,1)$ is the probability that each agent sends in the first stage of the game a message sampled from the full-support distribution $\tilde{\sigma} \in \Delta(S)$.

Remark 5. All of our results remain the same if one defines a perturbed game to include also a tremble at the second stage. As this alternative approach makes the notations more complicated, without affecting any of the results, we choose to present the simpler Definition 1.

Definition 2. A behaviour strategy of player i is a pair $(\sigma_i, \chi_i(s_i, s_j))$, where $\sigma_i \in \Delta(S)$ is a distribution over the set of messages and $\chi_i(s_i, s_j)$ is the distribution of efforts exerted at the second stage as a function of the observed messages in the first stage (s_1, s_2). Given a behaviour strategy $(\sigma_i, \chi_i(s_i, s_j))$, let $(\sigma_i, \chi_i(s_i, s_j))_\zeta = ((1-\epsilon)\cdot\sigma_i + \epsilon\cdot\tilde{\sigma}_i, \chi_i(s_i, s_j))$ be the perturbed strategy according to which player i chooses a message according to σ_i with probability $(1-\epsilon)$, and "trembles" and chooses a message according to $\tilde{\sigma}$ otherwise, and in the second stage she chooses an effort according to $\chi_i(s_i, s_j)$, where (s_i, s_j) is the realised message profile in stage one.

A strategy profile is a trembling-hand perfect equilibrium if it is the limit of Nash equilibria of a converging sequence of perturbed games. Formally,

Definition 3. A strategy profile $((\sigma_i, \chi_i(s_i, s_j)), (\sigma_j, \chi_j(s_i, s_j)))$ is a Nash equilibrium of the perturbed partnership game $\zeta = (\epsilon, \tilde{\sigma},)$ if each strategy $(\sigma_i, x_i(s_i, s_j))$ is a best reply against the opponent's perturbed strategy $(\sigma_j, \chi_j(s_i, s_j))_\zeta$.

Definition 4. A strategy profile $((\sigma_i, \chi_i(s_i, s_j)), (\sigma_j, \chi_j(s_i, s_j)))$ is a trembling-hand perfect equilibrium if there exist distributions $\tilde{\sigma}, \tilde{\chi} \in \Delta(S)$, converging sequences of positive numbers $(\epsilon_1^n)_n, (\epsilon_2^n)_n \to 0$, and a converging sequence of strategy profiles $\left((\sigma_i^n, \chi_i^n(s_i, s_j)), \left(\sigma_j^n, \chi_j^n(s_i, s_j)\right)\right)_n \to ((\sigma_i, \chi_i(s_i, s_j)), (\sigma_j, \chi_j(s_i, s_j)))$, such that each strategy profile $\left((\sigma_i^n, \chi_i^n(s_i, s_j)), \left(\sigma_j^n, \chi_j^n(s_i, s_j)\right)\right)$ is a Nash equilibrium of the perturbed partnership game $(\epsilon_1^n, \epsilon_2^n, \tilde{\sigma}, \tilde{\chi})$.

Fact 2. *Similar to the standard definition of trembling-hand perfection in games with a finite set of actions (Selten, 1975), one can show that: (1) each partnership game admits a trembling-hand perfect equilibrium, and (2) each trembling-hand perfect equilibrium satisfies subgame perfection. The arguments are standard and are omitted for brevity. These observations imply that any unique subgame perfect equilibrium also satisfies trembling-hand perfection.*

B Characterization of the Perfect Equilibria when $\lambda_j = 0$

In what follows we formally characterise the perfect equilibria of the partnership game in the case where one player has zero reneging aversion

Proposition 3. *Assume that* $\lambda_i > \lambda_j = 0$.

1. *If* $\lambda_i > \frac{1}{c(2-c^2)} - c$ *(which implies that* $\Theta_i < 0$ *and* $R_i = \infty$*), then there is a unique continuum of subgame perfect equilibria in which* $s_i = 1$ *and* $s_j \in [0, 1]$.

2. *If* $\lambda_i < \frac{1}{c(2-c^2)} - c$ *(which implies that* $\Theta_i > 0$ *and* $R_i = 0$*), then there is a unique continuum of subgame perfect equilibria in which* $s_i = 0$ *and* $s_j \in [0, 1]$.

3. *If* $\lambda_i = \frac{1}{c(2-c^2)} - c$ *(which implies that* $\Theta_i = 0$*), then for any* $(s_i, s_j) \in S \times S$*, there exists a subgame perfect equilibrium in which the messages are* (s_i, s_j).

In all cases, all subgame perfect equilibria are trembling-hand perfect.

References

Abeler, Johannes, Raymond, Collin, Abeler, Johannes, & Nosenzo, Daniele. 2016. *Preferences for Truth-Telling*. Mimeo.

Admati, Anat R., & Perry, Motty. 1991. Joint projects without commitment. *The Review of Economic Studies*, **58**(2), 259–276.

Alger, Ingela, & Weibull, Jörgen W. 2010. Kinship, incentives, and evolution. *The American Economic Review*, **100**(4), 1725–1758.

Alger, Ingela, & Weibull, Jörgen W. 2012. A generalization of Hamilton's rule: Love others how much? *Journal of Theoretical Biology*, **299**, 42–54.

Alger, Ingela, & Weibull, Jörgen W. 2013. Homo Moralis – Preference evolution under incomplete information and assortative matching. *Econometrica*, **81**(6), 2269–2302.

Bicchieri, Cristina, & Lev-On, Azi. 2011. Studying the ethical implications of e-trust in the lab. *Ethics and Information Technology*, **13**(1), 5–15.

Cahuc, Pierre, & Kempf, Hubert. 1997. Alternative time patterns of decisions and dynamic strategic interactions. *The Economic Journal*, **107**(445), 1728–1741.

Caruana, Guillermo, & Einav, Liran. 2008. A theory of endogenous commitment. *The Review of Economic Studies*, **75**(1), 99–116.

Charness, Gary. 2000. Self-serving cheap talk: A test of Aumann's conjecture. *Games and Economic Behavior*, **33**(2), 177–194.

Charness, Gary, & Dufwenberg, Martin. 2006. Promises and partnership. *Econometrica*, **74**(6), 1579–1601.

Cooper, Russell, & John, Andrew. 1988. Coordinating coordination failures in Keynesian models. *The Quarterly Journal of Economics*, **103**(3), 441–463.

Crawford, Vincent. 1998. A survey of experiments on communication via cheap talk. *Journal of Economic Theory*, **78**(2), 286–298.

Dekel, Eddie, Ely, Jeffrey C., & Yilankaya, Okan. 2007. Evolution of preferences. *The Review of Economic Studies*, **74**(3), 685–704.

Demichelis, Stefano, & Weibull, Jörgen W. 2008. Language, meaning, and games: A model of communication, coordination, and evolution. *The American Economic Review*, **98**(4), 1292–1311.

Ellingsen, Tore, & Johannesson, Magnus. 2004. Promises, threats and fairness. *The Economic Journal*, **114**(495), 397–420.

Ellingsen, Tore, & Miettinen, Topi. 2008. Commitment and conflict in bilateral bargaining. *The American Economic Review*, **98**(4), 1629–1635.

Eshel, Ilan. 1983. Evolutionary and continuous stability. *Journal of Theoretical Biology*, **103**(1), 99–111.

Farrell, J, & Rabin, M. 1996. Cheap talk. *Journal of Economic Perspectives*, **10**(3), 103–118.

Farrell, Joseph. 1988. Communication, coordination and Nash equilibrium. *Economics Letters*, **27**(3), 209–214.

Fischbacher, Urs, & Föllmi-Heusi, Franziska. 2013. Lies in disguise: An experimental study on cheating. *Journal of the European Economic Association*, **11**(3), 525–547.

Frenkel, S., Heller, Y., & Teper, R. 2017. *The endowment effect as a blessing*. Mimeo.

Gneezy, Uri. 2005. Deception: The role of consequences. *The American Economic Review*, **95**(1), 384–394.

Güth, Werner, & Yaari, Menahem. 1992. Explaining reciprocal behavior in simple strategic games: An evolutionary approach. *In:* Witt, Ulrich (ed), *Explaining Process and Change: Approaches to Evolutionary Economics*. University of Michigan Press, Ann Arbor.

Guttman, Joel M. 2003. Repeated interaction and the evolution of preferences for reciprocity. *The Economic Journal*, **113**(489), 631–656.

Heifetz, Aviad, Shannon, Chris, & Spiegel, Yossi. 2007a. The Dynamic Evolution of Preferences. *Economic Theory*, **32**, 251–286. 10.1007/s00199-006-0121-7.

Heifetz, Aviad, Shannon, Chris, & Spiegel, Yossi. 2007b. What to Maximize If You Must. *Journal of Economic Theory*, **133**(1), 31–57.

Heller, Yuval. 2014. Language, meaning, and games: A model of communication, coordination, and evolution: Comment. *The American Economic Review*, **104**(6), 1857–1863.

Heller, Yuval, & Winter, Eyal. 2016. Rule rationality. *International Economic Review*, **57**(3), 997–1026.

Herold, Florian, & Kuzmics, Christoph. 2009. Evolutionary stability of discrimination under observability. *Games and Economic Behavior*, **67**(2), 542–551.

Holmstrom, Bengt. 1982. Moral hazard in teams. *The Bell Journal of Economics*, **11**(2), 74–91.

Hurkens, Sjaak, & Kartik, Navin. 2009. Would I lie to you? On social preferences and lying aversion. *Experimental Economics*, **12**(2), 180–192.

Kartik, Navin. 2009. Strategic communication with lying costs. *Review of Economic Studies*, **76**(4), 1359–1395.

Kartik, Navin, Ottaviani, Marco, & Squintani, Francesco. 2007. Credulity, lies, and costly talk. *Journal of Economic Theory*, **134**(1), 93–116.

Kartik, Navin, Tercieux, Olivier, & Holden, Richard. 2014. Simple mechanisms and preferences for honesty. *Games and Economic Behavior*, **83**, 284–290.

Kerr, Norbert L., & Kaufman-Gilliland, Cynthia M. 1994. Communication, commitment, and cooperation in social dilemmas. *Journal of Personality and Social Psychology*, **66**(3), 513.

Lundquist, Tobias, Ellingsen, Tore, Gribbe, Erik, & Johannesson, Magnus. 2009. The aversion to lying. *Journal of Economic Behavior and Organization*, **70**(1–2), 81–92.

Marx, Leslie M, & Matthews, Steven. 2000. Dynamic voluntary contribution to a public project. *Review of Economic Studies*, **67**(2), 327–358.

Matsushima, Hitoshi. 2008. Role of honesty in full implementation. *Journal of Economic Theory*, **139**(1), 353–359.

Maynard Smith, John, & Price, George R. 1973. The logic of animal conflict. *Nature*, **246**, 15–18.

Milchtaich, Igal. 2016. *Static stability in symmetric and population games*. Mimeo.

Nachbar, John H. 1990. Evolutionary selection dynamics in games: convergence and limit properties. *International Journal of Game Theory*, **19**(1), 59–89.

Oechssler, Jörg, & Riedel, Frank. 2001. Evolutionary dynamics on infinite strategy spaces. *Economic Theory*, **17**(1), 141–162.

Ok, Efe A, & Vega-Redondo, Fernando. 2001. On the evolution of individualistic preferences: An incomplete information scenario. *Journal of Economic Theory*, **97**(2), 231–254.

Radner, Roy, Myerson, Roger, & Maskin, Eric S. 1986. Example of a repeated partnership game with discounting and with uniformly inefficient equilibria. *Review of Economic Design*, **53**(1), 59–69.

Ritzberger, Klaus, & Weibull, Jörgen W. 1995. Evolutionary selection in normal-form games. *Econometrica: Journal of the Econometric Society*, 1371–1399.

Sally, D. 1995. Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, **7**, 58–92.

Sánchez-Pagés, Santiago, & Vorsatz, Marc. 2007. An experimental study of truth-telling in a sender-receiver game. *Games and Economic Behavior*, **61**(1), 86–112.

Sandholm, William H. 2010. *Population Games and Evolutionary Dynamics*. MIT press.

Schelling, T. C. 1980. *The Strategy of Conflict*. Harvard University Press.

Selten, Reinhard. 1975. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, **4**(1), 25–55.

Shalvi, Shaul, Handgraaf, Michel J. J., & De Dreu, Carsten K. W. 2011. Ethical manoeuvring: Why people avoid both major and minor lies. *British Journal of Management*, **22**, 16–27.

Taylor, P. D., & Jonker, L. B. 1978. Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences*, **40**(1–2), 145–156.

Vanberg, Christoph. 2008. Why do people keep their promises? An experimental test of two explanations. *Econometrica*, **76**(6), 1–4.

Weibull, Jörgen W. 1995. *Evolutionary Game Theory*. The MIT press.

C Various Lemmas (For Online Publication)

C.1 Conditions for the Existence of Each Best-Response Type

Lemma 1. $\Theta_i \leq 0$ *(which implies that player i sends the maximum message in all cases) if and only if*

$$\lambda_i \geq \frac{1}{(c+\lambda_j)(2-c(c+\lambda_j))} - c \quad \textit{and} \quad \lambda_j < \frac{2}{c} - c$$

Proof. The best response function set out in Eq. (10) implies that each player i always best responds with the maximum message, for all s_j, if and only if $\Theta_i \leq 0$. By the definition of Θ_i:

$$\begin{aligned}
\Theta_i \leq 0 &\iff c(c+\lambda_j) + \frac{1}{(c+\lambda_i)(c+\lambda_j)} - 2 \leq 0 \\
&\iff c(c+\lambda_j)(c+\lambda_i) + \frac{1}{(c+\lambda_j)} - 2(c+\lambda_i) \leq 0 \\
&\iff \lambda_i(c(c+\lambda_j)-2) \leq 2c - \frac{1}{c+\lambda_j} - c^2(c+\lambda_j) \\
&\iff \lambda_i(c(c+\lambda_j)-2) \leq -\frac{1}{c+\lambda_j} - c(c(c+\lambda_j)-2)
\end{aligned}$$

Where the second " $\iff$ " is obtained by multiplying by $(c+\lambda_i)$ and the third and fourth by gathering terms in λ_i and rearranging. To solve for λ_i we then divide by $(c(c+\lambda_j)-2)$. There are two solutions: one for when $(c(c+\lambda_j)-2)$ is positive, and one for when it is negative:

$$\lambda_i \leq \frac{-1}{(c+\lambda_j)[c(c+\lambda_j)-2]} - c < 0, \text{ and } c(c+\lambda_j) - 2 > 0 \tag{12}$$

$$\lambda_i \geq \frac{1}{(c+\lambda_j)[2-c(c+\lambda_j)]} - c > 0, \text{ and } c(c+\lambda_j) - 2 < 0 \tag{13}$$

We can see that the solution given by Eq. (12) implies that $\lambda_i < 0$, which is ruled out by assumption. Therefore, we have that $\Theta_i \leq 0 \iff$ Eq. (13) holds. Rearranging the second inequality in Eq. (13) to give a condition in terms of λ_j yields the lemma. □

Lemma 2. $\frac{\lambda_j}{\Theta_i} > 1$ *(which implies that player i sends a message that is some multiple (greater*

than 1) of player j's message) if and only if:

$$\left(\frac{1}{\lambda_j^2(1-c)+\lambda_j(2-2c^2+c)+c(2-c^2)}-c<\lambda_i<\frac{1}{(c+\lambda_j)(2-c(c+\lambda_j))}-c \quad \textit{and} \quad \lambda_j<\frac{2}{c}-c\right)$$

$$OR, \quad \left(\lambda_i>\frac{1}{\lambda_j^2(1-c)+\lambda_j(2-2c^2+c)+c(2-c^2)}-c \quad \textit{and} \quad \frac{2}{c}-c\leq\lambda_j<\frac{2-c^2}{c-1}\right)$$

Proof. Eq. (10) implies that a player best responds by playing a (greater than 1) multiple of her opponent's message, s_j, if and only if $\frac{\lambda_j}{\Theta_i}>1$. By the definition of Θ_i:

$$\frac{\lambda_j}{\Theta_i}>1 \iff \frac{\lambda_j}{c(c+\lambda_{-i})+\frac{1}{(c+\lambda_i)(c+\lambda_j)}-2}>1 \tag{14}$$

Since $\lambda_j\geq 0$, this holds if and only if

$$\lambda_j>c(c+\lambda_j)+\frac{1}{(c+\lambda_i)(c+\lambda_j)}-2>0 \tag{15}$$

The second of these inequalities is the requirement that $\Theta_i>0$, which is the converse of the condition derived for Lemma 1 and holds when

$$\lambda_i<\frac{1}{(c+\lambda_j)[2-c(c+\lambda_j)]}-c \quad or \quad \lambda_j>\frac{2}{c}-c \tag{16}$$

It is straightforward to see that if the second inequality in Eq. (16) holds, then $2-c(c+\lambda_j)<0$ and hence the first inequality implies $\lambda_i<0$. Therefore, imposing $\lambda_i\geq 0$, we have that $\Theta_i>0$ if and only if

$$\lambda_i<\frac{1}{(c+\lambda_j)[2-c(c+\lambda_j)]}-c \quad \textit{and} \quad \lambda_j<\frac{2}{c}-c$$
$$OR, \quad \lambda_j\geq\frac{2}{c}-c \tag{17}$$

The first inequality in Eq. (15) holds if and only if

$$\begin{aligned}\lambda_j >& c(c+\lambda_j) + \frac{1}{(c+\lambda_i)(c+\lambda_j)} - 2 \\ \iff& \lambda_i(\lambda_j + 2 - c(c+\lambda_j)) > -c(\lambda_j + 2 - c(c+\lambda_j)) + \frac{1}{c+\lambda_j} \end{aligned} \tag{18}$$

This $\iff$ is obtained by multiplying by $(c+\lambda_i)$ and rearranging. To solve for λ_i, we divide by $(\lambda_j + 2 - c(c+\lambda_j))$. There are two solutions: one for when $(\lambda_j + 2 - c(c+\lambda_j))$ is positive and one for when it is negative:

$$\lambda_i > \frac{1}{\lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)} - c \ > 0 \text{ and } \lambda_j + 2 - c(c+\lambda_j) > 0 \tag{19}$$

$$\lambda_i \leq \frac{1}{\lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)} - c \ < 0 \text{ and } \lambda_j + 2 - c(c+\lambda_j) < 0 \tag{20}$$

We can see that the solution given by Eq. (20) implies that $\lambda_i < 0$. This is ruled out by assumption, and so we have that the first inequality in Eq. (15) $\iff$ Eq. (19) holds. Rearranging the second inequality in Eq. (19) to give a condition in terms of λ_j and combining with Eq. (17) yields the lemma. □

Lemma 3. $0 < \frac{\lambda_j}{\Theta_i} < 1$ *(which implies that player i sends a message that is some fraction (less than 1) of player j's message) if and only if*

$$\left(\lambda_i < \frac{1}{\lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)} - c \quad \text{and} \quad \lambda_j < \frac{2-c^2}{c-1}\right)$$
$$OR, \quad \lambda_j \geq \frac{2-c^2}{c-1}$$

Proof. To obtain Lemma 3, we can see that Eq. (10) implies that a player will best respond by sending a message lower than her opponent if and only if $0 < \frac{\lambda_j}{\Theta_i} < 1$. This again implies that $\Theta_i > 0$ and so Eq. (17) must hold. We also must have that $\frac{\lambda_j}{\Theta_i} < 1$. In the proof of Lemma 2 it was demonstrated that $\frac{\lambda_j}{\Theta_i} > 1 \iff$ Eq. (19) holds. By taking the converse of Eq. (19) we have that $\frac{\lambda_j}{\Theta_i} < 1$ if and only if

$$\lambda_i < \frac{1}{\lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)} - c \text{ or } \lambda_j + 2 - c(c+\lambda_j) \leq 0 \tag{21}$$

As was also demonstrated in the proof of Lemma 2, if both conditions in Eq. (21) hold simultaneously, this implies that $\lambda_i < 0$. Therefore, imposing $\lambda_i \geq 0$ and rearranging the second inequality in Eq. (21) yields that $\frac{\lambda_j}{\Theta_i} < 1$ if and only if

$$\lambda_i < \frac{1}{\lambda_j^2(1-c)+\lambda_j(2-2c^2+c)+c(2-c^2)} - c \text{ and } \lambda_j < \frac{2-c^2}{c-1} \tag{22}$$

$$OR, \quad \lambda_j \geq \frac{2-c^2}{c-1} \tag{23}$$

From the proof of Lemma 2, we have that $\Theta_i > 0$ if and only if

$$\lambda_i < \frac{1}{(c+\lambda_j)[2-c(c+\lambda_j)]} - c \quad \text{and} \quad \lambda_j < \frac{2}{c} - c \tag{24}$$

$$OR, \quad \lambda_j \geq \frac{2}{c} - c \tag{25}$$

It is straightforward to see that Eq. (23) implies Eq. (25). We can also see that Eq. (22) implies Eq. (24) as

$$\begin{aligned}
&\frac{1}{\lambda_j^2(1-c)+\lambda_j(2-2c^2+c)+c(2-c^2)} - c < \frac{1}{(c+\lambda_j)[2-c(c+\lambda_j)]} - c \\
\iff &\frac{1}{\lambda_j^2(1-c)+\lambda_j(2-2c^2+c)+c(2-c^2)} < \frac{1}{(c+\lambda_j)[2-c(c+\lambda_j)]} \\
&\iff (c+\lambda_j)[2-c(c+\lambda_j)] < \lambda_j^2(1-c)+\lambda_j(2-2c^2+c)+c(2-c^2) \\
\iff &2c-c^2-\lambda_j c^2+2\lambda_j-\lambda_j c^2-\lambda_j^2 c < \lambda_j^2(1-c)+\lambda_j(2-2c^2+c)+c(2-c^2) \\
&\iff 0 < \lambda_j^2+\lambda_j^2 c
\end{aligned}$$

Therefore, $\Theta_i > 0$ is implied by $\frac{\lambda_j}{\Theta_i} < 1$, when $\lambda_i \geq 0$ is imposed, and so Eq. (22) and Eq. (23) can be combined to yield the lemma. □

C.2 Additional Lemmas

C.2.1 Lemma 4 (Used in Proof of Theorem 2)

Lemma 4. *Fix $c \in (1, 1.25)$. For all $\lambda_j \geq 0$ there exists a $\lambda_i \geq 0$ such that in the unique perfect equilibrium of the game $G(\lambda_i, \lambda_j)$, player i achieves a strictly-positive material payoff, i.e., $\pi(\lambda_i, \lambda_j) > 0$.*

Proof. Theorem 1 and Proposition 3 say that if $R_i = \infty$, or $R_j = \infty$, or $R_i \cdot R_j > 1$, then either $s_i = 1$ or $s_j = 1$ in the unique equilibrium of $G(\lambda_i, \lambda_j)$. We show that for all $\lambda_j \geq 0$ there exists a $\lambda_i \geq 0$ such that at least one of these conditions holds.

We first show that if $\lambda_j > \frac{81}{140}$ then setting $\lambda_i = 0$ yields $\Theta_j < 0$, which, by definition, implies $R_j = \infty$. To see this, first use the definition of Θ_j to write the condition $\Theta_j < 0$ when $\lambda_i = 0$,

and rearrange it to yield a lower bound on λ_j:

$$c^2 + \frac{1}{(c+\lambda_j)c} - 2 < 0 \iff \frac{1}{c+\lambda_j} < c\left(2-c^2\right) \iff \frac{1}{c\left(2-c^2\right)} - c < \lambda_j \tag{26}$$

The first derivative of this lower bound with respect to c is

$$\frac{3c^2-2}{(2c-c^3)^2} - 1 \tag{27}$$

Eq. (27) is positive for $c < 1.25$. The lower bound on λ_j given by Eq. (26) therefore attains its highest value when $c = 1.25$. This value is $\frac{81}{140} \approx 0.578$. We therefore have that for all $\lambda_j > \frac{81}{140}$, $\lambda_i = 0$ implies $\Theta_j < 0$ and hence $R_j = \infty$.

We next show that for $\lambda_j \leq \frac{81}{140}$, then for λ_i sufficiently large, either $\Theta_i \leq 0$ or $R_i \cdot R_j > 1$. We take the limit of Θ_i as $\lambda_i \to \infty$ and find the conditions under which this is negative:

$$\lim_{\lambda_i \to \infty} \Theta_i \leq 0 \Longleftrightarrow c(c+\lambda_j) - 2 \leq 0 \Longleftrightarrow \lambda_j \leq \frac{2}{c} - c \tag{28}$$

Next, we check the condition for satisfying $\frac{\lambda_i \cdot \lambda_j}{\Theta_j \cdot \Theta_i} > 1$, which implies that $R_i \cdot R_j > 1$. We take the limit of $\frac{\lambda_i \cdot \lambda_j}{\Theta_i \cdot \Theta_j}$ as $\lambda_i \to \infty$:

$$\begin{aligned}
\lim_{\lambda_i \to \infty} \frac{\lambda_i \cdot \lambda_j}{\Theta_i \cdot \Theta_j} &= \lim_{\lambda_i \to \infty} \left[\frac{\lambda_i}{c(c+\lambda_j) + \frac{1}{(c+\lambda_i)(c+\lambda_j)} - 2} \cdot \frac{\lambda_j}{c(c+\lambda_i) + \frac{1}{(c+\lambda_i)(c+\lambda_j)} - 2} \right] \\
&= \lim_{\lambda_i \to \infty} \left[\frac{\lambda_i}{c(c+\lambda_j) - 2} \cdot \frac{\lambda_j}{c(c+\lambda_i) - 2} \right] = \lim_{\lambda_i \to \infty} \left[\frac{\lambda_i}{c(c+\lambda_j) - 2} \cdot \frac{\lambda_j}{c \cdot \lambda_i} \right] \\
&= \lim_{\lambda_i \to \infty} \left[\frac{\lambda_i \cdot \lambda_j}{c \cdot \lambda_i \left(c(c+\lambda_j) - 2\right)} \right] = \lim_{\lambda_i \to \infty} \left[\frac{\lambda_j}{c\left[c(c+\lambda_j) - 2\right]} \right] = \frac{\lambda_j}{c\left[c(c+\lambda_j) - 2\right]},
\end{aligned}$$

where the second equality is derived from neglecting the term $\frac{1}{(c+\lambda_i)(c+\lambda_j)}$, which converges to zero as $\lambda_i \to \infty$, in each denominator, and the third equality is derived by neglecting the term $c^2 - 2$, which is negligible with respect to $c \cdot \lambda_i$ when taking the limit $\lambda_i \to \infty$, in the second denominator.

We then determine the conditions under which this limit is greater than 1:

$$\begin{aligned}
\frac{\lambda_j}{c\left[c(c+\lambda_j) - 2\right]} > 1 &\Longleftrightarrow c(c+\lambda_j) - 2 > 0 \quad \textit{and} \quad \lambda_j < c\left[c(c+\lambda_j) - 2\right] \\
&\Longleftrightarrow \frac{2}{c} - c < \lambda_j < \frac{c}{c^2-1} - c
\end{aligned} \tag{29}$$

Observe that the first inequality in Eq. (29) holds precisely when Eq. (28) does not hold. The first derivative of the right-hand side of the second inequality in Eq. (29) is $\frac{c^2-c^4-2}{(c^2-1)^2}$, which is clearly negative for all $c > 1$. When evaluated at $c = 1.25$, the right-hand side of the second inequality in Eq. (29) is $\frac{35}{36} > \frac{81}{140}$. Therefore, for all $c < 1.25$ and $\lambda_j \leq \frac{81}{140}$, this second inequality holds. We therefore have that for all $c < 1.25$ and $\lambda_j \leq \frac{81}{140}$, either $\Theta_i \leq 0$ or $\frac{\lambda_i \cdot \lambda_j}{\Theta_j \cdot \Theta_i} > 1$, when λ_i is sufficiently high.

Therefore, for all $c < 1.25$ and for all $\lambda_j \geq 0$, there exists a $\lambda_i \geq 0$ such that either $s_i = 1$ or $s_j = 1$ in the unique equilibrium of the game $G(\lambda_i, \lambda_j)$. To demonstrate that player i achieves positive payoff in equilibrium, we first note that in each of the above cases, there is at least one player who both sends a positive message and has positive reneging aversion, which by Eq. (6) implies that both players exert positive effort in equilibrium. Remark 2 implies that the payoff to a player exerting positive effort in equilibrium is strictly positive. □

C.2.2 Lemma 5 (Used in Proof of Theorem 2 and Theorem 3)

Lemma 5. *Fix $c \in (1, 1.24)$. Let $\mathcal{C}_c$ be the set of pairs (λ_i, λ_j) such that the game $G_c(\lambda_i, \lambda_j)$ admits the perfect equilibrium with $s_i = s_j = 1$ (i.e., maximum message equilibrium) and let $\underline{\lambda}_c \equiv min\{\lambda : (\lambda, \lambda) \in \mathcal{C}_c\}$ and let $\lambda_c^+ \equiv max\{\lambda : (\lambda, \lambda) \in \mathcal{C}_c\}$. Then (1) for all $\lambda \in [\underline{\lambda}_c, \lambda_c^+)$, there exists $\delta_\lambda > 0$ such that for all $\lambda' \in [\lambda, \lambda + \delta_\lambda)$, $G_c(\lambda', \lambda)$ admits a maximum message equilibrium. (2) For all $\lambda' \neq \lambda_c^+$, $(\lambda', \lambda_c^+) \notin \mathcal{C}_c$.*

Proof. The proof of Theorem 1 derives Eq. (31) and the corresponding condition for player j, which together define $\mathcal{C}_c$. By the strict convexity of the right-hand side of the first inequality in Eq. (31), we know that there can be at most two solutions to Eq. (33) and that for $\lambda_j > \lambda_c^+$, the right-hand side of the first inequality in Eq. (31) is increasing in λ_j. This means that for $\lambda_i = \lambda_j > \lambda_c^+$, Eq. (31) does not hold. Analogously, for $\lambda_i = \lambda_j < \underline{\lambda}_c$, Eq. (31) does not hold either. Therefore, if the second inequality in Eq. (31) holds when $\lambda_i = \lambda_j = \underline{\lambda}_c$ and when $\lambda_i = \lambda_j = \lambda_c^+$, these two points are the maximum and minimum symmetric points in $\mathcal{C}_c$. Given $\lambda_c^+ > \underline{\lambda}_c$, the second inequality in Eq. (31) holds in both of these cases if and only if

$$
\begin{aligned}
&\lambda_c^+ < \frac{2 - c^2}{c - 1} \\
\iff &\frac{1 + 2c - 2c^2}{2(c-1)} + \frac{\sqrt{5 - 4c}}{2(c-1)} = \frac{1 + 2c - c^2 + \sqrt{5 - 4c}}{2(c-1)} < \frac{2 - c^2}{c - 1} \\
\iff &1 + 2c - c^2 + \sqrt{5 - 4c} < 4 - 2c^2 \\
\iff &3 - 2c + \sqrt{5 - 4c} > 0 \\
\iff &c < 1.25
\end{aligned}
$$

By the convexity of $\mathcal{C}_c$, we therefore have that for all $\lambda \in [\underline{\lambda}_c, \lambda_c^+]$, $(\lambda, \lambda) \in \mathcal{C}_c$. The *strict* convexity of the first inequality of Eq. (31) defining the boundary of $\mathcal{C}_c$, implies that for all $\lambda \in (\underline{\lambda}_c, \lambda_c^+)$, (λ, λ) is not on the boundary of $\mathcal{C}_c$ and is therefore in the interior of $\mathcal{C}_c$. By the definition of an interior point of a convex set, for all $\lambda \in (\underline{\lambda}_c, \lambda_c^+)$, there exists a $\delta_\lambda > 0$ such that for all $\lambda' \in [\lambda, \lambda + \delta_\lambda)$, $(\lambda', \lambda) \in \mathcal{C}_c$.

To complete the proof of (1), we show that there exists $\delta_\lambda > 0$ such that for all $\lambda' \in [\underline{\lambda}_c, \underline{\lambda}_c + \delta_\lambda)$, $(\lambda', \underline{\lambda}_c) \in \mathcal{C}_c$. This will be the case if and only if there is $\delta_\lambda > 0$ such that Eq. (31) holds whenever $\lambda_i = \underline{\lambda}_c$ and $\lambda_j \in [\underline{\lambda}_c, \underline{\lambda}_c + \delta_\lambda)$. This will be case if and only if Eq. (31) does not become "tighter" as λ_j increases, i.e., if and only if the derivative of the right-hand side of the first inequality of Eq. (31) is less than or equal to zero when evaluated at $\lambda_j = \underline{\lambda}_c$. The derivative of the right-hand side of the first inequality of Eq. (31) with respect to λ_j is

$$\frac{c(2c + 2\lambda_j - 1) - 2(1 + \lambda_j)}{(\lambda_j + c)^2[\lambda_j(c-1) + c^2 - 2]^2} \tag{30}$$

When evaluated at $\lambda_j = \underline{\lambda}_c$, Eq. (30) is non-positive if[9] $c > \sqrt{5} - 1 \approx 1.24$. Therefore, we have that for all $\lambda \in [\underline{\lambda}_c, \lambda_c^+)$, there exists a $\delta_\lambda > 0$ such that for all $\lambda' \in [\lambda, \lambda + \delta_\lambda)$, $(\lambda', \lambda) \in \mathcal{C}_c$, and therefore $G_c(\lambda', \lambda)$ admits a maximum message equilibrium.

Point (2) is easily established by noting first that the right-hand side of the first inequality of Eq. (31) must be increasing in λ_j when evaluated at λ_c^+ (and at any $\lambda > \lambda_c^+$) as this is the second point at which this strictly convex function crosses the 45 degree line (the first being $\underline{\lambda}_c$). Therefore, given that λ_c^+ satisfies Eq. (33), an increase in λ_j with λ_i fixed at λ_c^+ means that Eq. (31) does not hold. By symmetry, an increase in λ_i with λ_j fixed at λ_c^+ means that the equivalent condition on λ_j is violated. Secondly, it is straightforward to see that given that λ_c^+ satisfies Eq. (33), when λ_j is fixed at λ_c^+, any $\lambda_i < \lambda_c^+$ must violate the first inequality in Eq. (31). Therefore, for any $\lambda' \neq \lambda_c^+$, $(\lambda', \lambda_c^+) \notin \mathcal{C}$, and therefore $G_c(\lambda', \lambda_c^+)$ does not admit a maximum message equilibrium. □

D Proofs of Main Results (For Online Publication)

D.1 Proof of Theorem 1

1. If $\min(R_i, R_j) > 1$, then, by the definition of R_i and R_j, either (a) $\Theta_i > 0$ and $\Theta_j > 0$ and $\frac{\lambda_j}{\Theta_i} > 1$ and $\frac{\lambda_i}{\Theta_j} > 1$, or (b) $\Theta_i > 0$ and $\frac{\lambda_j}{\Theta_i} > 1$ and $\Theta_j = 0$, or (c) $\Theta_i > 0$ and $\frac{\lambda_j}{\Theta_i} > 1$ and $\Theta_j < 0$, or (d) $\Theta_i = \Theta_j = 0$, or (e) $\Theta_i = 0$ and $\Theta_j < 0$, or (f) $\Theta_i < 0$ and

[9] This final result is obtained using Mathematica. The code is available in the supplementary appendix of this paper.

$\Theta_j < 0$. In case (a), by the unique best-response function derived in Eq. (10), equilibrium messages in this class of games satisfy $s_i^* = min\{\frac{\lambda_j}{\Theta_i} s_j, 1\}$ and $s_j^* = min\{\frac{\lambda_i}{\Theta_j} s_i, 1\}$. These equations are simultaneously satisfied if and only if $s_i^* = s_j^* = 0$ or $s_i^* = s_j^* = 1$. In case (b), Eq. (10) implies that equilibrium messages satisfy $s_i^* = min\{\frac{\lambda_i}{\Theta_j} s_i, 1\}$ and $s_j^* = 1$ if $\mu_{\sigma_j} > 0$ and $s_j^* \in \Delta(S)$ if $\mu_{\sigma_j} = 0$. These equations are simultaneously satisfied if and only if $s_i^* = s_j^* = 1$. In case (c), Eq. (10) implies that equilibrium messages satisfy $s_i^* = min\{\frac{\lambda_i}{\Theta_j} s_i, 1\}$ and $s_j^* = 1$. These equations are simultaneously satisfied if and only if $s_i^* = s_j^* = 1$. In case (d), Eq. (10) implies that equilibrium messages satisfy $s_i^* = 1$ if $\mu_{\sigma_j} > 0$ and $s_i^* \in [0, 1]$ if $\mu_{\sigma_j} = 0$ and $s_j^* = 1$ if $\mu_{\sigma_i} > 0$ and $s_j^* \in [0, 1]$ if $\mu_{\sigma_i} = 0$. These equations are simultaneously satisfied if and only if $s_i^* = s_j^* = 0$ or $s_i^* = s_j^* = 1$. In case (e), Eq. (10) implies that equilibrium messages satisfy $s_i^* = 1$ if $\mu_{\sigma_j} > 0$ and $s_j^* = 1$. These equations are simultaneously satisfied if and only if $s_i^* = s_j^* = 1$. In case (f), Eq. (10) implies that equilibrium messages satisfy $s_i^* = 1$ and $s_j^* = 1$, which implies that $s_i^* = s_j^* = 1$.

This implies that in all six cases (a, b, c, d, e, and f) the strategy profile $((1, x_1^e(s_1, s_2)), (1, x_2^e(s_1, s_2)))$ is a subgame perfect equilibrium. It is unique (and, thus, also satisfies trembling-hand perfection) in cases (b), (c), (e), and (f). In cases (a) and (d), the strategy profile $((0, x_1^e(s_1, s_2)), (0, x_2^e(s_1, s_2)))$ is the only additional subgame perfect equilibrium. In what follows we show that the equilibrium $((0, x_1^e(s_1, s_2)), (0, x_2^e(s_1, s_2)))$ fails to satisfy trembling-hand perfection in case (a). Let $\epsilon > 0$ be sufficiently small such that $\frac{\lambda_j}{\Theta_i} \cdot (1 - \epsilon) > 1$. Let $\tilde{\sigma}$ be an arbitrary distribution of messages with full support. Let $((\hat{s}_1, x_1^e(s_1, s_2)), (\hat{s}_2, x_2^e(s_1, s_2)))$ be a Nash equilibrium of the perturbed game $\zeta = (\epsilon, \tilde{\sigma})$ that satisfies $\hat{s}_1, \hat{s}_2 < 1$. This implies that each message $\hat{s}_i$ is a best reply against the perturbed strategy $(\hat{s}_j, x_2^e(s_1, s_2))_\zeta$, which is possible only if the following equation is satisfied for each player i:

$$\hat{s}_i = \frac{\lambda_j}{\Theta_i} \cdot \mu_{(\hat{s}_j)_\zeta} = \frac{\lambda_j}{\Theta_i} \cdot ((1 - \epsilon) \cdot \hat{s}_j + \epsilon \cdot \mu_{\tilde{\sigma}})$$

Observe that $\hat{s}_i = \frac{\lambda_j}{\Theta_i} \cdot ((1 - \epsilon) \cdot \hat{s}_j + \epsilon \cdot \mu_{\tilde{\sigma}}) > \hat{s}_j$. By the same argument the analogous equation in which i is replaced by j yields $\hat{s}_j > \hat{s}_i$, and we get a contradiction for $((\hat{s}_1, x_1^e(s_1, s_2)), (\hat{s}_2, x_2^e(s_1, s_2)))$ being a Nash equilibrium. This implies that $((0, x_1^e(s_1, s_2)), (0, x_2^e(s_1, s_2)))$ does not satisfy trembling-hand perfection, and that $((1, x_1^e(s_1, s_2)), (1, x_2^e(s_1, s_2)))$ is the unique trembling-hand perfect equilibrium also in case (a).

Next, we show the additional results on the set of pairs (λ_i, λ_j) such that $\min(R_i, R_j) > 1$. By the definition of R_i, we recall that $R_i \geq 1$ if and only if (1) $\Theta_i \leq 0$ or (2) $\Theta_i > 0$ and

$\frac{\lambda_j}{\Theta_i} \geq 1$. We can recall from Lemma 1 that $\Theta_i \leq 0$ if and only if

$$\lambda_i \geq \frac{1}{(c+\lambda_j)(2-c(c+\lambda_j))} - c \quad \text{and} \quad \lambda_j < \frac{2}{c} - c$$

We can recall from Lemma 2 that $\Theta_i > 0$ and $\frac{\lambda_j}{\Theta_i} \geq 1$ if and only if

$$\frac{1}{\lambda_j^2(1-c)+\lambda_j(2-2c^2+c)+c(2-c^2)} - c \leq \lambda_i < \frac{1}{(c+\lambda_j)(2-c(c+\lambda_j))} - c \text{ and}$$

$$\lambda_j < \frac{2}{c} - c$$

or

$$\lambda_i \geq \frac{1}{\lambda_j^2(1-c)+\lambda_j(2-2c^2+c)+c(2-c^2)} - c \text{ and}$$

$$\frac{2}{c} - c \leq \lambda_j < \frac{2-c^2}{c-1}$$

Combining these conditions yields $R_i \geq 1$ if and only if

$$\lambda_i \geq \frac{1}{\lambda_j^2(1-c)+\lambda_j(2-2c^2+c)+c(2-c^2)} - c \quad \text{and} \quad \lambda_j < \frac{2-c^2}{c-1} \tag{31}$$

We will first show that the set of points that satisfies Eq. (31) is convex. First, observe that the second derivative of the right-hand side of the first inequality of Eq. (31) (the lower bound on λ_i) with respect to λ_j is

$$\frac{2[3c^4+(6\lambda_j-3)c^3+(3\lambda_j^2-9\lambda_j-5)c^2+3\lambda_j^2+6\lambda_j+4]}{(\lambda_j+c)[2-c^2-\lambda_j(c-1)]} \tag{32}$$

The numerator of this expression is positive for all $\lambda_j > 0$ and[10] $c > 1$. This expression is therefore positive if and only if the denominator is positive, which clearly holds if and only if the expression in square brackets is positive:

$$2-c^2-\lambda_j(c-1) > 0 \iff \lambda_j < \frac{2-c^2}{c-1}$$

This is the second inequality of Eq. (31). Therefore, the set of points that satisfy Eq.

[10]Eq. (32) and the conditions for the positive numerator are derived using Mathematica. The code is available in the supplementary appendix of this paper.

(31) lies above a strictly convex function and is therefore a convex set. By the symmetry of the conditions for player j, we have that the set of points such that $R_j > 1$ is also convex. The intersection of two convex sets is a convex set. Therefore the set of points such that $\min(R_i, R_j) > 1$ (denoted by $\mathcal{C}_c$) is convex. By the symmetry of Eq. (31) and its equivalent for j (which together define the set $\mathcal{C}_c$), we have that $\mathcal{C}_c$ is symmetric (in the sense that $(\lambda_i, \lambda_j) \in \mathcal{C}_c \iff (\lambda_j, \lambda_i) \in \mathcal{C}_c$).

We now establish the interval of c in which $\mathcal{C}_c$ is non-empty. By the convexity and symmetry of $\mathcal{C}_c$, if this set is non-empty there must be a maximum and a minimum λ such that $(\lambda, \lambda) \in \mathcal{C}_c$. We now show that when $c < 1.25$ such maximum and minimum elements exist. Clearly, the maximum and minimum λ such that $(\lambda, \lambda) \in \mathcal{C}_c$ are the largest and smallest values of λ such that Eq. (31) holds when $\lambda_i = \lambda_j = \lambda$. Given that $\mathcal{C}_c$ is convex, these maximum and minimum values must obtain when at least one of the inequalities in Eq. (31) holds with equality. To find the maximum and minimum values of λ that satisfy the first inequality in Eq. (31), we solve the corresponding equality when $\lambda_i = \lambda_j = \lambda$. We then show that these are the largest and smallest values satisfying both inequalities simultaneously. Imposing $\lambda_i = \lambda_j = \lambda$ on the first inequality in Eq. (31), we obtain

$$\lambda = \frac{1}{\lambda^2(1-c) + \lambda(2-2c^2+c) + c(2-c^2)} - c \tag{33}$$

Multiplying by $\lambda^2(1-c) + \lambda(2-2c^2+c) + c(2-c^2)$ and rearranging yields

$$\lambda^3\Big[1-c\Big] + \lambda^2\Big[2+2c-3c^2\Big] + \lambda\Big[4c-3c^3+c^2\Big] - \Big[c^2-1\Big]^2 = 0 \tag{34}$$

Eq. (34) has two solutions when λ is positive:

$$\lambda = \frac{1+2c-2c^2}{2(c-1)} - \frac{\sqrt{5-4c}}{2(c-1)} \equiv \underline{\lambda}_c \tag{35}$$

$$\lambda = \frac{1+2c-2c^2}{2(c-1)} + \frac{\sqrt{5-4c}}{2(c-1)} \equiv \lambda_c^+ \tag{36}$$

Clearly these two solutions are defined if and only if $c < 1.25$. By inspection of Eq. (35) and Eq. (36), it is straightforward to see that for all $1 < c < 1.25$, $0 < \underline{\lambda}_c < \lambda_c^+ < \infty$. To see that $\min(R_i, R_j) > 1$ implies that $\underline{\lambda}_c \leq max(\lambda_i, \lambda_j) \leq \lambda_c^+$, note first that by the definition of $\underline{\lambda}_c$ and λ_c^+ as the minimum and maximum λ such that $(\lambda, \lambda) \in \mathcal{C}_c$ and by the convexity of $\mathcal{C}_c$, we have that $(\lambda, \lambda) \notin \mathcal{C}_c$ for each $\lambda \in [0, \underline{\lambda}_c) \cup [\lambda_c^+, \infty)$. Assume that there

exist $\lambda_i, \lambda_j < \underline{\lambda}_c$ such that $(\lambda_i, \lambda_j) \in \mathcal{C}_c$. By the symmetry of $\mathcal{C}_c$, we have $(\lambda_j, \lambda_i) \in \mathcal{C}_c$. Let $\lambda_k = \frac{\lambda_i + \lambda_j}{2} < \underline{\lambda}_c$. By the convexity of $\mathcal{C}_c$, we have $(\lambda_k, \lambda_k) \in \mathcal{C}_c$, which is a contradiction. This establishes that $(\lambda_i, \lambda_j) \in \mathcal{C}_c$ implies that $\underline{\lambda}_c \leq max\,(\lambda_i, \lambda_j)$. By assuming that there exist a $\lambda_i, \lambda_j > \lambda_c^+$ such that $(\lambda_i, \lambda_j) \in \mathcal{C}_c$, we can derive a contradiction in an analogous way.

Finally, we consider the case where $\underline{\lambda}_c < \lambda_i < \lambda_c^+ < \lambda_j$. We have established above that the right-hand side of the first inequality in Eq. (31) is strictly convex and crosses the 45 degree line for the second time at $\lambda_i = \lambda_j = \lambda^+$, which implies that for all $\lambda_j \geq \lambda_c^+$ this function is increasing in λ_j and yields a lower bound on λ_i greater than λ_c^+, which is a contradiction. We therefore have that $\lambda_c^+ < max\,(\lambda_i, \lambda_j)$ implies $(\lambda_i, \lambda_j) \notin \mathcal{C}_c$ and hence $(\lambda_i, \lambda_j) \in \mathcal{C}_c$ implies $max\,(\lambda_i, \lambda_j) \leq \lambda^+$. Combining the two bounds on $max\,(\lambda_i, \lambda_j)$ and recalling that by definition $(\lambda_i, \lambda_j) \in \mathcal{C}_c \iff \min(R_i, R_j) > 1$ completes the proof.

2. If $R_i \cdot R_j > 1 > R_j$ then, by the definition of R_i and R_j, either (a) $\Theta_i < 0$ and $\Theta_j > 0$ and $\frac{\lambda_i}{\Theta_j} < 1$, or (b) $\Theta_i = 0$ and $\Theta_j > 0$ and $\frac{\lambda_i}{\Theta_j} < 1$, or (c) $\Theta_i > 0$ and $\Theta_j > 0$ and $\frac{\lambda_j}{\Theta_i} \cdot \frac{\lambda_i}{\Theta_j} > 1$. In case (a) Eq. (10) implies that equilibrium messages satisfy $s_i^* = 1$ and $s_j^* = \frac{\lambda_i}{\Theta_j} s_i$. These equations are simultaneously satisfied if and only if $1 = s_i^* > s_j^* > 0$. In case (b), Eq. (10) implies that equilibrium messages satisfy $s_i^* = 1$ if $\mu_{\sigma_j} > 0$ and $s_i^* \in [0, 1]$ if $\mu_{\sigma_j} = 0$ and $s_j^* = \frac{\lambda_i}{\Theta_j} s_i$. These equations are simultaneously satisfied if and only if $1 = s_i^* > s_j^* > 0$ or $s_i^* = s_j^* = 0$. In case (c) Eq. (10) implies that equilibrium messages satisfy $s_i^* = min\{\frac{\lambda_j}{\Theta_i} s_j, 1\}$ and $s_j^* = \frac{\lambda_i}{\Theta_j} s_i$. Given that $\frac{\lambda_j}{\Theta_i} \cdot \frac{\lambda_i}{\Theta_j} > 1$, these equations are simultaneously satisfied if and only if $1 = s_i^* > s_j^* > 0$ or $s_i^* = s_j^* = 0$. In all three cases (a, b and c), there exists a subgame perfect equilibrium in which $1 = s_i^* > s_j^* > 0$. This is the unique subgame perfect equilibrium in case (a) and therefore it must satisfy trembling-hand perfection. In cases (b) and (c) there exists also a subgame perfect equilibrium in which $s_i^* = s_j^* = 0$.
Next we show that this latter subgame perfect equilibrium
$((0, x_1^e(s_1, s_2)), (0, x_2^e(s_1, s_2)))$ fails to satisfy trembling-hand perfection in cases (b) and (c). Assume to the contrary that $((0, x_1^e(s_1, s_2)), (0, x_2^e(s_1, s_2)))$ satisfies trembling-hand perfection. This implies that there exists a distribution of messages with full support $\tilde{\sigma}$ and $\bar{\epsilon} > 0$, such that for each $0 < \epsilon < \bar{\epsilon}$, $((\hat{s}_1, x_1^e(s_1, s_2)), (\hat{s}_2, x_2^e(s_1, s_2)))$ is a Nash equilibrium of the perturbed game $\zeta = (\epsilon, \tilde{\sigma})$ that satisfies $\hat{s}_1, \hat{s}_2 < 1$. This implies that each message $\hat{s}_i$ is a best reply against the perturbed strategy $(\hat{s}_j, x_2^e(s_1, s_2))_\zeta$. We begin with case (b). Observe that the expected signal of player j is positive, which implies, due to the second condition in Eq. (10), that the unique best-reply of player i is the maximal message $\hat{s}_i = 1$, and we get a contradiction. Turning to case (c), let $0 < \epsilon < \bar{\epsilon}$ be sufficiently small such that $\frac{\lambda_j}{\Theta_i} \cdot \frac{\lambda_i}{\Theta_j} \cdot (1 - \epsilon)^2 > 1$. This implies that each message $\hat{s}_i$ is a best reply against

the perturbed strategy $\left(\hat{s}_j, x_2^e\left(s_1, s_2\right)\right)_\zeta$, which is possible only if the following equation is satisfied for each player i:

$$\hat{s}_i = \frac{\lambda_j}{\Theta_i} \cdot \mu_{(\hat{s}_j)_\zeta} = \frac{\lambda_j}{\Theta_i} \cdot \left((1-\epsilon) \cdot \hat{s}_j + \epsilon \cdot \mu_{\tilde{\sigma}}\right)$$

Observe that the right-hand side is strictly positive for any value of $\hat{s}_j \in [0,1]$, which implies that $\hat{s}_i > 0$ for each player i. Substituting the value of $\hat{s}_j$ from the analogous equation $\hat{s}_j = \frac{\lambda_j}{\Theta_i} \cdot \mu_{(\hat{s}_i)_\zeta}$ yields

$$\hat{s}_i = \frac{\lambda_j}{\Theta_i} \cdot \left((1-\epsilon) \cdot \frac{\lambda_i}{\Theta_j} \cdot \left((1-\epsilon) \cdot \hat{s}_i + \epsilon \cdot \mu_{\tilde{\sigma}}\right) + \epsilon \cdot \mu_{\tilde{\sigma}}\right)$$

Simplifying the equation yields

$$\hat{s}_i = \frac{(1-\epsilon) \cdot \frac{\lambda_i}{\Theta_j} \cdot \epsilon \cdot \mu_{\tilde{\sigma}} + \epsilon \cdot \mu_{\tilde{\sigma}}}{1 - (1-\epsilon)^2 \cdot \frac{\lambda_j}{\Theta_i} \cdot \frac{\lambda_i}{\Theta_j}},$$

which implies that $\hat{s}_i$ is negative (because the numerator is positive while the denominator is negative), and we get a contradiction.

Finally, we show that there exist $\lambda_i, \lambda_j > 0$ such that $R_i \cdot R_j > 1 > R_j$ if and only if $c < \sqrt{2}$. We first show that $R_i \cdot R_j > 1 > R_j \implies c < \sqrt{2}$. By the definition of R_i, we have that $R_i \cdot R_j > 1 > R_j$ implies either $\Theta_i \leq 0$ or $\frac{\lambda_j}{\Theta_i} > 1$. From Lemma 1 and Lemma 2, we see that this implies either $\lambda_j < \frac{2}{c} - c$ or $\lambda_j < \frac{2-c^2}{c-1}$, both of which imply $c < \sqrt{2}$, given that $\lambda_j > 0$. Next we show that $c < \sqrt{2}$ implies that there exist $\lambda_i, \lambda_j > 0$ such that $R_i \cdot R_j > 1 > R_j$. Again, $c < \sqrt{2} \implies \frac{2}{c} - c > 0$, and so Lemma 1 tells us that for any $\lambda_j < \frac{2}{c} - c$ and λ_i sufficiently large, we have $\Theta_i \leq 0$. The equivalent of Lemma 3 for player j tells us that if $\lambda_i \geq \frac{2-c^2}{c-1}$ then for any λ_j we have $\frac{\lambda_i}{\Theta_j} < 1$ (and hence $\Theta_j > 0$). Therefore for any $\lambda_j < \frac{2}{c} - c$ and λ_i sufficiently large we have $\Theta_i \leq 0$ and $\Theta_j > 0$ and $\frac{\lambda_i}{\Theta_j} < 1$, which, by the definition of R_i, together imply $R_i \cdot R_j > 1 > R_j$.

3. If $R_i \cdot R_j < 1$, then, by the definition of R_i and R_j, $\Theta_i > 0$ and $\Theta_j > 0$ and $\frac{\lambda_j}{\Theta_i} \cdot \frac{\lambda_i}{\Theta_j} < 1$. By the unique best-response function derived in Eq. (10), equilibrium messages in this class of games satisfy $s_i^* = \frac{\lambda_j}{\Theta_i} s_j$ and $s_j^* = \frac{\lambda_i}{\Theta_j} s_i$. Given that $\frac{\lambda_j}{\Theta_i} \cdot \frac{\lambda_i}{\Theta_j} < 1$, these equations are jointly satisfied if and only if $s_i^* = s_j^* = 0$, which is therefore the unique subgame perfect equilibrium pair of messages. This implies that the unique subgame perfect equilibrium is $\left(\left(0, x_1^e\left(s_1, s_2\right)\right), \left(0, x_2^e\left(s_1, s_2\right)\right)\right)$ (where $\left(x_1^e\left(s_1, s_2\right), x_2^e\left(s_1, s_2\right)\right)$ is the unique equilibrium in the second round following a message profile of (s_1, s_2) as defined in Section 2.2). As

observed above (Fact 2), a unique subgame perfect equilibrium must also satisfy trembling-hand perfection.

Finally, we prove that there always exists a pair (λ_i, λ_j) for which $R_i \cdot R_j < 1$. By the definition of R_i, any pair of parameters such that $\lambda_i = \lambda_j > 0$ implies $R_i = R_j$. If $R_i = R_j$ and it is not the case that $min(R_i, R_j) > 1$, we must have $R_i \cdot R_j < 1$. The results from point 1 of the theorem therefore imply that for all $c > 1$ there exists a $\lambda \notin [\underline{\lambda}_c, \lambda_c^+]$ such that $\lambda_i = \lambda_j = \lambda \iff R_i \cdot R_j < 1$.

D.2 Proof of Proposition 3

Proof. When $\lambda_j = 0$, the utility of player j is independent of s_j. To see this, we impose the condition $\lambda_j = 0$, on the expression for utility, taking subgame play as given (the analogue of Eq. (7) for player j). This yields

$$U_j(s_j, s_i, c) = \frac{c(\lambda_i s_i)^2}{2[c(c+\lambda_i) - 1]^2} \tag{37}$$

This expression is clearly independent of s_j and therefore any message sent by player j is a best response to any s_i. The utility of player i as a function of s_i, taking subgame play as given, is

$$U_i(s_i, s_j, c) = \frac{c(1 - \frac{c^2}{2})(\lambda_i s_i)^2}{[c(c+\lambda_i) - 1]^2} - \frac{\lambda_i}{2}\left[s_i - \frac{c\lambda_i s_i}{c(c+\lambda_i) - 1}\right]^2 \tag{38}$$

The first derivative of this function with respect to s_i is given by:

$$\left[2 - c^2 - \frac{1}{c(c+\lambda_i)}\right] s_i \tag{39}$$

Observe that Eq. (39) is negative iff $\lambda_i < \frac{1}{c(2-c^2)} - c$, and in this case: (I) $\Theta_i = -\left[2 - c^2 - \frac{1}{c(c+\lambda_i)}\right] > 0$, (II) the utility function is everywhere decreasing in s_i, and (III) the optimal choice of s_i, given any s_j, is 0. This implies that in this case there is a unique continuum of subgame perfect equilibria in which $s_i = 0$ and s_j can take any value in $[0, 1]$.

When Eq. (39) is positive, then: (I) $\Theta_i = -\left[2 - c^2 - \frac{1}{c(c+\lambda_i)}\right] < 0$ and, by definition, $R_i = \infty$, (II) the utility function is everywhere increasing in s_i, and (III) the optimal choice of s_i, given any s_j, is 1. This implies that in this case there is a unique continuum of subgame perfect equilibria in which $s_i = 1$ and s_j can take any value in $[0, 1]$.

Finally, when $\lambda_i = \frac{1}{c(2-c^2)} - c$, Eq. (39) is zero and any message $s_i \in [0,1]$ is optimal. This implies that for any $(s_i, s_j) \in S \times S$, there exists a subgame perfect equilibrium in which the messages are (s_i, s_j).

The argument that these equilibria satisfy trembling-hand perfection is analogous to the arguments presented in the proof of Theorem 1, and is omitted for brevity. □

D.3 Proof of Theorem 2

Proof. Corollary 1 shows that if players have identical positive reneging costs in the partnership game, then they play identical messages in its unique perfect equilibrium, and therefore the game admits either a no-effort or a maximum message equilibrium. If $\lambda_i = \lambda_j = 0$, Eq. (6) implies that $x_i = x_j = 0$ and the game admits only a no-effort equilibrium. Therefore, a symmetric pure-strategy Nash equilibrium of the population game corresponds to either a no-effort or a maximum message equilibrium of the partnership game. We consider these two sets of symmetric strategy profiles in turn.

For any λ such that the unique equilibrium in the corresponding partnership game $G(\lambda, \lambda)$ is a no-effort equilibrium, we have $\pi(\lambda, \lambda) = 0$. Lemma 4 shows that for all $\lambda \geq 0$, there exists a $\lambda' \geq 0$ such that $\pi\left(\lambda', \lambda\right) > 0$. Therefore, for all λ such that $\pi(\lambda, \lambda) = 0$, (λ, λ) cannot be a Nash equilibrium of the population game.

For any λ such that the unique equilibrium in the corresponding partnership game, $G(\lambda, \lambda)$, is a maximum message equilibrium, we say that such an equilibrium is either *interior* to the set of maximum message equilibria or *exterior* to that set. An equilibrium is *interior* if there exists a $\bar{\delta}$ such that for all $0 < \delta < \bar{\delta}$, the unique equilibrium of $G(\lambda + \delta, \lambda)$ is a maximum message equilibrium, and it is *exterior* otherwise. For all λ such that the unique equilibrium of $G(\lambda, \lambda)$ is a maximum message equilibrium, the equilibrium payoff to both players is obtained by substituting $s_i = s_j = 1$ in Eq. (7):

$$\pi_i(\lambda_i, \lambda_j) = \frac{[(c+\lambda_j)\lambda_i + \lambda_j][(c+\lambda_i)\lambda_j + \lambda_i]}{[(c+\lambda_i)(c+\lambda_j) - 1]^2} - \frac{c[(c+\lambda_j)\lambda_i + \lambda_j]^2}{2[(c+\lambda_i)(c+\lambda_j) - 1]^2} \tag{40}$$

The first derivative of this function with respect to λ_i is

$$\frac{(c-1)(1+c+\lambda_j)[\lambda_i c^3 + 2c^2\lambda_i\lambda_j + c\lambda_i(\lambda_j^2 - \lambda_j - 2) - \lambda_j(1+\lambda_i(2+\lambda_j))]}{[c^2 - 1 + \lambda_i\lambda_j + c(\lambda_i + \lambda_j)]^3} \tag{41}$$

Imposing the condition $\lambda_i = \lambda_j = \lambda$, we can simplify this expression to[11]

$$\frac{(c-1)[c(c+\lambda-1)-1-\lambda]\lambda}{[c+1+\lambda][c-1+\lambda]^3} \tag{42}$$

This expression is strictly positive if and only if

$$c(c+\lambda-1)-1-\lambda>0 \iff \lambda < \frac{1+c-c^2}{c-1} \tag{43}$$

Recall from Theorem 1 that a maximum message equilibrium exists only if $min(R_i, R_j) > 1$ and that this requires that either $\Theta_i \leq 0$ or $\frac{\lambda_j}{\Theta_i} > 1$ (and that the analogous conditions hold for j) and hence the conditions in either Lemma 1 or Lemma 2 must hold. Lemma 1 and Lemma 2 each imply that

$$\lambda_j < \frac{2-c^2}{c-1} \tag{44}$$

Therefore when $\lambda_i = \lambda_j = \lambda$, we have

$$\lambda < \frac{2-c^2}{c-1} < \frac{1+c-c^2}{c-1} \tag{45}$$

Where the second inequality clearly follows when $c > 1$. We can see that this yields us the second inequality in Eq. (43) and hence Eq. (42) is always positive in a maximum message equilibrium.

Therefore, for any strategy profile of the population game (λ, λ) such that $G(\lambda, \lambda)$ admits an interior maximum message equilibrium, there exists some $\lambda' > \lambda$ such that $\pi\left(\lambda', \lambda\right) > \pi(\lambda, \lambda)$ and hence no such strategy profile is a Nash equilibrium of the population game.

We have shown that the only potential symmetric pure Nash equilibria of the population game are those corresponding to symmetric exterior maximum message equilibria of the partnership game. Lemma 5 shows that there is a unique game $G(\lambda_c^+, \lambda_c^+)$ that admits such an equilibrium when $c \in (1, 1.24)$. We now show that $(\lambda_c^+, \lambda_c^+)$ is a Nash equilibrium of the population game.

We first show that any unilateral deviation from the candidate equilibrium to a lower reneging aversion yields a strictly lower payoff, i.e., $\pi\left(\lambda', \lambda_c^+\right) < \pi(\lambda_c^+, \lambda_c^+)$ for $\lambda' \in [0, \lambda_c^+)$. Point (2) of Lemma 5 says that for all $\lambda' \in [0, \lambda_c^+)$, the game $G\left(\lambda', \lambda_c^+\right)$ does not admit a maximum message equilibrium. Therefore for all such deviations, the unique equilibrium of the corresponding partnership game $G\left(\lambda', \lambda_c^+\right)$ is either a no-effort or a two-message equilibrium. In the former case, the effort levels of both players are zero and so we have $\pi\left(\lambda_c^+, \lambda_c^+\right) > \pi(\lambda', \lambda_c^+) = 0$. In the case of a two-message equilibrium, the payoff to the deviating player is obtained by substituting

[11]The derivative given by Eq. (41) and its simplification when $\lambda_i = \lambda_j$ is obtained using Mathematica. Code available in the supplementary appendix accompanying this paper.

the expression for equilibrium effort (Eq. 6) into the expression for material payoff (Eq. 1) and imposing the conditions $s_i = \frac{\lambda_j}{\Theta_i}$ and $s_j = 1$ and $\lambda_j = \lambda_c^+$ (player i is therefore the player making the deviation):

$$\pi_i(\lambda_i, \lambda_c^+) = \frac{[(c+\lambda_c^+)\lambda_i\frac{\lambda_j}{\Theta_i} + \lambda_c^+][(c+\lambda_i)\lambda_c^+ + \lambda_i\frac{\lambda_j}{\Theta_i}]}{[(c+\lambda_i)(c+\lambda_c^+) - 1]^2} - \frac{c[(c+\lambda_c^+)\lambda_i\frac{\lambda_j}{\Theta_i} + \lambda_c^+]^2}{2[(c+\lambda_i)(c+\lambda_c^+) - 1]^2} \tag{46}$$

The derivative of this expression with respect to λ_i is[12]

$$\frac{[\lambda_c^+]^2(c(\lambda_c^+ + c) - 1)^2}{[1 + (c+\lambda_c^+)(c+\lambda_i)(c(\lambda_c^+ + c) - 2)]^3} \tag{47}$$

Clearly the numerator of Eq. (47) is always positive. A *sufficient* condition for the denominator, and hence for the whole expression, to be strictly positive is that

$$c(\lambda_c^+ + c) - 2 > 0 \iff \lambda_c^+ > \frac{2}{c} - c \tag{48}$$

This always holds as

$$\begin{aligned}
&\lambda_c^+ = \frac{1 + 2c - 2c^2}{2(c-1)} + \frac{\sqrt{5-4c}}{2(c-1)} > \frac{2}{c} - c \\
\iff &1 + 2c - 2c^2 + \sqrt{5-4c} > \frac{4(c-1)}{c} - 2c(c-1) \\
\iff &-3 + \sqrt{5-4c} + \frac{4}{c} > 0 \\
\iff &c < 1.25
\end{aligned}$$

Where the final $\iff$ follows from the fact that $\sqrt{5-4c}$ is positive and defined if and only if $c < 1.25$, and $\frac{4}{c} - 3$ is positive for all $c < 1.33$.

To complete the proof, we show that a unilateral deviation from the candidate equilibrium to a higher reneging aversion yields a strictly lower payoff, i.e., $\pi\left(\lambda', \lambda_c^+\right) < \pi(\lambda_c^+, \lambda_c^+)$ for $\lambda' > \lambda_c^+$. By the definition of an exterior equilibrium, the unique equilibrium of all the corresponding partnership games $G\left(\lambda', \lambda_c^+\right)$ is either a no-effort or a two-message equilibrium. In the former case, the effort levels of both players are zero and so we have $\pi\left(\lambda_c^+, \lambda_c^+\right) > \pi(\lambda', \lambda_c^+) = 0$. In the case of a two-message equilibrium, the payoff to the deviating player is obtained by substituting the expression for equilibrium effort (Eq. 6) into the expression for material payoff (Eq. 1) and imposing the conditions $s_i = 1$ and $s_j = \frac{\lambda_i}{\Theta_j}$ and $\lambda_j = \lambda_c^+$ (player i is therefore the player making

[12]This derivative was calculated using Mathematica. Code is available in the supplementary appendix of this paper.

the deviation):

$$\pi_i(\lambda_i, \lambda_c^+) = \frac{[(c+\lambda_c^+)\lambda_i + \lambda_c^+ \frac{\lambda_i}{\Theta_j}][(c+\lambda_i)\lambda_c^+ \frac{\lambda_i}{\Theta_j} + \lambda_i]}{[(c+\lambda_i)(c+\lambda_c^+) - 1]^2} - \frac{c[(c+\lambda_c^+)\lambda_i + \lambda_c^+ \frac{\lambda_i}{\Theta_j}]^2}{2[(c+\lambda_i)(c+\lambda_c^+) - 1]^2} \tag{49}$$

In the supplementary appendix of this paper, we present the explicit formula for derivative of Eq. (49) with respect to λ_i and the Mathematica code proving that this derivative is strictly negative for all $\lambda_i > \lambda_c^+$. Hence for any $\lambda_i > \lambda_c^+$ such that the unique equilibrium of $G(\lambda_i, \lambda_c^+)$ is a two-message equilibrium we have $\pi(\lambda_i, \lambda_c^+) < \pi(\lambda_c^+, \lambda_c^+)$. Therefore, the population game admits a unique pure strategy Nash equilibrium, $(\lambda_c^+, \lambda_c^+)$.

We saw that any possible deviation from this equilibrium yields the deviating player a strictly lower payoff and hence this equilibrium is strict (point 1). By definition, λ_c^+ is the highest reneging cost, given c, such that the unique equilibrium of $G(\lambda_c^+, \lambda_c^+)$ is a maximum message equilibrium and therefore $s_i = s_j = 1$ in the equilibrium of this game (point 2).

To see point 3 (i.e., $\pi(\lambda_c^+, \lambda_c^+) > \pi\left(\lambda', \lambda'\right)$ for any $\lambda' \neq \lambda_c^+$), we recall again by Corollary 1 that the unique equilibrium of any $G(\lambda, \lambda)$, is symmetric and therefore either a no-effort equilibrium or a maximum message equilibrium. For any λ such that the unique equilibrium of $G(\lambda, \lambda)$ is a no-effort equilibrium, $s_i = s_j = x_i = x_j = 0$ and so $\pi(\lambda, \lambda) = 0$. To find the material payoff in a maximum message equilibrium, we recall Eq. (1) for the material payoff, and impose $x_i = x_j = x$, which yields

$$\pi(\lambda, \lambda) = x^2 - \frac{cx^2}{2} \tag{50}$$

Which is clearly positive and increasing in x for all $c < 2$. The reneging cost that maximises the material payoff in a symmetric game is therefore that which maximises equilibrium effort. Equilibrium effort in a maximum message equilibrium is obtained by imposing $s_i = s_j = 1$ and $\lambda_i = \lambda_j = \lambda$ on the equation for equilibrium effort (Eq. 6):

$$\frac{(c+\lambda)\lambda + \lambda}{(c+\lambda)(c+\lambda) - 1} = \frac{\lambda}{c+\lambda-1} \tag{51}$$

The derivative of this expression with respect to λ is

$$\frac{(c+\lambda-1) - \lambda}{[c+\lambda-1]^2} = \frac{(c-1)}{[c+\lambda-1]^2} \tag{52}$$

Which is clearly positive for all $c > 1$. Therefore, the reneging cost that maximises effort, and therefore the material payoff, in a symmetric game is the highest λ such that the unique equilibrium of $G(\lambda, \lambda)$ is a maximum message equilibrium. By definition, this is λ_c^+.

To see point 4 (i.e., λ_c^+ and $\pi\left(\lambda_c^+, \lambda_c^+\right)$ are decreasing in c), we recall that:

$$\lambda_c^+ = \frac{1 + 2c - 2c^2}{2(c-1)} + \frac{\sqrt{5-4c}}{2(c-1)} = \frac{1 + 2c(1-c) + \sqrt{5-4c}}{2(c-1)} \tag{53}$$

We can see that for $c \in (1, 1.25)$, the numerator of Eq. (53) is decreasing in c. The denominator of Eq. (53) is increasing in c and hence λ_c^+ is decreasing in c for all $c \in (1, 1.22)$. Given that we showed that for λ such that $G(\lambda, \lambda)$, $\pi(\lambda, \lambda)$ is increasing in λ and that λ_c^+ is decreasing in c, we have that $\pi(\lambda_c^+, \lambda_c^+)$ is decreasing in c.

To see point 5 (i.e., $\lim_{c\to 1} \lambda_c^+ = \infty$, and $\lim_{c\to 1} \pi\left(\lambda_c^+, \lambda_c^+\right) = \frac{1}{2}$), we note that as $c \to 1$, the numerator of Eq. (53) is increasing and the denominator of Eq. (53) converges to zero. Hence $\lim_{c\to 1} \lambda_c^+ = \infty$. To find the limit of the players' material payoff in the game $G(\lambda_c^+, \lambda_c^+)$ as $c \to 1$, we substitute the expression for effort in a maximum message equilibrium (Eq. 51) into that for material payoff in a symmetric equilibrium (Eq. (50)) when $\lambda = \lambda_c^+$:

$$\pi(\lambda_c^+, \lambda_c^+) = [\frac{\lambda_c^+}{c + \lambda_c^+ - 1}]^2 [1 - \frac{c}{2}] \tag{54}$$

As $c \to 1$, $\lambda_c^+ \to \infty$ and therefore the limit of Eq. (54) is given by

$$\lim_{c\to 1} \pi(\lambda_c^+, \lambda_c^+) = \lim_{c\to 1} [\frac{\lambda_c^+}{c + \lambda_c^+ - 1}]^2 [1 - \frac{c}{2}] = (1 - \frac{1}{2}) = \frac{1}{2} \tag{55}$$

□

D.4 Proof of Theorem 3

Proof. We have to prove that for each $c \in (1, 1.22)$, there exists $\bar{q} < 1$ such that $(\lambda_c^+, \lambda_c^+)$ is a strict Nash equilibrium of the population game with observability q for each $q \in [\bar{q}, 1)$, i.e., that $\pi_q\left(\lambda', \lambda_c^+ | \lambda_c^+\right) < \pi\left(\lambda_c^+, \lambda_c^+\right)$ for all $\lambda' \neq \lambda_c^+$. We first note that by Lemma 5, for any $\lambda' \neq \lambda_c^+$, the partnership game played after reneging costs are observed, $G(\lambda', \lambda_c^+)$, does not admit a maximum message equilibrium and so, by Theorem 1, must admit either a no-effort ($s_i = s_j = 0$) or a two-message ($s_i \neq s_j$) equilibrium. In the no-effort case, the material payoff to both players is zero and hence any mutant λ' such that $G(\lambda', \lambda_c^+)$ induces a no-effort equilibrium achieves a strictly lower material payoff than the incumbent type λ_c^+ in encounters where reneging costs are observed. In the proof of Theorem 2 it was shown that when $\lambda' \leq \lambda_c^+$ the derivative is equal to (the left derivative when $\lambda' = \lambda_c^+$):

$$\frac{\partial \pi\left(\lambda', \lambda_c^+\right)}{\partial \lambda'} = \frac{[\lambda_c^+]^2 (c(\lambda_c^+ + c) - 1)^2}{[1 + (c + \lambda_c^+)(c + \lambda')(c(\lambda_c^+ + c) - 2)]^3} \tag{56}$$

and that this expression is always *strictly* positive for $c \in (1, 1.25)$. In particular, this implies that

$$\lim_{\lambda' \nearrow \lambda_c^+} \frac{\partial \pi\left(\lambda', \lambda_c^+\right)}{\partial \lambda'} > 0 \tag{57}$$

The fact that the derivative of the material payoff function with respect to λ' is strictly positive for all $\lambda' < \lambda_c^+$ and that the left derivative at λ_c^+ is bounded away from zero implies that when reneging costs are observed and a two-message equilibrium is induced, there is a first order material payoff loss for a mutant with $\lambda' < \lambda_c^+$, compared to the incumbent type λ_c^+. Now, considering the case where $\lambda' > \lambda_c^+$ when $G(\lambda', \lambda_c^+)$ induces a two-message equilibrium, we note that, analogously, in the proof of Theorem 2 it was shown that when $\lambda' \geq \lambda_c^+$, the derivative of the payoff function with respect to λ' is strictly negative and that the right derivative of the payoff function, evaluated at λ_c^+, is strictly negative, i.e., the payoff increases as λ' decreases towards λ_c^+ (Mathematica code demonstrating this is in the online appendix). Therefore there is also a first-order loss for a mutant with $\lambda' > \lambda_c^+$ when reneging costs are observed. We have therefore demonstrated that any mutant achieves a strictly lower payoff in the partnership games played after reneging costs are observed than does an incumbent, i.e., $\pi\left(\lambda', \lambda_c^+\right) < \pi\left(\lambda_c^+, \lambda_c^+\right)$ for all $\lambda' \neq \lambda_c^+$, and, further, that the first-order loss of a mutant is bounded away from zero when $\lambda' \to \lambda_c^+$.

Next, we note that in the case where reneging costs are not observed, $\pi_q\left(\lambda', \lambda_c^+ | \lambda_c^+\right) - \pi_q\left(\lambda_c^+, \lambda_c^+\right)$ is bounded from above by a uniform bound. To see this, note that the maximum material payoff achievable in a partnership game is

$$\frac{1}{c} - \frac{c(\frac{1}{c})^2}{2} = \frac{1}{2c}$$

The payoff differential between a mutant of type λ', relative to the incumbents of type λ_c^+ when reneging costs are observed, can therefore be given by $q \cdot \left[\pi\left(\lambda', \lambda_c^+\right) - \pi\left(\lambda_c^+, \lambda_c^+\right)\right]$. The maximum positive payoff differential between a mutant of type λ', relative to λ_c^+ when reneging costs are not observed, is $(1-q) \cdot \frac{1}{2c}$. Therefore, the maximum payoff differential between a mutant type and an incumbent type under partial observability is:

$$q \cdot \left[\pi\left(\lambda', \lambda_c^+\right) - \pi\left(\lambda_c^+, \lambda_c^+\right)\right] + (1-q)\frac{1}{2c} \tag{58}$$

We therefore have that a mutant of type λ' is strictly outperformed by the incumbent type when Eq. (58) is strictly negative. Imposing this strict negativity and rearranging for q yields:

$$q > \frac{1}{1 + 2c[\pi\left(\lambda_c^+, \lambda_c^+\right) - \pi\left(\lambda', \lambda_c^+\right)]} \equiv \widetilde{q_{\lambda'}} \tag{59}$$

From the fact that the term in square brackets in the denominator of Eq. (59) is strictly

positive, it is immediate that $\widetilde{q_{\lambda'}} \in (0,1)$. We then define $\bar{q} \equiv sup\{\widetilde{q_{\lambda'}} : \lambda' \in \mathbb{R}^+\}$. It follows that for all $c \in (1, 1.22)$, there exists a $\bar{q}$ such that for all $q \in [\bar{q}, 1]$, $(\lambda_c^+, \lambda_c^+)$ is a strict Nash equilibrium of the population game with partial observability.

The stable population in the setting with partial observability has been proven to be identical to the population in the setting with perfect observability. Therefore, results (2) to (5) of Theorem 2 pertaining to the equilibrium play and the payoffs of this stable population, hold also in the partial observability setting. □

D.5 Proof of Proposition 1

Proof. If the incumbents have $\lambda = 0$, then they exert no effort due to Fact 1. If the incumbents have $\lambda > 0$, then, assume to the contrary that agents exert a positive level of effort on the equilibrium path. By Theorem 1, this implies that all agents send the maximum message 1 and, due to the payoff function being strictly convex, that they exert the same positive level of effort $x_i^e(1, 1, \lambda, \lambda, c) > 0$ on the equilibrium path in the second stage (see Eq. 6). Consider a mutant with zero reneging cost who sends message 1 and then exerts effort $\frac{1}{c} \cdot x_i^e(1, 1, \lambda, \lambda, c)$. It is immediate that such a mutant achieves strictly higher fitness than the incumbents because the mutant exerts the unique amount of effort that maximises the fitness, given that the partner exerts effort $x_i^e(1, 1, \lambda, \lambda, c)$. □

D.6 Proof of Proposition 2

Proof. We show that there can be no symmetric pure Nash equilibrium of the population game in which players exert no effort on the equilibrium path. Consider any symmetric population in which players have a level of reneging aversion λ and in which, in game $G(\lambda, \lambda)$, the unique equilibrium is a no-effort equilibrium and hence all players achieve a material payoff of zero, i.e., $\pi(\lambda, \lambda) = 0$. Lemma 4 implies that for any such λ, there exists λ' such that in $G(\lambda, \lambda')$ – the partnership game played where players of type λ and type λ' meet and observe their opponent's level of reneging aversion - both players exert positive effort in equilibrium and achieve strictly positive material payoffs. As any player can always guarantee a payoff of at least zero in any interaction, a player of type λ' achieves a weakly positive payoff from the partnership game played after players of types λ and λ' meet but do not observe their opponent's level of reneging aversion. Therefore, when $q > 0$ (players in a population observe each other's level of reneging aversion at least some of the time) any mutant of the type λ' achieves a strictly positive fitness in the population game with partial observability, i.e., $\pi_q\left(\lambda', \lambda | \lambda\right) > 0 = \pi(\lambda, \lambda)$. □