Promises and Endogenous Reneging Costs

Heller, Yuval and Sturrock, David

Bar Ilan University, Institute for Fiscal Studies

April 2017

Promises and Endogenous Reneging Costs*

Yuval Heller[†] David Sturrock[‡]

February 27, 2020

Final pre-print of a manuscript accepted in *Journal of Economic Theory*.

Abstract

We present a novel mechanism that explains how nonenforceable communication about future actions has the capacity to improve efficiency. We explore a two-player partnership game where each player, before choosing a level of effort to exert on a joint project, makes a cheap talk promise to his partner about his 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 costs induce second-best optimal outcomes.

Keywords: Promises, strategic complements, lying costs, input games, partnership games.

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 (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 for 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

*We thank Vincent Crawford, Rachel Griffith, Nick Netzer, Alon Raviv, Nils Rochowicz, Peyton Young, two anonymous referees and an associate editor, and seminar audiences at the University of Oxford, Bar Ilan University, the Institute for Fiscal Studies, and University of Zurich for valuable discussions and suggestions. This paper is based on David Sturrock's M.Phil. thesis, which was submitted to 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).

[‡]University College London and Institute for Fiscal Studies, david.sturrock.17@ucl.ac.uk. The author is grateful for funding from the Economic and Social Research Council (award #ES/J500112/1) and Lincoln College, University of Oxford.

in an important class of strategic interactions. Taken together, these contributions present a novel explanation for the way in which preplay communication can foster cooperation in one-shot strategic interactions when agents' interests are only partially aligned.

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 experimental evidence discussed in Section 2, and indeed much of daily experience, demonstrates 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.

We study a class of partnership games (also known as input games; see, e.g., Holmstrom, 1982; Cooper & John, 1988) with preplay communication. In the setting we examine, agents simultaneously communicate promised levels of effort, and, following this, they simultaneously choose their levels of effort from an interval. 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 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. Each player's level of reneging cost is parameterised by his level of reneging aversion. This reneging cost can be seen as representing the guilt or bad feeling that agents experience when going back on promises they have made. The subjective utility of each agent is equal to his material payoff (which depends only on the agents' effort choices) minus his reneging cost. Reneging aversion transforms what is ordinarily modeled as a cheap talk promise into a *partially* self-enforcing commitment.

We characterise the perfect equilibria of the partnership game with reneging costs. The game admits an essentially unique perfect equilibrium, the properties of which depend on the players' levels of reneging aversion. One of the main insights of the paper is that if both players have either very high or very low levels of reneging aversion then efforts are zero in equilibrium and promises play no role in promoting efficiency, but if both players have intermediate levels of reneging aversion they will promise and exert effort. The reason for this is that when both players have low levels of reneging aversion, they are very willing to go back on their promises and "undercut" their partner when choosing their level of effort and so neither is willing to promise to exert effort. When both players have sufficiently high levels of reneging aversion, they have little flexibility in their second round choice of effort. Knowing that they will have to exert effort close to what they promise, and that the same is true of their partner, both agents seek to "undercut" each other's promises, leading to no effort being promised or exerted in equilibrium.

When both players have intermediate levels of reneging aversion, they promise and exert effort. Intuitively, a player is willing to strategically promise high effort if and only if (1) he has sufficiently large reneging aversion that his promises are credible and (2) his partner has sufficiently low reneging aversion that he can react to the player's credible promise of high effort by exerting more effort himself. This implies that both agents are induced to promise high effort if and only if both players' levels of reneging aversion are neither too low nor too high. The highest symmetric level of reneging aversion consistent with both players making maximal promises has several appealing properties. This intermediate level of reneging aversion induces the best equilibrium outcome among all symmetric partnership games (and this outcome converges to the first best as the cost of effort converges to zero). The outcome is also better than that induced in a "Stackelberg" equilibrium without reneging costs (i.e., the equilibrium when effort levels are chosen sequentially), if the cost of effort is not too high.

We provide an evolutionary grounding for the emergence and stability of this intermediate level of reneging aversion by studying the endogenous determination of players' levels of reneging aversion. We consider an infinite population of players 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 level of reneging aversion before starting the two-stage partnership game described above and receiving the equilibrium payoffs. A gradual process of learning (or imitation of successful agents) makes levels of reneging aversion that induce higher payoffs to their agents become more frequent in the population. We show that the homogeneous population state in which all agents have the same intermediate level of reneging aversion is a strict equilibrium of the population game, which implies that it is evolutionarily stable. Moreover, there does not exist any other symmetric pure equilibrium of the population game, which implies that there exists no other symmetric homogeneous stable state. This demonstrates the strong tendency of evolutionary processes to select for agents who incur intermediate reneging costs. In variants and extensions of our main model, we show that our main conclusions are robust to allowing for sequential communication of promises, "one-sided" reneging costs (i.e. when there is no reneging cost if exerting more effort than promised), fixed reneging costs, and a more general form of utility function.

The paper is organised as follows. Section 2 discusses the related literature and the contributions made by this paper. Section 3 presents our model and analyses partnership games. Section 4 shows the appealing properties of intermediate reneging aversion. In Section 5 we discuss the evolutionary stability of intermediate reneging aversion. Section 6 describes the robustness of our main results to the relaxation of various assumptions in our model. We conclude in Section 7. All the appendices appear in the online supplementary material. The formal definition of a trembling-hand perfect equilibrium with a continuum of strategies is presented in Appendix A. We discuss a few technical aspects of our evolutionary interpretation in Appendix B. Additional illustrative figures are presented in Appendix C. Our extensions are formally presented in Appendix D. Formal proofs appear in Appendix E.

2 Related Literature and Contribution

Our paper contributes to several strands of literature, which we discuss in this section. The theoretical literature on signaling intentions through cheap talk explores the potential for preplay 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 nonequilibrium outcomes (Kerr & Kaufman-Gilliland, 1994; Sally, 1995; Ellingsen & Johannesson, 2004; Bicchieri & Lev-On, 2011).

In particular, players often make and keep promises to cooperate in two-player partnership games where the unique equilibrium involves no such cooperation (Charness & Dufwenberg, 2006; Vanberg, 2008; Ederer & Stremitzer, 2017; Di Bartolomeo *et al.*, 2018). We advance the theoretical analysis by presenting a novel mechanism (intermediate reneging costs) by which preplay communication is able to sustain such cooperative but apparently nonequilibrium action, and demonstrate its evolutionary stability.

Our analysis of direct psychological costs of going back on one's word is related to the theoretical literature incorporating exogenously given (and, typically, small) psychological lying costs into strategic models. Kartik *et al.* (2007) and Kartik (2009) study sender-receiver games with convex lying costs and 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 by analysing bilateral communication about agents' own future actions rather than unilateral communication about an exogenously given state of the world. Additionally, we endogenise the reneging costs, and allow them to be determined as part of a stable population state.[1]

We contribute to the literature on partnerships with strategic complementarities (pioneered by Holmstrom, 1982) by introducing reneging costs. Radner *et al.* (1986) show the capacity for repeated interaction to sustain effort in a partnership game when this is efficient but not an equilibrium of the one-shot game (see also related models 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 nonenforceable effort choices.

The theoretical model that we present rationalises the main stylised facts of the experimental literature on promising and lying. Intrinsic costs of lying or reneging on one's promise have been examined in a number of laboratory setups including: (1) trust games (Ellingsen & Johannesson, 2004; Charness & Dufwenberg, 2006; Vanberg, 2008; Ederer & Stremitzer, 2017; Di Bartolomeo *et al.*, 2018), (2) sender-receiver games (Gneezy, 2005; Sánchez-Pagés & Vorsatz, 2007; Hurkens & Kartik, 2009; Lundquist *et al.*, 2009), and (3) reporting the outcome of a private dice roll (Fischbacher & Föllmi-Heusi, 2013; Shalvi *et al.*, 2011; Gneezy *et al.*, 2018; Abeler *et al.*, 2019). Experimental evidence suggests that subjects do not always lie to gain money, even when their doing so cannot be detected.[2] In experiments on promises, subjects only sometimes renege on promises to carry out actions that are socially beneficial but reduce their own payoff and, on average, achieve more efficient outcomes than when promises cannot be made (Charness & Dufwenberg, 2006; Vanberg, 2008; Ederer & Stremitzer, 2017; Di Bartolomeo *et al.*, 2018).

We defer further discussion of the relation between our model and the experimental evidence to Section 7. We note here that the stylised facts from these experiments suggest that the intrinsic costs of lying/reneging are intermediate, and are increasing (potentially convexly) with one or more of the following : (I) the difference between the reported/exerted outcome and the true/promised outcome, (II) the damage caused to the partner by the lying/reneging, and (III) others' perceptions of the agent's

[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 reneging 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 reneging costs *per se*.

[2] In the case of reporting a private dice roll, Abeler *et al.*'s Finding 1 demonstrates that subjects obtain only about a quarter of the payoff they could obtain by reporting the die's maximal outcome. When subjects lie, they sometimes do so by using a nonmaximal lie (see, e.g., Abeler *et al.*, Finding 5), suggesting that bigger lies induce higher intrinsic costs.

behaviour (e.g., the agent derives disutility in proportion to the extent to which he is perceived to cheat; see Dufwenberg & Dufwenberg, 2018). In our model the intrinsic cost of reneging is proportional to the difference between the promised and exerted effort, which directly captures (I). In a richer model, in which others observe effort with some random noise, this difference can also capture (III). In Section 6, we describe a variant of our model in which an intrinsic cost of reneging is incurred only if the promised effort is smaller than the exerted effort (formal analysis of this variant is in Appendix D). This captures (II), as in this variant the partner suffers a utility loss proportional to the extent to which promised effort was higher than exerted effort. Our model therefore captures the central findings of these studies, but also allows the level of reneging aversion to be endogenously determined by an evolutionary process, providing a theoretical foundation and explanation for the stylised experimental facts.

In our theoretical exploration of the potential evolutionary determinants of reneging aversion, we build on the "indirect" evolutionary approach, 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), and Alger & Weibull (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.[3] Second, our main result is qualitatively different from the stylised result in this literature, according to which if preferences are observable, 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.

Finally, 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 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 his interest in making a strong commitment in order to induce high effort from his partner, against his conflicting desire to retain some flexibility to exert less effort.

3 The Partnership Game

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

[3] Heifetz *et al.* (2007a,b) find that under payoff-monotonic evolution of preferences with perfect observability, any "distortion" of preferences (divergences between the subjective utility and the material payoffs) will be stable at some non-zero level. Their setup imposes particular assumptions not satisfied by our model but we nevertheless show results that are consistent with their findings and, in addition, explicitly characterise the unique stable level of reneging aversion.

3.1 Model

There are two players (i and j) and two stages of the game. In the first stage, both players simultaneously send a message $s_k \in [0,1]$ 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, $x_k \in [0,1]$.

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.

We focus on a specific family of quadratic payoff functions. We define the material payoff of a player who exerts effort x_i and whose partner exerts effort x_j as the following function:

$$\pi(x_i, x_j, c) = x_i \cdot x_j - \frac{c \cdot x_i^2}{2} \quad : \quad c \in (1,2)\,. \tag{1}$$

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 $c \in (1,2)$ governs the cost of effort.[4]

The subjective utility of each player i is defined as follows:

$$U_i(x_i, x_j, s_i, c) = x_i \cdot x_j - \frac{c \cdot x_i^2}{2} - \frac{\lambda_i}{2}(s_i - x_i)^2. \tag{2}$$

Subjective utility is the sum of a player's material payoff and a term representing the psychological cost of breaking a promise (reneging). 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*. We assume that each player perfectly observes his partner's level of reneging aversion, i.e., the parameters λ_i, λ_j are common knowledge.

We extend the material payoffs and the subjective utility to mixed strategies in the usual linear way (i.e., players are expected utility maximisers). It turns out that, essentially, all perfect equilibria are pure; thus, we focus in the main text on pure strategies. (We formally deal with mixed strategies in Lemma 1 and Footnote 15.)

3.2 Unique Second-Stage Equilibrium and First-Stage Best-Reply Functions

In the second stage of the game, player i's first-order condition for his choice of x_i is given by[5]

$$x_j - cx_i + \lambda_i(s_i - x_i) = 0. \tag{3}$$

[4] We restrict attention to $c \in (1,2)$ as this is the interval in which (1) players exerting maximal effort is efficient and (2), as shown below, the game with simultaneous effort choices encourages shirking above an effort of zero. Note that when $c > 2$, the efficient outcome is for both players to exert zero effort. When $c < 1$, the unique Nash equilibrium in the game with simultaneous effort choices is $x_i = x_j = 1$.

[5] 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.

The strict concavity of the utility function in x_i implies that the second-stage best reply is a unique pure strategy, which implies that we can focus in the second stage, without loss of generality, on pure strategies. The unique best-reply strategy is given by the function

$$x_i^*(x_j, s_i, s_j, \lambda_i, \lambda_j, c) = \frac{x_j + \lambda_i s_i}{c + \lambda_i}. \tag{4}$$

This function is a weighted average of the level of effort which maximises the player's material payoff ($\frac{x_j}{c}$) and that which fulfills his promise (s_i), with greater weight placed on s_i as reneging aversion increases. This equation therefore embodies the way that a player trades-off undercutting (exerting less effort than) his partner and coming closer to keeping his promise.

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

$$x_i^e(s_i, s_j, \lambda_i, \lambda_j, c) = \frac{(c + \lambda_j)\lambda_i s_i + \lambda_j s_j}{(c + \lambda_i)(c + \lambda_j) - 1}. \tag{5}$$

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 at the same time they 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 toward $\frac{x_j}{c}$, 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[6] s_i, s_j, and 0. As a player's level of reneging aversion increases, he will exert effort closer to his own promise.

The subgame-perfect equilibrium of the game is easily obtained using backward induction. Given the unique Nash equilibrium strategies in each subgame, we can derive the player's utility $U_i(s_i, s_j, c)$ as a function of the messages sent by the agent and his partner (assuming that both players follow the unique Nash equilibrium in the second stage of the game).

$$U_i(s_i, s_j, c) \equiv U_i\left(x_i^e(s_i, s_j, \lambda_i, \lambda_j, c), x_j^e(s_i, s_j, \lambda_i, \lambda_j, c), s_i, c\right). \tag{6}$$

Clearly, if $\lambda_i = 0$, then a player's choice of message has no impact upon his own or his partner's choices and any message is a best reply. When $\lambda_i > 0$, it turns out that the derived utility function $U_i(s_i, s_j, c)$ leads to a unique pure best reply in all but a measure zero of cases,[7] which implies that, without loss of

[6] 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.

[7] The choice of best reply in these measure-zero cases plays no role in our analysis.

generality, we can focus on pure strategies (formal details for this argument are presented in the proof of Proposition 1 in Appendix E.2).

Our first result characterises the first-stage best-reply functions. Let $s_i^*(s_j|\lambda_i,\lambda_j,c)$ denote the best reply of agent i (with reneging aversion λ_i) to a partner's message of s_j, where the cost of effort is c. We show that there exists a threshold $\bar{\lambda}_c \equiv \frac{2-c^2}{c-1}$, such that an agent overcuts his partner (i.e., $s_i^*(s_j|\lambda_i,\lambda_j,c) > s_j$ for each $s_j \in (0,1)$) if and only if the partner's reneging aversion is below this threshold and the agent's reneging aversion is sufficiently high. Formally:

Proposition 1. *There exists a function $g : R^+ \times [1,2] \to R^+$, such that:*

1. *Overcutting: $s_i^*(s_j|\lambda_i,\lambda_j,c) > s_j$ if $\lambda_j < \frac{2-c^2}{c-1}$, $\lambda_i > g(\lambda_j,c)$ and $s_j \in (0,1)$.*
2. *Undercutting: $s_i^*(s_j|\lambda_i,\lambda_j,c) < s_j$ if $s_j > 0$ and $\left(\lambda_j \geq \frac{2-c^2}{c-1} \text{ or } \lambda_i \in (0, g(\lambda_j,c))\right)$.*

Observe that $c \geq \sqrt{2}$ implies $\frac{2-c^2}{c-1} \leq 0$ and that each agent will undercut his partner's message. The division of the parameter space into these best-reply types (undercutting vs. overcutting) is illustrated in Figure 1 for the effort costs of $c = 1.1$ and $c = 1.2$.

3.3 Unique Perfect Equilibrium

We now characterise the subgame-perfect equilibria of the partnership game. Recall that a strategy profile $\left(s_i^*, s_j^*, x_i^*, x_j^*\right)$ (where $s_i^*, s_j^* \in [0,1]$ are the first-stage promises and the functions $x_i^*, x_j^* : [0,1]^2 \to [0,1]$ describe the second-stage efforts as a function of the observed promises) is a subgame-perfect equilibrium if for each player i (1) $x_i^*(\overrightarrow{s}) \equiv x_i^*(s_i,s_j) = x_i^e(s_i,s_j,\lambda_i,\lambda_j,c)$ (i.e., best replying in the second stage), and (2) $U_i(s_i^*,s_j^*,c) \geq U_i(s_i',s_j^*,c)$ for each message $s_i' \in [0,1]$ (i.e., best replying in the first stage), where the derived utility $U_i(s_i^*,s_j^*,c)$ is as defined in (6).

We show that all subgame-perfect equilibria can be classified into three types:

1. *Maximum-message equilibrium,* in which agents send maximal promises, i.e., $s_i^* = s_j^* = 1$.
2. *No-effort equilibrium*, in which agents exert no effort, i.e., $x_i^*\left(\overrightarrow{s^*}\right) = x_j^*\left(\overrightarrow{s^*}\right) = 0$. In this equilibrium any agent with a positive reneging aversion promises nothing, i.e., $\lambda_i > 0 \Rightarrow s_i^* = 0$.
3. *Two-message equilibrium*, in which the agent with higher reneging aversion sends the maximal message, while his partner undercuts the agent's message, i.e., either $s_i = 1 > s_j$ or $s_j = 1 > s_i$.

In some parameterisations of the game, the subgame-perfect equilibrium is unique. In all the remaining cases (except the "measure zero" set of pairs with multiple equilibria discussed below), the game admits two subgame-perfect equilibria, where only one of these equilibria satisfies trembling-hand perfection (see the formal definition, à la Selten, 1975; Simon & Stinchcombe, 1995, in Appendix A). The imperfect equilibrium is characterised by each agent sending a zero message. However, any small perturbation (e.g., with a small probability, ϵ, each player trembles and sends a positive message) induces at least one of the agents to overcut his partner, and this equilibrium is eliminated from the perturbed game.

Theorem 1, below, shows that:

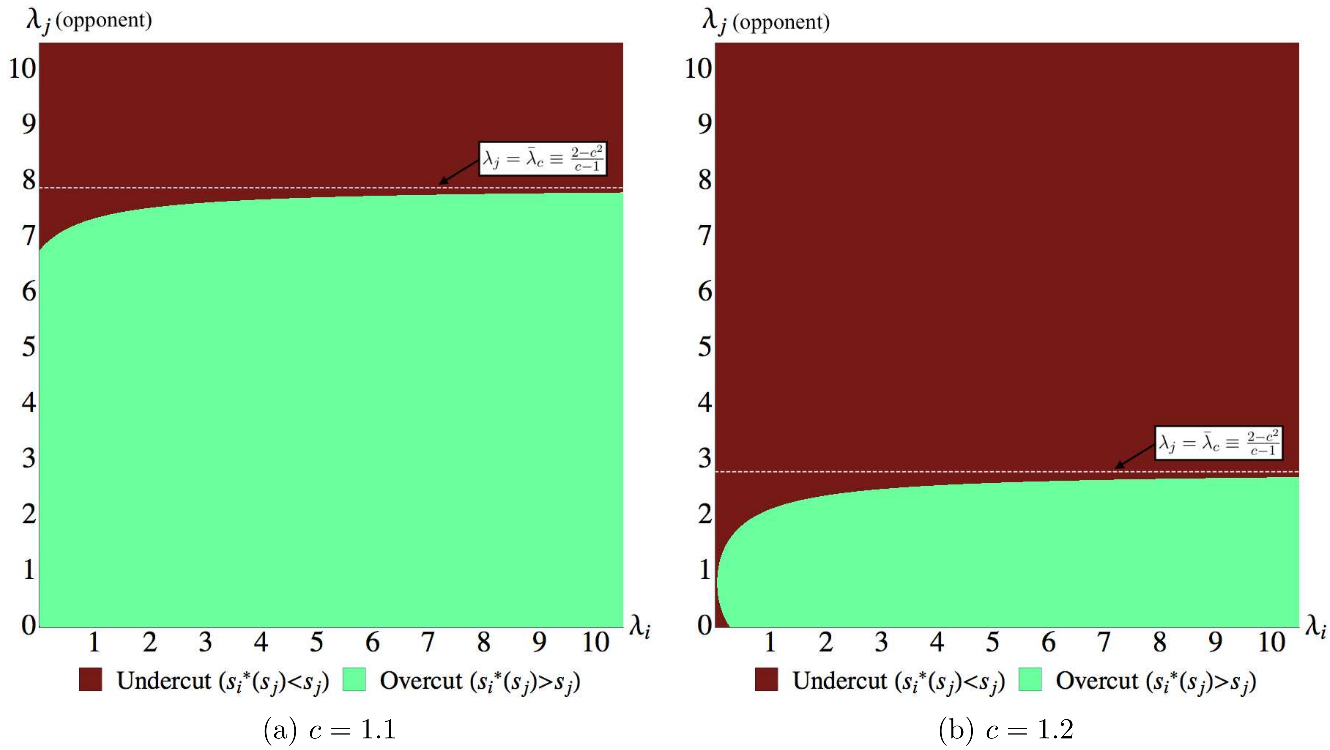


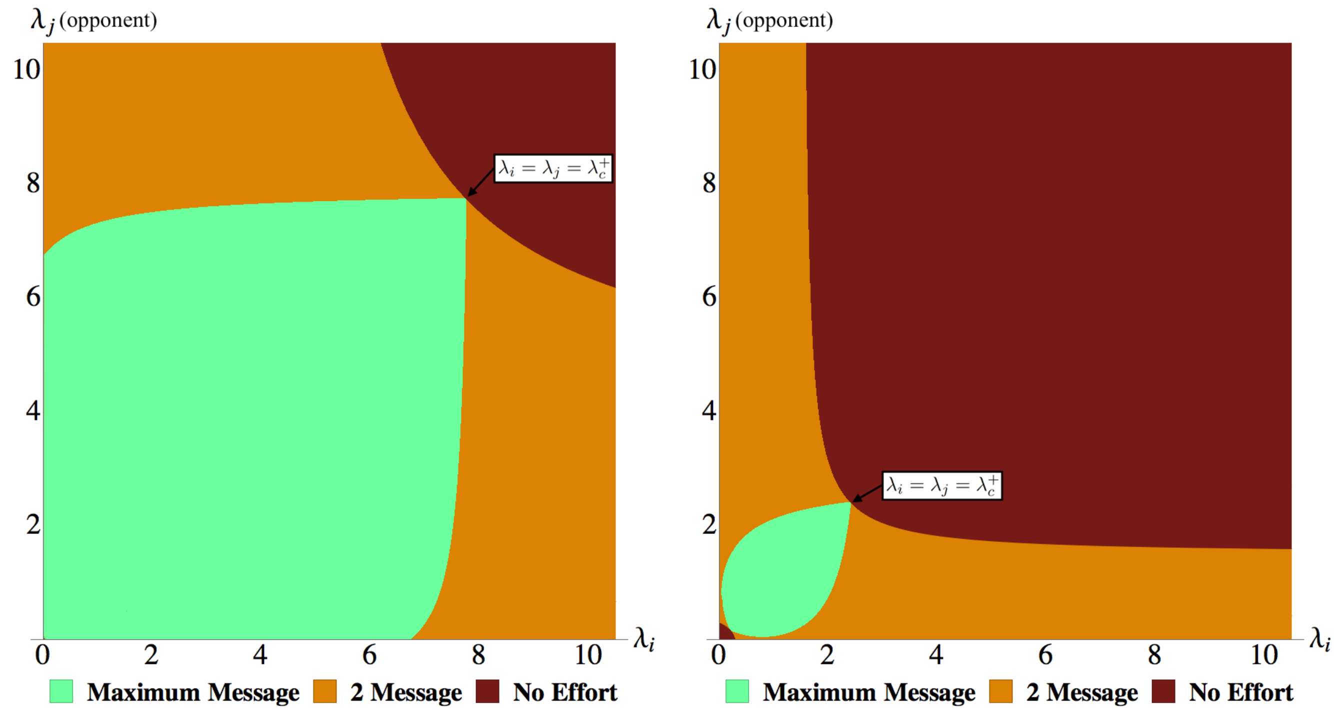
Figure 1: **Best-Reply Types for Player *i* in a Reneging Aversion Parameter Space**. The x axis in each figure presents the player's level of reneging aversion (λ_i) and the y axis presents the partner's level of reneging aversion (λ_j). The left panel deals with a cost of effort of $c = 1.1$ and the right figure deals with $c = 1.2$. The dark area in each panel is the region in which player i's best reply is to undercut his partner, i.e., $s_i^*(s_j) < s_j$; the light area in each panel is the region in which player i's best reply is to overcut his partner, i.e., $s_i^*(s_j) > s_j$. The dashed line in each figure shows the value $\lambda_j = \bar{\lambda}_c \equiv \frac{2-c^2}{c-1} > 0$ presented in Proposition 1, above which player i's best reply is to undercut his partner regardless of the value of λ_i.

1. There is a convex symmetric region of intermediate levels of reneging aversion, Λ^c_{max}, in which the game admits only a maximum-message equilibrium. This region is nonempty if and only if $c < 1.25$ and, in this case, there is a symmetric point $(\lambda_c^+, \lambda_c^+)$ in the boundary of this region, which is higher than any point in the region.

2. There are two disjoint areas in which one agent has a sufficiently low level of reneging aversion and his partner has a sufficiently high level of reneging aversion, and the game admits only a two-message equilibrium. This region, Λ^c_{2ms}, is nonempty if and only if $c < \sqrt{2}$.

3. In the remaining region, Λ^c_{0ef}, the partnership game admits only a no-effort equilibrium. This region includes two areas: (1) an area in which both agents have sufficiently high levels of reneging aversion and (2) an area in which both agents have sufficiently low levels of reneging aversion.

Figure 2 illustrates the division of the reneging aversion parameter space into the three classes of unique perfect equilibria for the effort costs of $c = 1.1$ and $c = 1.2$.



(a) $c = 1.1$ (b) $c = 1.2$

Figure 2: **Unique Perfect Equilibrium Types in a Reneging Aversion Parameter Space**. The x axis in each figure presents the player's level of reneging aversion (λ_i) and the y axis presents the partner's level of reneging aversion (λ_j). The left panel deals with a cost of effort of $c = 1.1$ and the right panel deals with $c = 1.2$. The dark areas in each figure are the regions in which both agents exert no effort in equilibrium ("No Effort"). The light area in each figure is the region in which both agents promise maximal efforts in the unique perfect equilibrium ("Maximum Message"). The remaining areas are the regions in which one of the agents sends a maximal promise ("2 Message").

Let $PE(\lambda_i, \lambda_j, c)$ denote the set of all trembling-hand perfect equilibria (as defined Appendix A) in of the partnership game with levels of reneging aversion λ_i and λ_j and effort cost c. Let $Cl(\Lambda)$ denote closure of the set Λ, i.e., the set Λ together with all its limit points. A formal statement of our result is as follows.

Theorem 1. *For each $c > 1$, there exist disjoint symmetric sets $\Lambda^c_{0ef}, \Lambda^c_{max}, \Lambda^c_{2ms} \subseteq [0,\infty)^2$ with an exhaustive union of closures (i.e., $Cl\left(\Lambda^c_{0ef}\right) \cup Cl\left(\Lambda^c_{max}\right) \cup Cl\left(\Lambda^c_{2ms}\right) = [0,\infty)^2$) that satisfy the following properties:*

1. *Region of maximum-message equilibrium Λ^c_{max}:*

   (a) *If $(\lambda_i, \lambda_j) \in \Lambda^c_{max}$ then there is a unique $\left(s^*_i, s^*_j, x^*_i, x^*_j\right) \in PE(\lambda_i, \lambda_j, c)$, and $s^*_i = s^*_j = 1$.*

   (b) *Λ^c_{max} is a convex set, which is nonempty iff $c \in (1, 1.25)$.*

   (c) *For each $c < 1.25$, there exists $\left(\lambda^+_c, \lambda^+_c\right) \in Cl\left(\Lambda^c_{max}\right)$ such that $(\lambda_i, \lambda_j) \in \Lambda^c_{max} \Rightarrow \lambda_i, \lambda_j < \lambda^+_c$.*

2. *Region of two-message equilibrium Λ^c_{2ms}:*

(a) *If* $(\lambda_i, \lambda_j) \in \Lambda^c_{2ms}$ *and* $\lambda_j < \lambda_i$ *then there is a unique perfect equilibrium* $\left(s^*_i, s^*_j, x^*_i, x^*_j\right) \in PE(\lambda_i, \lambda_j, c)$*, and* $s^*_j < s^*_i = 1$.

(b) Λ^c_{2ms} *is nonempty iff* $c \in \left(1, \sqrt{2}\right)$.

(c) *If* $0 < \lambda_j < \frac{2}{c} - c$ *and* λ_i *is sufficiently high, then* $(\lambda_i, \lambda_j) \in \Lambda^c_{2ms}$.

3. *Region of no-effort equilibrium* Λ^c_{0ef}*:*

(a) *If* $(\lambda_i, \lambda_j) \in \Lambda^c_{0ef}$ *then there is a unique perfect equilibrium* $\left(s^*_i, s^*_j, x^*_i, x^*_j\right) \in PE(\lambda_i, \lambda_j, c)$*, and* $x^*_i(\overrightarrow{s}^*) = x^*_j(\overrightarrow{s}^*) = s^*_i = s^*_j = 0$.

(b) *There exist* $0 < \underline{\lambda}_c < \overline{\lambda}_c$ *such that* $(\lambda_i, \lambda_j) \in \Lambda^c_{0ef}$ *if either* $0 < \lambda_i, \lambda_j < \underline{\lambda}_c$ *or* $\lambda_i, \lambda_j > \overline{\lambda}_c$.

The intuition for these results is as follows. When both λ_i and λ_j are low or when both are high, the unique equilibrium is a no-effort equilibrium. Too low reneging costs induce too little commitment power and, as a result, each agent undercuts his partner's effort in the second round regardless of the promises. Too high reneging costs leave too little flexibility for the second round, which induces each agent to undercut his partner's promise in the first round.

When one player has a high level of reneging aversion and the other a low level and $c \in \left(1, \sqrt{2}\right)$, the unique equilibrium is a two-message equilibrium. The intuition is that only the agent with the high reneging cost has a substantial commitment power, while their partner's promise has a very small impact on either player's effort choice. As a result, the agent with the high reneging cost is essentially a Stackelberg leader (he essentially chooses his effort by the committing promise he makes in the first round), while the partner is essentially a Stackelberg follower (her promise in the first round has little influence on her choice of effort in the second round). The lower the cost of effort is, the higher the effort that the Stackelberg follower will exert in reply to a given promise by the leader. When the effort cost is low enough, it therefore becomes worthwhile for the leader to make the promise of high effort. Similarly to a standard setup of duopoly with strategic complements, the Stackelberg follower has the higher payoff.

Finally if both players' levels of reneging aversion are intermediate (and sufficiently similar) and $c \in (1, 1.25)$, then we have the maximum-message equilibrium. If the partner's level of reneging aversion is not too high, the indirect benefit of overcutting the partner's message (i.e. the greater material payoff that is achieved because promising more effort induces the partner to exert more effort in the second stage) is increasing in the agent's level of reneging aversion, as higher reneging aversion makes his promise more credible. If the agent's level of reneging aversion is sufficiently high, this benefit outweighs the direct cost of restricting his ability to shirk in the second stage. Therefore, if both players have a level of reneging aversion that is high enough to give them committing power but is not so high that they do not have some flexibility in the second stage, they will wish to overcut each other. This happens in a convex region of intermediate levels of reneging aversion. In this region, both players are sufficiently bound by their message to be able to strategically induce high effort in their partner, but are also flexible enough to respond to their partner's promise.

4 Appealing Properties of Intermediate Reneging Aversion, λ_c^+

Induced population game Theorem 1 has shown that almost all partnership games have a unique trembling-hand perfect equilibrium. Multiple perfect equilibria may occur only on a "measure-zero" of pairs of λ_i, λ_j that are located on the boundaries between the open sets Λ_{0ef}^c, Λ_{max}^c, and Λ_{2ms}^c. In the remaining "measure-one" set of pairs of levels of reneging aversion, we define $\pi_c(\lambda_i, \lambda_j)$ to be the unique trembling-hand perfect equilibrium payoff of an agent with reneging aversion λ_i who is matched with a partner with reneging aversion λ_j.

Recall, that, for each $c \in (1, 1.25)$, one of the pairs in these measure-zero boundaries is $(\lambda_c^+, \lambda_c^+)$, which is the upper limit of all pairs in the set Λ_{max}^c of intermediate levels of reneging aversion that induce maximal messages. In Corollary 2 of Theorem 1 (Appendix E.5), we show that the pair of levels of reneging aversion $(\lambda_c^+, \lambda_c^+)$ induces a continuum of perfect equilibria. Specifically, for each message $s^* \in [0, 1]$, there is a trembling-hand perfect equilibrium in which both agents send message s^*. We define $\pi_c(\lambda_c^+, \lambda_c^+)$ as the highest equilibrium payoff among these equilibria (i.e., the payoff induced in the equilibrium in which the agents send the maximal message, $s^* = 1$). We discuss this equilibrium selection in Remark 2 below. Given any other pair (λ_i, λ_j) with multiple equilibria, we can apply any arbitrary equilibrium selection function (without affecting our results), and we let $\pi_c(\lambda_i, \lambda_j)$ $(U_c(\lambda_i, \lambda_j))$ be the material (subjective) payoff induced by the arbitrarily selected equilibrium.

The payoff function $\pi_c : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ defined above induces a symmetric two-player population game $\Gamma = (\mathbb{R}^+, \pi_c)$. This population game can be interpreted as being played between two principals, where each principal simultaneously chooses a reneging cost for his agent, the two agents are matched to play the partnership game (where each agent observes his partner's reneging cost), and they play the trembling-hand perfect equilibrium of the partnership game (applying the equilibrium selection function mentioned above when multiple perfect equilibria exist). In Section 5 we discuss an evolutionary interpretation of the population game and of our results.

A pure (mixed) strategy in this game corresponds to a level of reneging aversion (a distribution over levels of reneging aversion). We say that (λ, λ) is a symmetric (strict) pure Nash equilibrium of the population game if $\pi_c(\lambda, \lambda) \geq \pi_c(\lambda', \lambda)$ $(\pi_c(\lambda, \lambda) > \pi_c(\lambda', \lambda))$ for each $\lambda' \neq \lambda$.

Appealing properties of λ_c^+ In the following result we focus on the case of low costs of effort, in which maximum-message equilibria exist (i.e., we focus on the case of $c < 1.25$). We show that the maximum-message equilibrium induced by the symmetric pair of intermediate levels of reneging aversion $(\lambda_c^+, \lambda_c^+)$ has various appealing properties:

1. "Second-best" symmetric outcome: The trembling-hand perfect equilibrium induced by $(\lambda_c^+, \lambda_c^+)$ yields the best equilibrium outcome among all trembling-hand perfect equilibrium outcomes of symmetric partnership games. This holds both for the material payoff, as well as for the subjective payoff, i.e., $\pi_c(\lambda_c^+, \lambda_c^+) > \pi_c\left(\lambda', \lambda'\right)$ and $U_c(\lambda_c^+, \lambda_c^+) > U_c\left(\lambda', \lambda'\right)$ for any $\lambda' \neq \lambda_c^+$.

2. As c converges to 1, the equilibrium material payoff and equilibrium subjective payoff both converge to the maximum feasible payoff, achieved by both agents promising and exerting the maximum effort of one. This maximal payoff is equal to $1 - \frac{c}{2}$, and it converges to 0.5 as c converges to 1.

3. It is a strict equilibrium of the population game (i.e., $(\pi_c(\lambda_c^+, \lambda_c^+) > \pi_c(\lambda', \lambda_c^+))$.

4. The equilibrium material payoff $\pi_c(\lambda_c^+, \lambda_c^+)$ is larger than the average of the two players' payoffs induced in a "Stackelberg" equilibrium without reneging costs (i.e., the equilibrium when effort levels are chosen sequentially), if the cost of effort is low ($c < 1.22$).

5. The population game does not admit any other symmetric pure equilibrium.

Theorem 2. *Fix $c \in (1, 1.25)$. Let $(\lambda_c^+, \lambda_c^+)$ be the highest symmetric pair of levels of reneging aversion inducing a maximum-message equilibrium. The equilibrium induced by $(\lambda_c^+, \lambda_c^+)$ satisfy:*

1. *"Second-best" symmetric outcome: $\pi_c(\lambda_c^+, \lambda_c^+) > \pi_c\left(\lambda', \lambda'\right)$ and $U_c(\lambda_c^+, \lambda_c^+) > U_c\left(\lambda', \lambda'\right)$* for any *$\lambda' \neq \lambda_c^+$.*

2. *Convergence to "first-best" outcome: $\lim_{c \to 1} \pi_c(\lambda_c^+, \lambda_c^+) = \lim_{c \to 1} U_c(\lambda_c^+, \lambda_c^+) = \frac{1}{2}$ (which is the best symmetric feasible payoff).*

3. *Strict equilibrium of the population game: $\pi_c(\lambda_c^+, \lambda_c^+) > \pi_c(\lambda', \lambda_c^+)$ for each $\lambda' \neq \lambda_c^+$.*

4. *Better outcome than the sequential-game equilibrium outcome when effort costs are low: Let π_i^s be the payoff to player i in the unique equilibrium of the game where efforts are chosen sequentially (and there are no reneging costs). Then if $c < 1.22$, $\pi_c(\lambda_c^+, \lambda_c^+) > \frac{1}{2} \cdot \left(\pi_i^s + \pi_j^s\right)$.*

5. *Unique pure symmetric equilibrium: If $c < 1.24$ and there is λ^* such that $\pi_c(\lambda^*, \lambda^*) \geq \pi_c(\lambda', \lambda^*)$ for each λ', then*[8] *$\lambda^* = \lambda_c^+$.*

Sketch of proof. □

1. *"Second-best" symmetric outcome:* Theorem 2 implies that any higher symmetric level of reneging aversion $\lambda > \lambda_c^+$ induces the no-effort equilibrium with zero payoffs. We show that the weaker commitment power induced by lower symmetric levels of reneging aversion $\lambda < \lambda_c^+$ induces agents to exert less effort relative to the equilibrium induced by λ_c^+, and thus to achieve a lower material payoff. Higher levels of reneging aversion have both a positive "direct" effect of increasing players' chosen levels of efforts (due to the the desire to avoid the reneging cost) and an "indirect" strategic effect as the players each anticipate the higher effort that will be exerted by the other. These positive impacts on effort levels are great enough to outweigh the increased psychological cost of reneging, such that λ_c^+ also induces the "second-best" level of subjective utility.

2. *Convergence to "first-best" outcome:* Players material payoffs are maximised if they exert $\frac{1}{c}$ times their partner's effort. As c converges to one, this converges to the partner's effort choice and so the incentives to shirk are diminished, and, as a result, the equilibrium effort levels converge to one.

3. *Strict equilibrium of the population game:* For any $\lambda > \lambda_c^+$ the game induced by (λ, λ_c^+) admits only the no-effort equilibrium with zero payoffs. When $\lambda < \lambda_c^+$, the lower commitment power of

[8] Our proof technique allows us to prove the uniqueness results only for $c \in (1, 1.24)$. Numeric simulations suggest that the result also holds for $c \in (1.24, 1.25)$.

the λ-player implies that the players exert less effort in equilibrium, and the payoff of both players is lower than the equilibrium payoff of the game induced by $(\lambda_c^+, \lambda_c^+)$.

4. *Better outcome than the sequential-game equilibrium outcome:* In the sequential effort setting (with no reneging costs) the "Stackelberg leader" will choose effort level 1 and the follower will choose effort level $\frac{1}{c}$. In the equilibrium outcome under λ_c^+, both agents promise to exert an effort of 1, and in the second stage due to the substantial reneging costs, the agents choose an effort that is much closer to one than to $\frac{1}{c}$, when the cost of effort is sufficiently low.[9] The intuition for why the average payoff with reneging costs converges "faster" to the first best as effort costs decrease (as compared to sequential effort choices) is that, whereas with sequential choices lower effort costs mean simply that the second player has a smaller incentive to shirk, and so puts in more effort as effort costs fall (with no change in the leader's action), with communication there is a positive reinforcing mechanism whereby the knowledge that his partner is going to put in more effort means that a player will put in more effort himself, leading his partner to exert more effort, and so on.

5. *Unique pure symmetric equilibrium:* We show that an agent can gain by having a higher reneging cost than his partner for every level of the partner's reneging aversion $\lambda < \lambda_c^+$, which implies that (λ, λ) is not a Nash equilibrium of the population game for any $\lambda < \lambda_c^+$. The intuition is that the indirect gain induced by the stronger commitment power of the agent (which induces the partner to exert more effort) outweighs the loss induced by the smaller flexibility in the second stage. Observe that Theorem 1 implies that for any $\lambda > \lambda_c^+$ the game induced by (λ, λ) admits the no-effort equilibrium, which yields zero payoffs. One can show that if an agent deviates to a sufficiently low level of reneging aversion, then the players play a two-message equilibrium that yields the deviator a positive payoff. Thus (λ, λ) is not a Nash equilibrium for any $\lambda > \lambda_c^+$.

5 Evolutionary Interpretation of Our Results

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 level of reneging aversion before starting the two-stage partnership game described above. We assume that in each partnership game, the players play the unique perfect equilibrium (and they follow the equilibrium selection function described above when there are multiple equilibria).

Consider first the case in which the set of feasible levels of reneging aversion are discrete (e.g., the set of feasible λs are 0, 0.01, 0.02, 0.03, ...). This discreteness might be due to having a finite, albeit very large, set of feasible genotypes in biological evolutionary processes, or due to some constraints in social evolutionary processes (e.g., each agent follows a simple rule of thumb to guide his behaviour, and the set of simple rules is finite). It is well known that stable population states in this setup correspond to symmetric equilibria of the population game, given a smooth and payoff-monotone dynamic process by which the levels of reneging aversion in the population evolve, such as the replicator dynamics (Taylor & Jonker, 1978; see Weibull, 1995; Sandholm, 2010 for a textbook introduction). Specifically:

[9] The higher payoff in the equilibrium induced by $(\lambda_c^+, \lambda_c^+)$ holds for any $c < 1.22$, but it does not hold for $c \in (1.22, 1.25)$.

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

Thus, part (3) of Theorem 2 (which states that $(\lambda_c^+, \lambda_c^+)$ is a strict equilibrium of the population game) implies that the homogeneous population state in which all agents have the same intermediate level of reneging aversion λ_c^+ is dynamically stable. Part (5) of Theorem 2 (which states that the population game does not admit any other symmetric pure equilibrium) implies that this state is the unique homogeneous stable state. This suggests a tendency of evolutionary processes to select the level of reneging aversion λ_c^+ when players each observe their partner's type.

When the set of feasible levels of reneging aversion is a continuum (i.e., without the discretization described above), then, as argued by Eshel (1983) and Oechssler & Riedel (2002), a 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 Appendix B we discuss the relevant notions of continuous stability proposed by these authors, and briefly sketch how our equilibrium can satisfy a somewhat weaker version of these these refinements.

Remark 2 (Alternative equilibrium selection). Theorem 2 depends on the equilibrium selection function choosing the most efficient equilibrium in the game in which both agents have reneging aversion λ_c^+. In what follows, we discuss two informal arguments that could justify such an assumption.

1. **Discrete set of feasible levels of reneging aversion**: Consider the setup in which the set of feasible levels of reneging aversion are discrete (as described above). Suppose that none of these discrete levels of reneging aversion is λ_c^+. Let $\lambda^* = \lambda_c^+ - \varepsilon$ be the largest level of reneging aversion such that (λ^*, λ^*) induces a maximum message equilibrium in the partnership game. By Theorem 1, this equilibrium is unique. Further, as payoffs and utilities are continuous in reneging aversion within each of the three regions of the parameter space, (λ^*, λ^*) will be an equilibrium of the population game, given ε is sufficiently small. Though we do not provide a formal analysis, it is intuitive that results analogous to Theorem 2, with λ^* in place of λ_c^+, should therefore hold in the discrete case without requiring the equilibrium selection function assumption.

2. **Focality of the efficient equilibrium**: We do not formalise the dynamic process leading a population to the homogeneous state in which everyone has the level of reneging aversion λ_c^+. Intuitively, one plausible way in which the population can converge to λ_c^+ is as follows. The populations begins in a state in which agents have a lower intermediate level of reneging aversion $\lambda < \lambda_c^+$, and they

play the unique perfect equilibrium induced by the state λ in which the messages are maximal. As argued in the proof of part (4), the state λ is vulnerable to a few agents ("mutants") experimenting with a higher level of reneging aversion $\lambda' \in (\lambda, \lambda_c^+)$, where the mutants also play the unique perfect equilibrium (with maximal messages) against the incumbents. Such a sequence of invasions of mutants will take the population to the state in which all agents have a reneging aversion of λ_c^+, and along this sequence the agents play the unique perfect equilibrium (with maximal messages). Arguably, it is plausible that after the population converges to every agent having reneging aversion λ_c^+ (and multiple equilibria exist) the agents will continue to play the "focal" equilibrium, which is similar to the unique maximum-message equilibrium played against the previous incumbents with $\lambda < \lambda_c^+$.

Remark 3 (Delegation interpretation of our results). One can also apply our result to a setup of strategic delegation. The literature on strategic delegation (see, e.g., Fershtman *et al.* 1991; Dufwenberg & Güth 1999; Fershtman & Gneezy 2001) deals with players who strategically use other agents to play on their behalf, where the agents so used may have different preferences than the players using them. One can interpret our setup as an environment in which the principals can choose agents to play the partnership games on their behalf. The principals receive the material payoff from the partnership game (i.e. principals have no reneging aversion) and they can choose to delegate to one of a number of agents, who have different levels of reneging aversion and whose payoff is the corresponding subjective utility. The results of this section show the strong tendency of principals in this setup to choose agents with the intermediate level of reneging aversion λ_c^+.

6 Variants and Extensions

Our main model makes the following assumptions: (1) agents send their promises simultaneously, (2) an agent incurs a reneging cost when his effort is higher than his promise (as well as when it is lower), (3) the reneging costs of the agent are continuous around zero reneging, and (4) the utility function has a specific quadratic form. In this section we describe how our main results are robust to relaxation each of these assumptions. Formal statement of the results, and their proofs, are confined to Appendix D and Appendix E, respectively.

We examine a variant of the partnership game in which there is sequential, rather than simultaneous, making of promises. In stage 0 of the game, nature chooses at random which player (denoted by i) will be the first to communicate (where each player has a probability of 50% of being the first). In stage 1, player i sends a message $s_i \in [0,1]$ to player j and player j observes this. In stage 2, player j chooses a message $s_j \in [0,1]$ to send to player i and player i observes this. In stage 3, the players simultaneously choose effort levels $x_i, x_j \in [0,1]$. Utility levels and material payoffs are the same functions of messages and effort levels as in the baseline model with simultaneous communication.

We show (Proposition 2) that the equilibrium induced by the partnership game with sequential communication in which both players have reneging aversion λ_c^+ induces a unique perfect equilibrium that satisfies the same appealing properties as in the baseline model.[10] The reason for this is that

[10]We do not determine whether the population game with sequential communication admits additional symmetric pure

the combinations of levels of reneging aversion that induce maximum-message equilibria or two-message equilibria in the game with simultaneous communication induce equilibria with the same promises, effort levels and payoffs under sequential communication.

Next, we consider "one-sided" reneging costs where an agent incurs a cost of reneging only if their effort is less than was promised. The subjective utility function of each player i is redefined as follows:

$$U_i(x_i, x_j, s_i, c) = x_i \cdot x_j - \frac{c \cdot x_i^2}{2} - \mathbf{1}_{s_i > x_i} \frac{\lambda_i}{2}(s_i - x_i)^2. \tag{7}$$

This "one-sided" cost can be seen as representing "guilt" that is proportional to the damage caused to the partner due to the fact that the agent broke his promise (see, e.g., Charness & Dufwenberg, 2006). When the agent exerts more effort than promised, there is no damage to the partner, and thus no guilt. When the agent i's exerted effort is less than his promise, then the loss to the partner is $x_j \cdot (s_i - x_i)$, which is proportional to the difference between i's promised and exerted effort. We show (Prop. 3) that $(\lambda_c^+, \lambda_c^+)$ is a strict Nash equilibrium of the population game with one-sided reneging costs, it induces the second-best symmetric outcome, it converges to the first-best outcome in the limit when c converges to one, and it induces a better outcome than the sequential-game equilibrium without reneging costs.

Next, we examine a variant of the model in which an agent incurs a fixed reneging cost whenever the exerted effort is different from the promised effort, regardless of the size of the difference. That is, agents care about perfectly keeping their promises. Any reneging on a promise, regardless of the size of the reneging, incurs the same intrinsic cost to the agent. For each $\beta_i, \beta_j \geq 0$ we define *the partnership game with fixed reneging costs* β_i, β_j in the same way as the baseline model except that we change the reneging cost term such that the subjective utility function of each player i is redefined as follows:

$$U_i(x_i, x_j, s_i, c) = x_i \cdot x_j - \frac{c \cdot x_i^2}{2} - \beta_i \cdot 1_{s_i \neq x_i}. \tag{8}$$

We interpret $\beta_i \geq 0$ as the *fixed reneging aversion* of player i (i.e., the intrinsic cost he incurs by reneging on a promise, regardless of the extent of the reneging). We show (Proposition 4) that there exists an intermediate level of fixed reneging aversion, β_c^+, that induces the players to promise and exert maximal effort as part of a trembling-hand perfect equilibrium. This equilibrium induces the first-best outcome.[11]

Finally, we consider more general forms of the utility function. We analyse a general class of games with strategic complements and show that if both agents have either too low or too high reneging costs then essentially no effort is exerted by either player in the game. Effort can only be exerted when reneging costs are intermediate. This demonstrates that the intuition of our baseline model carries over to a more general setup. We formally define this game and formally set out this result in Appendix D.

equilibria and leave this question for future research.

[11] It is difficult to adapt the evolutionary analysis of the baseline model to this setup, because the partnership game with fixed reneging costs may admit multiple trembling-hand perfect equilibria, which makes it difficult to define the payoffs in the induced population game, and to study evolutionary stability of population states.

7 Conclusion

We have demonstrated that an intermediate level of reneging aversion is evolutionarily stable and has a number of appealing properties: it induces a second-best outcome, which converges to the first-best in the limit of small costs of exerting efforts, and it induces a socially better outcome than the Stackelberg-leader setup. Moreover, our main results are robust to relaxations of the model's assumptions to allow for sequential communication, one-sided reneging costs, reneging costs that are discontinuous around zero, and non-quadratic utility.

These results demonstrate a strong tendency of evolution to select preferences for the partial keeping of promises. In stable populations, players make slightly "overoptimistic" promises and, while these are not fully realised, the outcome is welfare-maximising among symmetric equilibria of the game. This 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 cannot 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 finding of several experimental studies (Charness & Dufwenberg 2006; Vanberg 2008; Ederer & Stremitzer 2017; Di Bartolomeo *et al.* 2018) that players are significantly more likely to make "cooperative" choices in a partnership setting when they can communicate before playing, and that players making promises are particularly likely to cooperate.[12] 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. The four aforementioned studies all find that: (1) not all pairs make choices that achieve the cooperative outcome and (2) most players keep promises to play the cooperative or efficient action but some players break their promise. Additionally, Charness & Dufwenberg (2006), who vary the value of the outside option from not engaging in the partnership, find that players are less likely to promise and achieve cooperation when the return from not cooperating is high. The setup in Ederer & Stremitzer (2017) allows agents to choose to "perform" a promised action to varying degrees (where higher performance reduces their own payoff but increases the social payoff) and they find substantial amounts of partial reneging, consistent with our modeling of convex reneging costs.

There is in the experimental literature a debate over whether individuals are inclined to keep their promises because they have an aversion to breaking their promises *per se* or because they are averse to letting down others' payoff expectations (so-called *guilt aversion*).[13] We lay the theoretical foundations for both of these accounts of promise-keeping. In our baseline model, the reneging costs can be interpreted as a cost of promise-breaking *per se*. In the one-sided cost variant, individuals suffer a cost of reneging

[12] The appendix of Charness & Dufwenberg (2006) provides the text of the messages sent by players and demonstrates that they were indeed often used to make explicit promises about their own future action. Ederer & Stremitzer (2017) and Di Bartolomeo *et al.* (2018) classify communication according to whether or not it constituted a promise and show that promises are associated with higher total payoffs relative to general forms of communication.

[13] Ederer & Stremitzer (2017) and Di Bartolomeo *et al.* (2018) further distinguish between agents who care about letting down others' payoff expectations in general (guilt aversion) and agents who care about letting down others' expectations when their promise has caused those expectations to be raised (conditional guilt aversion). In our model, these notions coincide as payoffs are a function solely of players' actions.

only if they exert less effort than promised, and hence cause their partner to have a lower payoff than if they did not renege, and the cost they experience is proportional to the impact on their partner's payoff, such that it can be interpreted as guilt aversion. In both cases, we demonstrate the evolutionary stability and efficiency of intermediate levels of reneging aversion.

Finally, experiments on promises find heterogeneity in preferences for promise-keeping *within* the populations studied. Additionally, Cohn *et al.* (2019) find heterogeneity in preferences for honesty *between* populations in different countries. In our model, the stable level of reneging aversion is a function of the parameters of the game played. If different populations have faced different environments (i.e. different games) and evolution takes place within populations (for example if cultural exchange and imitation mainly takes place between those in the same country) then our model could rationalise heterogeneity in preferences between populations. While our evolutionary results pertain to homogenous populations, the existence of two-message equilibria in partnership games with heterogeneous levels of reneging aversion suggests that the evolution of heterogeneous preferences *within* populations could be a fruitful subject for future research.

This research lends 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 and when allowing for the evolution of the form of the reneging cost function, not only the reneging aversion parameter. Finally, following Alger & Weibull (2013), we conjecture that evolution under positive assortative matching could support the stability of non-cheap talk preferences even when preferences are unobserved.

References

Abeler, Johannes, Nosenzo, Daniele, & Raymond, Collin. 2019. Preferences for Truth-Telling. *Econometrica*, **87**(4), 1115–1153.

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.

Cohn, Alain, Maréchal, Michel André, Tannenbaum, David, & Zünd, Christian Lukas. 2019. Civic honesty around the globe. *Science*, **365**(6448), 70–73.

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.

Di Bartolomeo, Giovanni, Dufwenberg, Martin, Papa, Stefano, & Passarelli, Francesco. 2018. Promises, expectations & causation. *Games and Economic Behavior*.

Dufwenberg, Martin, & Dufwenberg, Martin A. 2018. Lies in disguise–A theoretical analysis of cheating. *Journal of Economic Theory*, **175**, 248–264.

Dufwenberg, Martin, & Güth, Werner. 1999. Indirect Evolution vs. Strategic Delegation: A Comparison of Two Approaches to Explaining Economic Institutions. *European Journal of Political Economy*, **15**(2), 281–295.

Ederer, Florian, & Stremitzer, Alexander. 2017. Promises and expectations. *Games and Economic Behavior*, **106**, 161–178.

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.

Fershtman, Chaim, & Gneezy, Uri. 2001. Strategic delegation: An experiment. *RAND Journal of Economics*, **32**(2), 352–368.

Fershtman, Chaim, Judd, Kenneth L., & Kalai, Ehud. 1991. Observable contracts: Strategic delegation and cooperation. *International Economic Review*, **32**(3), 551–559.

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

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

Gneezy, Uri, Kajackaite, Agne, & Sobel, Joel. 2018. Lying aversion and the size of the lie. *American Economic Review*, **108**(2), 419–53.

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.

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.

Levin, Jonathan. 2003. Supermodular games. Lectures Notes, Department of Economics, Stanford University.

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.

Milgrom, Paul, & Roberts, John. 1990. Rationalizability, learning, and equilibrium in games with strategic complementarities. *Econometrica*, **58**(6), 1255–1277.

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. 2002. On the dynamic foundation of evolutionary stability in continuous models. *Journal of Economic Theory*, **107**(2), 223–252.

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*, **63**(6), 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.

Simon, Leo K., & Stinchcombe, Maxwell B. 1995. Equilibrium refinement for infinite normal-form games. *Econometrica*, **63**(6), 1421–1443.

Taylor, Peter D., & Jonker, Leo 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*. MIT Press, Cambridge, MA.

Appendices (for Online Publication)

A Trembling-Hand Perfection

In this section we formally define the refinement of trembling-hand perfection. This refinement requires that the equilibrium behaviour should be a limit of equilibria of perturbed environments in which the players occasionally make mistakes ("tremble").

As discussed in Section 3.2 the second stage of the game (in which players exert efforts, given the realised promises) admits a unique Nash equilibrium. Thus, we can simplify the notation of our definition of trembling-hand perfection by focusing only on trembles in the first stage, and assuming that players always follow the unique Nash equilibrium of the second stage. Specifically, we study a one-shot game (the *promise game*), in which agents simultaneously choose promises $s_i, s_j \in [0,1]^2$, and the utility of the players $U_i(s_i, s_j, c)$ is determined by assuming that in the second stage the players must follow the unique second-stage Nash equilibrium (as defined in Eq. (6)). Originally, Selten (1975) defined the notion of trembling-hand perfection only for finite games. Because the set of promises in our setup is a continuum, we follow Simon & Stinchcombe's (1995) adaptation of trembling-hand perfection to infinite games (called strong perfect equilibrium in Simon & Stinchcombe, Definition 1.2).

Fix $c \in (1,2)$. Let $\Delta^{fs}([0,1])$ be the set of (Borel) probability measures on $[0,1]$ assigning strictly positive mass to every nonempty open subset of $[0,1]$. Given a strategy $\sigma_j \in \Delta^{fs}([0,1])$, let $BR_i^c(\sigma_{-i})$ be the set of distributions over promises (mixed promises) that are best replies to σ_{-i} (where the players are assumed to follow the unique Nash equilibrium when choosing their effort levels in the second-stage):

$$BR_i^c(\sigma_j) = \left\{ \text{argmax}_{\sigma_i \in [0,1]} \left(U_i(\sigma_i, \sigma_{-i}, c) \equiv \int_{[0,1]^2} (\sigma_i(s_i) \cdot \sigma_j(s_j) \cdot U_i(s_i, s_{-i}, c))\, ds_i ds_j \right) \right\}.$$

An ϵ-perfect equilibrium is a full-support strategy profile in which each player assigns a probability of at least $1-\epsilon$ to best replies to the opponent's strategy. Formally:

Definition 1. An *ϵ-perfect equilibrium* is a pair $(\sigma_1^\epsilon, \sigma_2^\epsilon) \in \left(\Delta^{fs}([0,1])\right)^2$ such that for each $i \in \{1,2\}$,

$$\inf_{\tilde{\sigma}_i \in BR_i^c(\sigma_j^\epsilon)} \sup\left(|\sigma_i^\epsilon(B) - \tilde{\sigma}_i(B)\,|B \text{ measurable}|\right) < \epsilon.$$

A perfect equilibrium is a limit of ϵ-perfect equilibria as ϵ converges to zero. Formally:

Definition 2. A pair of mixed promises $(\sigma_1^*, \sigma_2^*) \in (\Delta([0,1]))^2$ is a *trembling-hand perfect equilibrium* if it is the weak limit as $\epsilon_n \to 0$ of a sequence of *ϵ_n-perfect equilibria.*

Simon & Stinchcombe (1995, Thm. 2.1) show that the set of perfect equilibria is a closed, nonempty subset of the set of Nash equilibria of the promise game. Our definition of the promise game implies that any Nash equilibrium $\left(s_i^*, s_j^*\right)$ of the promise game (and, thus, any trembling-hand perfect equilibrium of the promise game) induces a subgame-perfect equilibrium of the two-stage partnership game $\left(s_i^*, s_j^*, x_i^e, x_j^e\right)$, where $\left(x_i^e(s_i, s_j), x_j^e(s_i, s_j)\right)$ describes the unique second-stage Nash equilibrium in any subgame. The arguments presented in the proof of Lemma 1 below imply that all Nash equilibria (and hence all perfect equilibria) of the promise game are pure strategy profiles.

Finally, we say that a subgame-perfect equilibrium $\left(s_i^*, s_j^*, x_i^*, x_j^*\right)$ of the partnership game is trembling-hand perfect if the pair of promises $\left(s_i^*, s_j^*\right)$ is a trembling-hand perfect equilibrium of the induced one-shot promise game.

Remark 4. We follow the main solution concept introduced by Simon & Stinchcombe (1995), namely, strong perfect equilibrium. Simon & Stinchcombe, at the end of Section 1.1, argue that this notion best captures the strategic structure of infinite games. Their alternative notion, *weak perfect equilibrium*, replaces the strong metrics with the weak metrics in Def. 1. Weak perfection has no bite in our setup: any subgame-perfect equilibrium of the partnership game satisfies weak perfection. Specifically, consider the region Λ_{max}^c in which each player's best reply is overcutting his partner's promise (i.e., $BR_i^c(s_j) = \min(a_i \cdot s_j, 1)$ for some $a_i > 0$). The intuitively unstable Nash equilibrium of the promise game $(0,0)$ satisfies weak perfection: if the partner uses a totally mixed strategy with expectation $\frac{\epsilon}{a_i}$, then the message 0 is ϵ away from the unique best reply message ϵ, which is sufficient for $(0,0)$ to be a weak trembling-hand perfect equilibrium. By contrast, the message 0 is never a best reply to a totally mixed message sent by the partner, which implies that it is not a (strong) trembling-hand perfect equilibrium.

B Further Discussion of Our Evolutionary Model

In this appendix we discuss two issues related to our evolutionary interpretation of the population game: (1) mixed and asymmetric equilibria, and (2) refinements of continuous stability.

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 we allow also for mixed equilibria and asymmetric equilibria).[14] We leave the analysis of this conjecture for future research.

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 level of reneging aversion. Eshel (1983) argues that in some setups one should also require stability against perturbations in which many (or all) agents slightly change their reneging aversion. Eshel presents the notion of a *continuous stable strategy* to capture stability also against the latter class of perturbations, and Oechssler & Riedel (2002) 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

[14]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 in Ritzberger & Weibull, 1995).

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.,

$$\sum_{\lambda \in \Delta(\mu)} \mu(\lambda) \cdot \pi(\lambda^*, \lambda) > \sum_{\lambda, \lambda' \in \Delta(\mu)} \mu(\lambda) \cdot \mu(\lambda') \cdot \pi(\lambda, \lambda'). \tag{9}$$

One can show that the population state $(\lambda_c^+, \lambda_c^+)$ satisfies a somewhat weaker version of the evolutionary robustness refinement of (9). Specifically, it satisfies the weak inequality counterpart of Eq. (9) 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^+ .

C Additional Figures

The appendix presents additional figures demonstrating how the value of λ_c^+, level of effort, material payoff, and subjective utility change as a function of the cost of effort c.

Figure 3 shows how the value of $\lambda_c^+ = \frac{1+2c-2c^2}{2(c-1)} + \frac{\sqrt{5-4c}}{2(c-1)}$ depends on the cost of effort c.

Figure 3: The Intermediate Reneging Aversion λ_c^+ as a Function of the Cost of Effort c

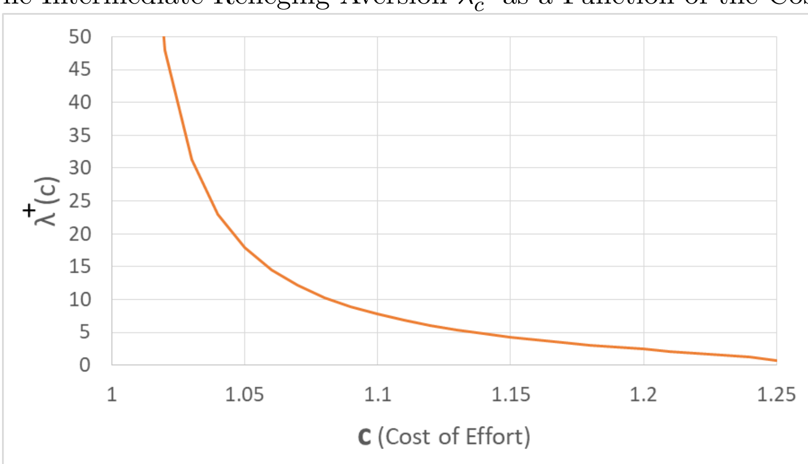
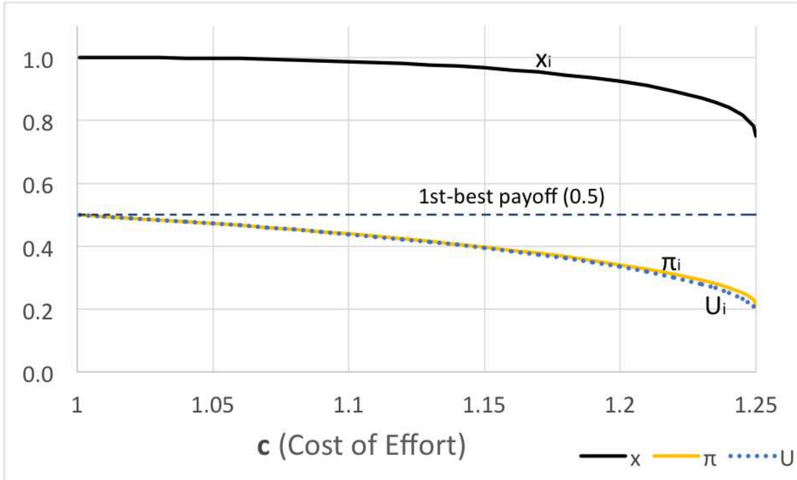


Figure 4 shows the level of effort, x_i, the material payoff of each player, π_i, and the subjective utility of each player, U_i, as a function of the cost of effort, c, in the unique equilibrium induced by the partnership game in which both players have the intermediate level of reneging aversion λ_c^+.

Figure 4: Equilibrium Effort, Material Payoff and Subjective Utility as a Function of the Cost of Effort



D Formal Statement of Results of Variants and Extensions

Result for sequential communication The choices and information sets of players is as described in Section 6. Material payoff and utility functions and all other aspects of the game remain the same as in the baseline model. We now state the formal result.

Proposition 2. *Fix $c \in (1, 1.25)$. Let $(\lambda_c^+, \lambda_c^+)$ be the highest symmetric pair of levels of reneging aversion inducing a maximum-message equilibrium in the simultaneous communication game. The unique subgame-perfect equilibrium induced by $(\lambda_c^+, \lambda_c^+)$ under sequential communication satisfies:*

1. *The agents promise maximal efforts, and exert the same level of effort as in the baseline model.*
2. *"Second-best" outcome: If $c < 1.18$ then $\pi_c(\lambda_c^+, \lambda_c^+) > \pi_c(\lambda', \lambda')$ and $U_c(\lambda_c^+, \lambda_c^+) > U_c(\lambda', \lambda')$ for any $\lambda' \neq \lambda_c^+$.*
3. *Convergence to "first-best" outcome: $\lim_{c \to 1} \pi_c(\lambda_c^+, \lambda_c^+) = \lim_{c \to 1} U_c(\lambda_c^+, \lambda_c^+) = \frac{1}{2}$.*
4. *Strict equilibrium of population game: $\pi_c(\lambda_c^+, \lambda_c^+) > \pi_c(\lambda', \lambda_c^+)$ for each $\lambda' \neq \lambda_c^+$.*
5. *Better outcome than that of the sequential-game equilibrium without reneging costs: Let π_i^s be the payoff to player i in the unique equilibrium of the game where efforts are chosen sequentially (and there are no reneging costs). Then, if $c < 1.22$ then $\pi_c(\lambda_c^+, \lambda_c^+) > \frac{1}{2} \cdot (\pi_i^s + \pi_j^s)$.*

Result for one-sided reneging costs The utility function is as defined in Eq. (7) in Section 6. All other aspects of the partnership game remain the same as in the baseline model. Before stating the result, we make two assumptions regarding the equilibrium selection function in cases in which the partnership game admits multiple equilibria:

1. It turns out that the set of equilibria in the symmetric partnership game $(\lambda_c^+, \lambda_c^+)$ with one-sided reneging costs coincides with the set of equilibria in the baseline model with two-sided reneging costs, and, thus, we apply in this case the same equilibrium selection function as in the baseline model.

2. Unlike in the baseline model, with one-sided reneging costs, some symmetric partnership games (λ,λ) (with $\lambda \neq \lambda_c^+$) have multiple equilibria. We allow in this case an arbitrary equilibrium selection function. If an asymmetric equilibrium is selected (in which one of the agents is assigned to the role of player one, while the partner is assigned to the role of player two), we define $\pi_c(\lambda,\lambda)$ as the mean payoff of the two players' roles. This corresponds to a homogeneous population of agents with reneging aversion λ, in which each agent has equal probability of being assigned to each role in the selected asymmetric equilibrium.

Proposition 3. *Fix $c \in (1, 1.25)$. Let $(\lambda_c^+, \lambda_c^+)$ be the highest symmetric pair of levels of reneging aversion inducing a maximum-message equilibrium in the partnership game with two-sided reneging costs (as defined in Theorem 1). The equilibrium induced by $(\lambda_c^+, \lambda_c^+)$ with one-sided reneging costs has the following properties:*

1. *The agents promise maximal efforts, and exert the same level of effort as in the baseline model.*
2. *"Second-best" outcome: If $c < 1.22$ then $\pi_c(\lambda_c^+, \lambda_c^+) > \pi_c\left(\lambda', \lambda'\right)$ for any $\lambda' \neq \lambda_c^+$.*
3. *Convergence to "first-best" outcome: $\lim_{c\to 1} \pi_c(\lambda_c^+, \lambda_c^+) = \lim_{c\to 1} U_c(\lambda_c^+, \lambda_c^+) = \frac{1}{2}$ (which is the best feasible material payoff).*
4. *Strict equilibrium of the population game: $\pi_c(\lambda_c^+, \lambda_c^+) > \pi_c(\lambda', \lambda_c^+)$ for each $\lambda' \neq \lambda_c^+$.*
5. *Better outcome than that of the sequential-game equilibrium without reneging costs: Let π_i^s be the payoff to player i in the unique equilibrium of the game where efforts are chosen sequentially (and there are no reneging costs). Then, if $c < 1.22$ then $\pi_c(\lambda_c^+, \lambda_c^+) > \frac{1}{2} \cdot \left(\pi_i^s + \pi_j^s\right)$.*

Result for fixed reneging costs The utility function is as defined in Eq. (8) in Sec. 6. All other aspects of the game remain the same as in the baseline model. We now state the formal result.

Proposition 4. *For any $c \in (1, 2)$, there exists an intermediate level of reneging aversion β_c^+, for which there exists a trembling-hand perfect equilibrium of the partnership game with fixed reneging costs $\beta_i = \beta_j = \beta_c^+$, in which both agents promise and exert the maximal effort.*

Result for general utility functions Finally, we generalize the model to deal with general games with strategic complements and show that if both agents have either too low or too high reneging costs then essentially no effort is exerted by either player in the game.

Let $\pi(x_i, x_j)$ be the material payoff of an agent exerting effort $x_i \in [0,1]$, given that his partner exerts effort $x_j \in [0,1]$. Throughout this extension we assume that the material payoff function, π, is twice continuously differentiable, has positive externalities, i.e., $\pi(x_i, x_j)$ is increasing in its second argument ($\frac{\partial \pi(x_i,x_j)}{\partial x_j} > 0$), and has strategic complements, i.e., $\frac{\partial^2 \pi_i(x_i,x_j)}{\partial x_i \cdot \partial x_j} > 0$ for each $x_i, x_j \in [0,1]$. Recall (see, e.g., Milgrom & Roberts, 1990; Levin, 2003) that a game with strategic complements admits a highest pure Nash equilibrium, which we denote by $(\bar{x}, \bar{x})$ (i.e., $x, x' \leq \bar{x}$ for each Nash equilibrium (x, x')).

We assume that each player i is endowed with a level of reneging aversion λ_i. The players' levels of reneging aversion are common knowledge. The subjective utility of each player i is the sum of the material payoff and a term representing the psychological cost of breaking a promise (reneging). Formally:

$$U_i(x_i, x_j, s_i, \lambda_i) = \pi(x_i, x_j) - \lambda_i \cdot D(|s_i - x_i|).$$

Hence, 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 function $D : [0,1] \to R^+$ determines the shape of the reneging cost function. We assume that this function is weakly increasing (i.e., $x \geq y$ implies $D(x) \geq D(y)$), and that $D(x) > D(0)$ for each $x > 0$. That is, any difference between the promise and the exerted effort induces a positive intrinsic cost. To simplify notation, we normalise D such that $D(0) = 0$.

Our final result shows that agents do not exert effort above $\bar{x}$ in any pure subgame-perfect equilibrium of the partnership game whenever the reneging costs are either too low or too high.

Proposition 5. *For any $\epsilon > 0$, there exist $\overline{\lambda}_\epsilon > \underline{\lambda}_\epsilon > 0$, such that the effort level exerted by any agent in any pure subgame-perfect equilibrium is at most $\bar{x} + \epsilon$ if either (1) $\lambda_i, \lambda_j < \underline{\lambda}_\epsilon$ or (2) $\lambda_i, \lambda_j > \overline{\lambda}_\epsilon$.*

E Proofs

E.1 Definitions of Notation Used in the Proofs

For ease of exposition, we define the following notation, used in several of the subsequent proofs:

$$\Theta_i \equiv c(c+\lambda_j) + \frac{1}{(c+\lambda_i)(c+\lambda_j)} - 2, \qquad R_i \equiv \begin{cases} \dfrac{\lambda_j}{\Theta_i} & \Theta_i > 0 \\ \infty & \Theta_i \leq 0. \end{cases}$$

Throughout the proofs we define the product of ∞ and 0 to be equal to ∞. That is, when $R_i = \infty$ and $R_j = 0$, we define $R_i \cdot R_j = \infty$.

E.2 Proof of Proposition 1

This section consists of several lemmas used in the proof of Proposition 1, followed by the proof itself.

E.2.1 Lemma Characterising the Best-Reply Correspondence

Lemma 1. *Let μ_{σ_j} denote i's expectation of s_j in the first stage of the partnership game when player j chooses a mixed strategy $\sigma_j \in \Delta([0,1])$ in the first stage (i.e., a distribution over the set of messages). The best-reply correspondence in the first stage is*[15]

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

[15] 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, thus, both players choose pure messages in all equilibria. This justifies the focus on pure strategies in the main text.

Proof. To derive player i's first stage best reply, we substitute the equations for equilibrium second-stage effort levels (Eq. (5)) into the utility function to obtain utility as a function of s_i and s_j:

$$\begin{aligned} U_i(s_i, s_j, c) =& \frac{[(c+\lambda_j)\lambda_i s_i + \lambda_j s_j][(c+\lambda_i)\lambda_j s_j + \lambda_i s_i]}{[(c+\lambda_i)(c+\lambda_j)-1]^2} \\ & - \frac{c[(c+\lambda_j)\lambda_i s_i + \lambda_j s_j]^2}{2[(c+\lambda_i)(c+\lambda_j)-1]^2} - \frac{\lambda_i}{2}\left[s_i - \frac{(c+\lambda_j)\lambda_i s_i + \lambda_j s_j}{(c+\lambda_i)(c+\lambda_j)-1}\right]^2 \end{aligned} \tag{11}$$

When $\lambda_i = 0$, player i's choice of message has no bearing on his optimal effort choice or that of his partner and thus does not impact his utility. Therefore, any s_i is a best reply to any μ_{σ_j} (and indeed any s_j). 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}:

$$\frac{\partial U_i(s_i, \mu_{\sigma_j}, c)}{\partial s_i} = \frac{\lambda_i(c+\lambda_i)(c+\lambda_j)}{[(c+\lambda_i)(c+\lambda_j)-1]^2}\left(\left[2 - c(c+\lambda_j) - \frac{1}{(c+\lambda_i)(c+\lambda_j)}\right]s_i + \lambda_j \cdot \mu_{\sigma_j}\right) \tag{12}$$

$$\equiv \Upsilon_i \cdot \left(-\Theta_i s_i + \lambda_j \cdot \mu_{\sigma_j}\right) \tag{13}$$

When $\lambda_i > 0$, we have $\Upsilon_i > 0$. Given that λ_j and μ_{σ_j} are constrained to be (weakly) positive, the second term inside the brackets in Eq. (13) is also (weakly) positive. Therefore, when $\Theta_i > 0$ (and hence the term multiplying s_i in Eq. (13) 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:

$$s_i(\mu_{\sigma_j}, \lambda_i, \lambda_j, c) = \frac{\lambda_j}{\Theta_i} \cdot \mu_{\sigma_j} \tag{14}$$

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. (13) 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 reply to the opponent's message. □

E.2.2 Conditions for the Existence of Each Best-Reply "Type"

Lemma 2. $\Theta_i \leq 0$ *(which implies that player i's best reply is to send the maximum message) 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. 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:

$$\lambda_i \leq \frac{-1}{(c+\lambda_j)[c(c+\lambda_j)-2]} - c < 0, \text{ and } c(c+\lambda_j)-2 > 0, \tag{15}$$

$$\lambda_i \geq \frac{1}{(c+\lambda_j)[2-c(c+\lambda_j)]} - c > 0, \text{ and } c(c+\lambda_j)-2 < 0. \tag{16}$$

We can see that the solution given by Eq. (15) implies that $\lambda_i < 0$, which is ruled out by assumption. Therefore, we have that $\Theta_i \leq 0 \iff$ Eq. (16) holds. Rearranging the second inequality in Eq. (16) to give a condition in terms of λ_j yields the lemma. □

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

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

$$AND \left(\left(\lambda_i < \frac{1}{(c+\lambda_j)(2-c(c+\lambda_j))} - c\right) \text{ or } \left(\frac{2}{c} - c \leq \lambda_j < \frac{2-c^2}{c-1}\right)\right).$$

Proof. By the definition of Θ_i,

$$\frac{\lambda_j}{\Theta_i} > 1 \iff \frac{\lambda_j}{c(c+\lambda_{-i}) + \frac{1}{(c+\lambda_i)(c+\lambda_j)} - 2} > 1. \tag{17}$$

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

$$\lambda_j > c(c+\lambda_j) + \frac{1}{(c+\lambda_i)(c+\lambda_j)} - 2 > 0. \tag{18}$$

The second of these inequalities is the requirement that $\Theta_i > 0$, which is the converse of the condition derived for Lemma 2, and this second inequality holds when

$$\lambda_i < \frac{1}{(c+\lambda_j)[2-c(c+\lambda_j)]} - c \quad or \quad \lambda_j \geq \frac{2}{c} - c. \tag{19}$$

The first inequality in Eq. (18) holds if and only if

$$\begin{aligned}
\lambda_j >& c(c+\lambda_j) + \frac{1}{(c+\lambda_i)(c+\lambda_j)} - 2 \\
\Longleftrightarrow & \lambda_j + 2 - c(c+\lambda_j) > \frac{1}{(c+\lambda_i)(c+\lambda_j)} \iff (c+\lambda_i)\left(\lambda_j + 2 - c(c+\lambda_j)\right) > \frac{1}{(c+\lambda_j)} \\
\Longleftrightarrow & \lambda_i(\lambda_j + 2 - c(c+\lambda_j)) > -c(\lambda_j + 2 - c(c+\lambda_j)) + \frac{1}{c+\lambda_j}.
\end{aligned} \tag{20}$$

The second $\Longleftrightarrow$ is obtained by multiplying by $(c+\lambda_i)$, and the first and third by 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:

$$\lambda_i > \frac{1}{(\lambda_j + 2 - c(c+\lambda_j))\,(c+\lambda_j)} - c \;\; > 0 \text{ and } \lambda_j + 2 - c(c+\lambda_j) > 0, \tag{21}$$

$$\lambda_i < \frac{1}{(\lambda_j + 2 - c(c+\lambda_j))\,(c+\lambda_j)} - c \;\; < 0 \text{ and } \lambda_j + 2 - c(c+\lambda_j) < 0. \tag{22}$$

We can see that the solution given by Eq. (22) implies that $\lambda_i < 0$. This is ruled out by assumption, and so we have that the first inequality in Eq. (18) $\iff$ Eq. (21) holds. Rearranging the inequalities in Eq. (21) yields

$$\lambda_i > \frac{1}{\lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)} - c \;\; > 0 \text{ and } \lambda_j < \frac{2-c^2}{c-1}. \tag{23}$$

Combining the inequalities in Eqs. (23) and (19) and observing that $\frac{2}{c} - c = \frac{2-c^2}{c} < \frac{2-c^2}{c-1}$, and that therefore $0 < \lambda_i < \dfrac{1}{(c+\lambda_j)(2-c(c+\lambda_j))} - c \Rightarrow \lambda_j < \frac{2-c^2}{c-1}$, yields the lemma. ☐

Lemma 4. $0 < \frac{\lambda_j}{\Theta_i} < 1$ *(which implies that player i sends a message that is some fraction (less than 1) of player j's message) if and only if*

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

Proof. The inequality $0 < \frac{\lambda_j}{\Theta_i} < 1$ implies that $\Theta_i > 0$ and so Eq. (19) must hold. We also must have that $\frac{\lambda_j}{\Theta_i} < 1$. In the proof of Lemma 3 it was demonstrated that $\frac{\lambda_j}{\Theta_i} > 1 \iff$ Eq. (23) holds. By taking the converse of Eq. (23) we have that $\frac{\lambda_j}{\Theta_i} < 1$ if and only if

$$\lambda_i < \frac{1}{\lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)} - c \text{ or } \lambda_j \geq \frac{2-c^2}{c-1}. \tag{24}$$

From the proof of Lemma 3, we have that $\Theta_i > 0$ if and only if

$$\lambda_i < \frac{1}{(c+\lambda_j)[2-c(c+\lambda_j)]} - c \quad \textit{or} \quad \lambda_j \geq \frac{2}{c} - c. \tag{25}$$

To see that Eq. (24) implies that $\Theta_i > 0$, first note that as $\frac{2}{c} - c = \frac{2-c^2}{c} < \frac{2-c^2}{c-1}$, the second inequality in Eq. (24) implies the second inequality in Eq. (25). Next, we see that if $\lambda_j \geq \frac{2}{c} - c$ then we clearly have the second inequality in Eq. (25). If instead $\lambda_j < \frac{2}{c} - c$ then, given $\frac{2}{c} - c = \frac{2-c^2}{c} < \frac{2-c^2}{c-1}$, Eq. (24) implies that the first inequality in Eq. (24) holds, which in turn implies the first inequality in Eq. (25):

$$\frac{1}{\lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)} - c < \frac{1}{(c+\lambda_j)[2-c(c+\lambda_j)]} - c$$
$$\iff \frac{1}{\lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)} < \frac{1}{(c+\lambda_j)[2-c(c+\lambda_j)]}$$
$$\iff (c+\lambda_j)[2-c(c+\lambda_j)] < \lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)$$
$$\iff 2c - c^2 - \lambda_j c^2 + 2\lambda_j - \lambda_j c^2 - \lambda_j^2 c < \lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)$$
$$\iff 0 < \lambda_j^2 + \lambda_j^2 c.$$

Therefore, $\Theta_i > 0$ is implied by $\frac{\lambda_j}{\Theta_i} < 1$ and so we obtain the lemma. □

E.2.3 Proof of Proposition 1

Proof. Let

$$g(\lambda_j, c) = \frac{1}{\lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)} - c.$$

We prove each point in turn:

1. Overcutting: Assume that $\lambda_j < \frac{2-c^2}{c-1}$ and $\lambda_i > g(\lambda_j, c)$. Lemma 2 and Lemma 3 imply that either $\Theta_i \leq 0$ or $\frac{\lambda_j}{\Theta_i} > 1$. In both of these cases, Lemma 1 implies that $s_i^*(s_j|\lambda_i, \lambda_j, c) > s_j$, given that $0 < s_j < 1$, and this proves part 1.

2. Undercutting: Lemma 4 shows that if $\lambda_j \geq \frac{2-c^2}{c-1}$ or $\lambda_i \in (0, g(\lambda_j, c))$, then $0 < \frac{\lambda_j}{\Theta_i} < 1$. Lemma 1 implies that if $\lambda_i > 0$ and $0 < \frac{\lambda_j}{\Theta_i} < 1$ and $s_j > 0$, then $s_i^*(s_j|\lambda_i, \lambda_j, c) < s_j$ for each $s_j > 0$ and $\lambda_i > 0$, which proves part 2.

□

E.3 Proof of Theorem 1

Proof. Observe that each partnership game is identified by a pair (λ_i, λ_j) and, by the definition of R_i (in Appendix E.1), each partnership game (and each pair (λ_i, λ_j)) corresponds to a unique pair (R_i, R_j). Let

$$\Lambda_{0ef}^c \equiv \left\{(\lambda_i, \lambda_j) \subseteq (0, \infty)^2 \colon R_i \cdot R_j < 1\right\},$$

$$\Lambda_{max}^c \equiv \left\{(\lambda_i, \lambda_j) \subseteq (0, \infty)^2 \colon \min(R_i, R_j) > 1\right\}, \text{ and}$$

$$\Lambda_{2ms}^c \equiv \left\{(\lambda_i, \lambda_j) \subseteq (0, \infty)^2 \colon \min(R_i, R_j) < 1 < R_i \cdot R_j\right\}.$$

Recall that when $R_i = \infty$ and $R_j = 0$, we define $R_i \cdot R_j$ to be equal to ∞. Note that these sets are disjoint and symmetric, and that

$$Cl\left(\Lambda^c_{0-eff}\right) \cup Cl\left(\Lambda^c_{max}\right) \cup Cl\left(\Lambda^c_{2-msg}\right) = [0,\infty)^2.$$

We now prove each point of the theorem in turn.

1. Region of maximum-message equilibrium Λ^c_{max}:

   (a) Let $(\lambda_i, \lambda_j) \in \Lambda^c_{max}$ and let $\left(s^*_i, s^*_j, x^*_i, x^*_j\right) \in PE\left(\lambda_i, \lambda_j, c\right)$ be a trembling-hand perfect equilibrium. We have to prove that $s^*_i = s^*_j = 1$. Observe that $(\lambda_i, \lambda_j) \in \Lambda^c_{max} \Longrightarrow \min\left(R_i, R_j\right) > 1$. Then, by the definition of R_i and R_j, either (i) $\Theta_i, \Theta_j > 0$, and $\frac{\lambda_j}{\Theta_i}, \frac{\lambda_i}{\Theta_j} > 1$, or (ii) $\Theta_i > 0 = \Theta_j$ and $\frac{\lambda_j}{\Theta_i} > 1$, or (iii) $\Theta_i > 0 > \Theta_j$ and $\frac{\lambda_j}{\Theta_i} > 1$, or (iv) $\Theta_i = \Theta_j = 0$, or (v) $\Theta_i = 0 > \Theta_j$, or (vi) $\Theta_i, \Theta_j < 0$. In case (i), by the best-reply correspondence derived in Lemma 1, 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 (ii), Lemma 1 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 (iii), Lemma 1 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 (iv), Lemma 1 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 (v), Lemma 1 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 (vi), Lemma 1 implies that equilibrium messages satisfy $s^*_i = 1$ and $s^*_j = 1$.

   This implies that in all six cases (i, ii, iii, iv, v, and vi) the strategy profile $(1, 1, x^e_1, x^e_2)$ is a subgame-perfect equilibrium. It is unique (and thus satisfies trembling-hand perfection) in cases (ii), (iii), (v), and (vi). In cases (i) and (iv), the strategy profile $(0, 0, x^e_1, x^e_2)$ is the only additional subgame-perfect equilibrium. Finally, we have to show that the additional equilibrium $(0, 0, x^e_1, x^e_2)$ fails to satisfy trembling-hand perfection in cases (i) and (iv).

   Assume to the contrary that $(0, 0, x^e_1, x^e_2)$ satisfies trembling-hand perfection. This implies that $(0, 0)$ is the weak limit as $\epsilon_n \to 0$ of a sequence of ϵ_n-perfect equilibria $\left(\sigma^n_i, \sigma^n_j\right)$ of the promise game (defined in Appendix A). This implies, in particular, that for each $\epsilon > 0$, there exists an ϵ-perfect equilibrium $(\sigma_i, \sigma_j) \in \Delta^{fs}\left([0,1]\right)^2$ such that $\sigma_i\left(1\right), \sigma_j\left(1\right) < \epsilon$. We begin by considering case (iv). The fact that σ_j has full support implies that $\mu_{\sigma_j} > 0$ and that $BR^c_i\left(\sigma_j\right) = \{1\}$. The definition of an ϵ-perfect equilibrium implies that $\sigma_i\left(1\right), \sigma_j\left(1\right) > 1 - \epsilon$, and we get a contradiction for each $\epsilon < 0.5$. We are left with case (i), in which $\Theta_i, \Theta_j > 0$, and $\frac{\lambda_j}{\Theta_i}, \frac{\lambda_i}{\Theta_j} > 1$, in which, in particular, $R_i \cdot R_j > 1$. The fact that $(0, 0)$ is the weak limit of a sequence of ϵ_n-perfect equilibria $\left(\sigma^n_i, \sigma^n_j\right)$ when $\epsilon_n \to 0$ implies that for each $\epsilon > 0$, there exists an ϵ-perfect equilibrium $(\sigma_i, \sigma_j) \in \left(\Delta^{fs}\left([0,1]\right)\right)^2$ such that $\sigma_i\left(\left[\frac{1}{R_i}, 1\right]\right), \sigma_j\left(\left[\frac{1}{R_i}, 1\right]\right) < \epsilon$.

The fact that σ_j has full support implies that $\mu_{\sigma_j} > 0$. Observe that $BR_i^c(\sigma_j) = \left\{R_i \cdot \mu_{\sigma_j}\right\}$. The fact that (σ_i, σ_j) is an ϵ-perfect equilibrium implies that $\sigma_i\left(R_i \cdot \mu_{\sigma_j}\right) \geq 1 - \epsilon$, which implies that $\mu_{\sigma_i} \geq (1-\epsilon) \cdot R_i \cdot \mu_{\sigma_j}$. The fact that (σ_i, σ_j) is an ϵ-perfect equilibrium implies that $\sigma_j(R_j \cdot \mu_{\sigma_i}) \geq 1 - \epsilon$, which implies that

$$\mu_{\sigma_j} \geq (1-\epsilon) \cdot R_j \cdot \mu_{\sigma_i} \geq (1-\epsilon)^2 \cdot R_i \cdot R_j \cdot \mu_{\sigma_j},$$

which yields the contradiction $\mu_{\sigma_j} > \mu_{\sigma_j}$ for a sufficiently small ϵ that satisfies $(1-\epsilon)^2 \cdot R_i \cdot R_j > 1$.

(b) Next we prove that Λ_{max}^c is a convex set, which is nonempty iff $c \in (1, 1.25)$. By the definition of R_i, we recall that $R_i > 1$ if and only if (1) $\Theta_i \leq 0$ or (2) $\Theta_i > 0$ and $\frac{\lambda_j}{\Theta_i} > 1$. We can recall from Lemma 2 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 3 that $\Theta_i > 0$ and $\frac{\lambda_j}{\Theta_i} > 1$ if and only if

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

$$AND \ \left(\left(\lambda_i < \frac{1}{(c+\lambda_j)(2-c(c+\lambda_j))} - c\right) \ or \ \left(\frac{2}{c} - c \leq \lambda_j < \frac{2-c^2}{c-1}\right)\right).$$

Combining these conditions yields $R_i > 1$ if and only if

$$\lambda_i > \frac{1}{\lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)} - c \quad \text{and} \quad \lambda_j < \frac{2-c^2}{c-1}. \tag{26}$$

We will now show that the set of points that satisfy Eq. (26) is convex. First, observe that the second derivative of the right-hand side of the first inequality of Eq. (26) (the lower bound on λ_i) with respect to λ_j is

$$\frac{2[3c^4 + (6\lambda_j - 3)c^3 + (3\lambda_j^2 - 9\lambda_j - 5)c^2 + 3\lambda_j^2 + 6\lambda_j + 4]}{(\lambda_j + c)[2 - c^2 - \lambda_j(c-1)]}. \tag{27}$$

The numerator of this expression is positive for all $\lambda_j > 0$ and[16] $c > 1$. 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. (26). Therefore, the set of points that satisfy Eq. (26)

[16] Eq. (27) and the conditions for the positive numerator are derived using Mathematica. The code is available in the supplementary appendix of this paper.

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 $\min(R_i, R_j) > 1$ (Λ^c_{max}) is convex.

Next, we show that Λ^c_{max} is nonempty iff $c \in (1, 1.25)$. By the convexity and symmetry of Λ^c_{max}, if this set is nonempty there must be a maximum and a minimum λ such that $(\lambda, \lambda) \in Cl(\Lambda^c_{max})$. We now show that such maximum and minimum elements exist if and only if $c < 1.25$. Clearly, the maximum and minimum λ such that $(\lambda, \lambda) \in Cl(\Lambda^c_{max})$ are the largest and smallest values of λ such that the weak counterpart of Eq. (26) holds when $\lambda_i = \lambda_j = \lambda$. Given that $Cl(\Lambda^c_{max})$ is convex and closed, these maximum and minimum values must obtain when at least one of the inequalities in Eq. (26) holds with equality. To find the maximum and minimum values of λ that satisfy the first inequality in Eq. (26), we solve the corresponding equation 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. (26), we obtain

$$\lambda = \frac{1}{\lambda^2(1-c) + \lambda(2-2c^2+c) + c(2-c^2)} - c. \tag{28}$$

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

$$\lambda^3\Big[1-c\Big] + \lambda^2\Big[2+2c-3c^2\Big] + \lambda\Big[4c-3c^3+c^2\Big] - \Big[c^2-1\Big]^2 = 0. \tag{29}$$

Eq. (29) has two solutions when λ is positive:

$$\lambda = \frac{1+2c-2c^2}{2(c-1)} - \frac{\sqrt{5-4c}}{2(c-1)} \equiv \lambda_c^- \tag{30}$$

$$\lambda = \frac{1+2c-2c^2}{2(c-1)} + \frac{\sqrt{5-4c}}{2(c-1)} \equiv \lambda_c^+ \tag{31}$$

Clearly, these two solutions are defined if and only if $c < 1.25$ (and, thus, $\Lambda^c_{max} = \emptyset$ when $c \geq 1.25$). By inspection of Eq. (30) and Eq. (31), it is straightforward to see that for all $1 < c < 1.25$, $0 < \lambda_c^- < \lambda_c^+ < \infty$ (and, in particular, that $\Lambda^c_{max} \neq \emptyset$ when $c < 1.25$).

(c) For each $c \in (1, 1.25)$, there exists $(\lambda_c^+, \lambda_c^+) \in Cl(\Lambda^c_{max})$ such that $(\lambda, \lambda') \in \Lambda^c_{max} \Rightarrow \lambda, \lambda' < \lambda_c^+$. Let λ_c^- and λ_c^+ be defined as in Eq. (30) and Eq. (31) above. It is immediate from the definition of λ_c^+ in (30) that $(\lambda_c^+, \lambda_c^+) \in Cl(\Lambda^c_{max})$. The definition of λ_c^- and λ_c^+ as the minimum and maximum λ (respectively) such that $(\lambda, \lambda) \in Cl(\Lambda^c_{max})$ and the convexity of Λ^c_{max}, further imply that $(\lambda, \lambda) \notin Cl(\Lambda^c_{max})$ for each $\lambda \in [0, \lambda_c^-) \cup (\lambda_c^+, \infty)$. Assume that there exist $\lambda_i, \lambda_j > \lambda_c^+$ such that $(\lambda_i, \lambda_j) \in \Lambda^c_{max}$. By the symmetry of Λ^c_{max}, we have that $(\lambda_j, \lambda_i) \in \Lambda^c_{max}$. Let $\lambda_k = \frac{\lambda_i + \lambda_j}{2} > \lambda_c^+$. By the convexity of Λ^c_{max}, we have that $(\lambda_k, \lambda_k) \in \Lambda^c_{max}$, which is a contradiction. Finally, we consider the case where $\lambda_i \leq \lambda_c^+ \leq \lambda_j$. We have established in the proof of the previous part that the right-hand side of the first inequality in Eq. (26) (which gives the condition for $R_i > 1$) is strictly convex and crosses the 45 degree line for the second time at $\lambda_i = \lambda_j = \lambda_c^+$, which implies that for all $\lambda_j \geq \lambda_c^+$ this function is increasing in λ_j and so $R_i > 1 \implies \lambda_i > \lambda_c^+$, which is a contradiction. We therefore have that $\lambda_c^+ \leq max\,(\lambda_i, \lambda_j)$ implies that $(\lambda_i, \lambda_j) \notin \Lambda_{max}^c$ and hence $(\lambda_i, \lambda_j) \in \Lambda_{max}^c$ implies that $max\,(\lambda_i, \lambda_j) < \lambda^+$.

2. Region of two-message equilibrium Λ_{2ms}^c:

    (a) Let $(\lambda_i, \lambda_j) \in \Lambda_{2ms}^c$ and $\left(s_i^*, s_j^*, x_i^*, x_j^*\right) \in PE\,(\lambda_i, \lambda_j, c)$. Assume that $0 < \lambda_j < \lambda_i$. We have to show that $s_j^* < s_i^* = 1$. We prove this claim in the following three steps:

        i. We first show that $R_j < \infty$. Assume to the contrary that $R_j = \infty$. This implies that $\Theta_j \leq 0$. The fact that $\lambda_j < \lambda_i$ implies that $\Theta_i \leq \Theta_j \leq 0$, which, in turn, implies that $R_i = \infty$, and we get a contradiction to $\min\,(R_i, R_j) < 1$.

        ii. Next, we show that $R_j < 1 < R_i$. Assume to the contrary that $R_i < 1 < R_j < \infty$. Observe that

$$R_i - R_j = \frac{\lambda_j}{\Theta_i} - \frac{\lambda_i}{\Theta_j} = \frac{\lambda_j \Theta_j - \lambda_i \Theta_i}{\Theta_i \cdot \Theta_j},$$

which implies that $R_i < R_j$ iff

$$0 > \lambda_j \Theta_j - \lambda_i \Theta_i = \lambda_j c(c+\lambda_i) + \frac{\lambda_j}{(c+\lambda_i)(c+\lambda_j)} - 2\lambda_j - \lambda_i c(c+\lambda_j) - \frac{\lambda_i}{(c+\lambda_i)(c+\lambda_j)} + 2\lambda_i$$

$$= c\,(\lambda_j(c+\lambda_i) - \lambda_i(c+\lambda_j)) - \frac{\lambda_i - \lambda_j}{(c+\lambda_i)(c+\lambda_j)} + 2\,(\lambda_i - \lambda_j)$$

$$= c\,((\lambda_j - \lambda_i)\,c) - \frac{\lambda_i - \lambda_j}{(c+\lambda_i)(c+\lambda_j)} + 2\,(\lambda_i - \lambda_j) = (\lambda_i - \lambda_j)\left(2 - c^2 - \frac{1}{(c+\lambda_i)(c+\lambda_j)}\right)$$

$$= (\lambda_i - \lambda_j)\,(c\lambda_i - \Theta_j) >= (\lambda_i - \lambda_j)\,(\lambda_i - \Theta_j) > 0,$$

and we get a contradiction (where the last inequality is due to $1 < R_j < \infty \Rightarrow \lambda_i > \Theta_j$).

        iii. The previous step and the definition of Λ_{2ms}^c imply that $R_j < 1 < R_i \cdot R_j < R_i$. By the definition of R_i and R_j, either (I) $\Theta_i < 0$, $\Theta_j > 0$, and $0 < \frac{\lambda_i}{\Theta_j} < 1$, or (II) $\Theta_i = 0$, $\Theta_j > 0$, $\lambda_j > 0$, and $\frac{\lambda_i}{\Theta_j} < 1$, or (III) $\Theta_i > 0$, $\Theta_j > 0$, and $\frac{\lambda_j}{\Theta_i} \cdot \frac{\lambda_i}{\Theta_j} > 1$. In case (I) Lemma 1 implies that equilibrium messages satisfy $s_i^* = 1$. Then by Lemma 1 $s_j^* = \frac{\lambda_i}{\Theta_j} s_i$, and these equations are simultaneously satisfied if and only if $1 = s_i^* > s_j^* > 0$. In case (II), Lemma 1 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 (III), Lemma 1 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 (I, II, and III), there exists a subgame-perfect equilibrium in which $1 = s_i^* > s_j^* > 0$. This is the unique subgame-perfect equilibrium in case (I) and therefore it must satisfy trembling-hand perfection.

In cases (II) and (III) there exists also a subgame-perfect equilibrium in which $s_i^* = s_j^* = 0$. The proof that this subgame-perfect equilibrium is not trembling-hand perfect is essentially the same as in the analogous proof in the end of part (1-a) above (for the region Λ^c_{max}) and is omitted for brevity.

(b) Λ^c_{2ms} is nonempty iff $c \in \left(1, \sqrt{2}\right)$. Proposition 1 implies that if $c \geq \sqrt{2}$, then each agent undercuts the partner's promise whenever the partner's promise is positive, which implies that $\Lambda^c_{0ef} = [0, \infty)^2 \Rightarrow \Lambda^c_{2ms} = \emptyset$. The fact that Λ^c_{2ms} is nonempty if $c \in \left(1, \sqrt{2}\right)$ is implied by part (c) below (and the observation that $c < \sqrt{2}$ implies that $0 < \frac{2}{c} - c$).

(c) Assume that $c < \sqrt{2}$ and $0 < \lambda_j < \frac{2}{c} - c$. Let

$$g\left(\lambda_j, c\right) = \max\left\{\frac{1}{(c+\lambda_j)(2-c(c+\lambda_j))} - c, \frac{2-c^2}{c-1}\right\} < \infty.$$

Assume that $\lambda_i > g\left(\lambda_j, c\right)$. In what follows we show that $(\lambda_i, \lambda_j) \in \Lambda^c_{2ms}$. Recall that Lemma 2 says that if $\lambda_j < \frac{2}{c} - c$ and $\lambda_i > \dfrac{1}{(c+\lambda_j)(2-c(c+\lambda_j))} - c$, then $\Theta_i \leq 0$, and hence $R_i = \infty$ (using the first part of the maximum function defining $g\left(\lambda_j, c\right)$). Given that $\lambda_i > \frac{2-c^2}{c-1} > 0$, we have by Lemma 4 that $1 > R_j > 0$ and so $R_i \cdot R_j > 1 > R_j$ and so $(\lambda_i, \lambda_j) \in \Lambda^c_{2-msg}$.

3. Region of no-effort equilibrium Λ^c_{0ef}:

(a) Let $(\lambda_i, \lambda_j) \in \Lambda^c_{0ef}$ and $\left(s_i^*, s_j^*, x_i^*, x_j^*\right) \in PE\left(\lambda_i, \lambda_j, c\right)$. We have to show that $s_i^* = s_j^* = 0$ (which immediately implies that $x_i^*\left(\overrightarrow{s}^*\right) = x_j^*\left(\overrightarrow{s}^*\right) = 0$ because $x_i^e\left(0,0\right) = x_j^e\left(0,0\right) = 0$). The fact that $(\lambda_i, \lambda_j) \in \Lambda^c_{0-eff}$implies that $R_i \cdot R_j < 1$. By the definition of R_i and R_j, this implies that $\Theta_i > 0$ and $\Theta_j > 0$ and $\frac{\lambda_j}{\Theta_i} \cdot \frac{\lambda_i}{\Theta_j} < 1$. By the best-reply correspondence derived in Lemma 1, 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 (and, thus, also the unique trembling hand perfect equilibrium).

(b) We have to show that there exist $0 < \underline{\lambda}_c < \overline{\lambda}_c$, such that $(\lambda_i, \lambda_j) \in \Lambda^c_{0ef}$ if either $0 < \lambda_i, \lambda_j < \underline{\lambda}_c$ or $\lambda_i, \lambda_j > \overline{\lambda}_c$. Let $\overline{\lambda}_c = \frac{2-c^2}{c-1}$ and observe that Lemma 4 implies that if $\lambda_i, \lambda_j > \overline{\lambda}_c$ then $0 < \frac{\lambda_j}{\Theta_i} < 1$ and $0 < \frac{\lambda_i}{\Theta_j} < 1$ and so $R_i \cdot R_j < 1$ and so $(\lambda_i, \lambda_j) \in \Lambda^c_{0ef}$. In particular, if $c \geq \sqrt{2}$, then $\Lambda^c_{0ef} = [0, \infty)$. We are left with showing that for each $c \in \left(1, \sqrt{2}\right)$, there exists $\underline{\lambda}_c > 0$, such that $(\lambda_i, \lambda_j) \in \Lambda^c_{0ef}$ for each $\lambda_i, \lambda_j < \underline{\lambda}_c$. Recall, that Lemma 4 implies that if

$$\lambda_i < \frac{1}{\lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)} - c \tag{32}$$

and the expression holds also with i and j interchanged then $R_i \cdot R_j < 1$ and $(\lambda_i, \lambda_j) \in \Lambda^c_{0-eff}$. Consider the right-hand side. of Eq. (32) when $\lambda_j = 0$: $h\left(c\right) = \dfrac{1}{c(2-c^2)} - c$. Observe that

$h(1)=0$, and that h is increasing in c for each $c \in \left(1,\sqrt{2}\right)$ as

$$h'(c) = -\left(c\left(2-c^2\right)\right)^{-2}\left(2-c^2\right)c\left(-2c\right)-1 = \frac{\left(2-c^2\right)2c^2}{\left(c\left(2-c^2\right)\right)^2}-1 = \frac{2}{\left(2-c^2\right)}-1 >_{\forall c\in\left(1,\sqrt{2}\right)} 0.$$

This implies that $h(c)>0$ for each $c \in \left(1,\sqrt{2}\right)$, which implies by continuity, that there exists a sufficiently small $\underline{\lambda}_c > 0$, such that the right-hand side of Eq. (32) is larger than $\underline{\lambda}_c$ for any $\lambda_j < \underline{\lambda}_c$. This, in turn, implies that $(\lambda_i,\lambda_j) \in \Lambda^c_{0ef}$ for each $\lambda_i,\lambda_j < \underline{\lambda}_c$.

□

E.4 Corollary of Theorem 1

We formalise one corollary of Theorem 1, which says that if players' levels of reneging aversion are identical and positive, they send the same message in the unique perfect equilibrium of the partnership game. This corollary is used in some subsequent proofs.

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 in the Λ^c_{2-msg}) where $s_i \neq s_j$ are those where $R_i \cdot R_j > 1 > R_j$. This implies that $R_i \neq R_j$. By the definition of Θ_i, we see that $\lambda_i=\lambda_j \Rightarrow \Theta_i=\Theta_j$. By 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 $s_i=s_j$. □

E.5 Corollary of Theorem 1 and Lemma 1

Theorem 1 characterises unique equilibria in all but a "measure-zero" set of points of the reneging aversion space that correspond to the boundaries of the three sets defined in the theorem. We demonstrate that at the two points $(\lambda^-_c,\lambda^-_c)$ and $(\lambda^+_c,\lambda^+_c)$ (where $\lambda^-_c = min\{\lambda : (\lambda,\lambda) \in Cl(\Lambda^c_{max})\}$ and $\lambda^+_c = max\{\lambda : (\lambda,\lambda) \in Cl(\Lambda^c_{max})\}$), any pair of identical messages sent by the players can be supported as a perfect equilibrium when $c<1.25$. This result is used in the results of Section 4. The other boundary points do not play a role in our analysis and we refrain from analysing them for the sake of brevity.

Corollary 2. *Let $c \in (1,1.25)$ and let $\lambda_i=\lambda_j=\lambda$. (1) If $\lambda=\lambda^-_c$ or $\lambda=\lambda^+_c$ then $\left((s,s'),\left(x^*_i(\overrightarrow{s}),x^*_j(\overrightarrow{s})\right)\right)$ is a perfect equilibrium of the partnership game if and only if $s=s'$. (2) If $\left(1,x^*_i,1,x^*_j\right)$ is a perfect equilibrium of the partnership game then $(\lambda,\lambda) \in Cl(\Lambda^c_{max})$.*

Proof. Part 1: The proof of part 1 (b) of Theorem 1 demonstrates that by the definitions $\lambda^-_c = min\{\lambda : (\lambda,\lambda) \in Cl(\Lambda^c_{max})\}$ and $\lambda^+_c = max\{\lambda : (\lambda,\lambda) \in Cl(\Lambda^c_{max})\}$, we have that both $\lambda_i=\lambda_j=\lambda^+_c$ and $\lambda_i=\lambda_j=\lambda^-_c$ imply that $R_i=R_j=1$. The best-reply correspondence derived in Lemma 1 then implies that $s^*_i=\mu_{\sigma_j}$ and $s^*_j=\mu_{\sigma_i}$, which are jointly satisfied if and only if $s^*_i=s^*_j$. In order to see that $\left(s^*_i,s^*_j,x^*_i,x^*_j\right)$ is a trembling-hand perfect equilibrium, observe that for each $\epsilon>0$ there exists an ϵ-perfect equilibrium $(\sigma_\epsilon,\sigma_\epsilon) \in \left(\Delta^{fs}([0,1])\right)^2$ satisfying $\sigma_\epsilon(s^*)=1-\epsilon$ and $\mu_{\sigma_\epsilon}=s^*$, which implies that

(s^*, s^*) is a trembling-hand perfect equilibrium of the promise game.
Part 2: This follows from the definitions of λ_c^- and λ_c^+ ($\lambda_c^- = min\{\lambda : (\lambda, \lambda) \in Cl(\Lambda_{max}^c)\}$ and $\lambda_c^+ = max\{\lambda : (\lambda, \lambda) \in Cl(\Lambda_{max}^c)\}$) and from the fact that Λ_{max}^c is a convex set (part 1 (b) of Theorem 1) and so its closure is too. □

E.6 Proof of Theorem 2

This section consists of several lemmas used in the proof of Theorem 2, followed by the proof itself.

E.6.1 Lemma: Positive Payoff Always Possible in the Population Game

Lemma 5. *Fix $c \in (1, 1.25)$. For all $\lambda_j \geq 0$ there exists $\lambda_i \geq 0$ such that in any perfect equilibrium of the partnership game (λ_i, λ_j), player i achieves a strictly positive material payoff, i.e., $\pi(\lambda_i, \lambda_j) > 0$.*

Proof. Theorem 1 and the definition of Λ_{max}^c and Λ_{2ms}^c imply 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 the partnership game when $\lambda_i, \lambda_j > 0$ and in any equilibrium when $\lambda_i > \lambda_j = 0$. We show that for all $\lambda_j \geq 0$ there exists $\lambda_i \geq 0$ 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:

$$c^2 + \frac{1}{(c + \lambda_j)c} - 2 < 0 \iff \frac{1}{c + \lambda_j} < c\left(2 - c^2\right) \iff \frac{1}{c\,(2 - c^2)} - c < \lambda_j. \tag{33}$$

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

$$\frac{3c^2 - 2}{(2c - c^3)^2} - 1. \tag{34}$$

Eq. (34) is positive for $c < 1.25$. The lower bound on λ_j given by Eq. (33) 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 that $\Theta_j < 0$ 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:

$$\lim_{\lambda_i \to \infty} \Theta_i \leq 0 \iff c(c + \lambda_j) - 2 \leq 0 \iff \lambda_j \leq \frac{2}{c} - c. \tag{35}$$

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 $\lambda_i \to \infty$:

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

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:

$$\frac{\lambda_j}{c\left[c(c+\lambda_j)-2\right]}>1\iff c(c+\lambda_j)-2>0\quad and\quad \lambda_j<c\left[c(c+\lambda_j)-2\right]$$
$$\iff\frac{2}{c}-c<\lambda_j<\frac{c}{c^2-1}-c. \tag{36}$$

Observe that the first inequality in Eq. (36) holds precisely when Eq. (35) does not hold. The first derivative of the right-hand side of the second inequality in Eq. (36) 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. (36) is $\frac{35}{36}>\frac{81}{140}$. Therefore, for all $c<1.25$ and $\lambda_j\le\frac{81}{140}$, this second inequality holds. We therefore have that for all $c<1.25$ and $\lambda_j\le\frac{81}{140}$, either $\Theta_i\le 0$ 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 $\lambda_j\ge 0$, there exists a $\lambda_i\ge 0$ such that either $s_i=1$ or $s_j=1$ in the unique equilibrium of the game (λ_i,λ_j) (or in any equilibrium of the game when either $\lambda_i=0$ or $\lambda_j=0$). 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 (due to part 3 of Theorem 1) has strictly positive reneging aversion, which by Eq. (5) implies that both players exert strictly positive effort in equilibrium. Observe that a player can always guarantee a utility level of zero by playing $s_i=x_i=0$. Further, observe that if $\lambda_i>0$ then, by Lemma 1, the uniqueness of the best reply implies that either $\Theta_i>0$ or $[\Theta_i<0$ or $(\Theta_i=0$ *and* $\lambda_j\cdot\mu_{\sigma_j}>0)]$, and therefore the utility function is either strictly concave or strictly increasing (respectively) in s_i. This implies that if $\lambda_i>0$ and the best reply s_i^* is positive (i.e., $s_i^*>0$) and unique, then it must yield strictly positive utility for player i. In the case where $\lambda_i=0$, given that $x_i^*>0$ and the strict concavity of the utility function in x_i and the fact that playing $x_i=0$ guarantees a utility level of zero, the utility of player i must be strictly positive in equilibrium. □

E.6.2 Lemma: Additional Properties of Λ^c_{max}

Lemma 6. *Fix $c\in(1,1.24)$. (1) For all $\lambda\in[\lambda^-_c,\lambda^+_c)$, there exists $\delta_\lambda>0$ such that for all $\lambda'\in(\lambda,\lambda+\delta_\lambda)$, $(\lambda',\lambda)\in\Lambda^c_{max}$ (2) For all $\lambda'\neq\lambda^+_c$, $(\lambda',\lambda^+_c)\notin Cl(\Lambda^c_{max})$.*

Proof. The proof of Theorem 1 yields Eq. (26) and the corresponding condition for player j, which together define Λ^c_{max}. The *strict* convexity of the first inequality of Eq. (26) defining the boundary of Λ^c_{max}, implies that for all $\lambda \in (\lambda^-_c, \lambda^+_c)$, (λ, λ) is not on the boundary of Λ^c_{max} and is therefore in the interior of Λ^c_{max} (i.e., it is in Λ^c_{max} but not in $Cl(\Lambda^c_{max})$) . By the definition of an interior point of a convex set, for all $\lambda \in (\lambda^-_c, \lambda^+_c)$, there exists $\delta_\lambda > 0$ such that for all $\lambda' \in [\lambda, \lambda + \delta_\lambda)$, $(\lambda', \lambda) \in \Lambda^c_{max}$. We next show that there exists $\delta_\lambda > 0$ such that for all $\lambda' \in (\lambda^-_c, \lambda^-_c + \delta_\lambda)$, $(\lambda', \underline{\lambda}_c) \in \Lambda^c_{max}$. This will be the case if and only if there is $\delta_\lambda > 0$ such that Eq. (26) holds whenever $\lambda_i = \lambda^-_c$ and $\lambda_j \in [\lambda^-_c, \lambda^-_c + \delta_\lambda)$. This will be the case if and only if Eq. (26) 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. (26) is less than or equal to zero when evaluated at $\lambda_j = \lambda^-_c$. The derivative of the right-hand side of the first inequality of Eq. (26) with respect to λ_j is

$$\frac{c(2c + 2\lambda_j - 1) - 2(1 + \lambda_j)}{(\lambda_j + c)^2[\lambda_j(c-1) + c^2 - 2]^2}. \tag{37}$$

When evaluated at $\lambda_j = \lambda^-_c$, Eq. (37) is nonpositive if[17] $c > \sqrt{5} - 1 \approx 1.24$. Therefore, we have that for all $\lambda \in [\lambda^-_c, \lambda^+_c)$, there exists $\delta_\lambda > 0$ such that for all $\lambda' \in (\lambda, \lambda + \delta_\lambda)$, $(\lambda', \lambda) \in \Lambda^c_{max}$.

Point (2) is established by noting first that the right-hand side of the first inequality of Eq. (26) 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 λ^-_c). Therefore, given that λ^+_c satisfies Eq. (28), an increase in λ_j with λ_i fixed at λ^+_c means that the weak counterpart of Eq. (26) 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. (28), when λ_j is fixed at λ^+_c, any $\lambda_i < \lambda^+_c$ must violate the weak counterpart of the first inequality in Eq. (26). Therefore, for any $\lambda' \neq \lambda^+_c$, $min(R_i, R_j) < 1$ and so $(\lambda', \lambda^+_c) \notin Cl(\Lambda^c_{max})$. $\square$

E.6.3 Proof of Theorem 2

Proof. We prove each part of the theorem in turn.

1. By the definition of R_i, $\lambda_i = \lambda_j \implies R_i = R_j$, which in turn implies that either $max\{R_i, R_j\} < 1$ and hence $R_i \cdot R_j < 1$ and $(\lambda_i, \lambda_j) \in \Lambda^c_{0ef}$ or $min\{R_i . R_j\} \geq 1$ and hence $(\lambda_i, \lambda_j) \in Cl(\Lambda^c_{max})$. Therefore, $\lambda_i = \lambda_j \implies (\lambda_i, \lambda_j) \in Cl(\Lambda^c_{max}) \cup \Lambda^c_{0ef}$. For any λ such that $(\lambda, \lambda) \in \Lambda^c_{0ef}$, $x_i = x_j = 0$ and so $\pi(\lambda, \lambda) = 0$. To find the material payoff and subjective utility of each player when $(\lambda, \lambda) \in Cl(\Lambda^c_{max})$, we impose $s_i = s_j = 1$ and $\lambda_i = \lambda_j = \lambda$ on the equation for equilibrium effort (Eq. 5):

$$\frac{(c + \lambda)\lambda + \lambda}{(c + \lambda)(c + \lambda) - 1} = \frac{\lambda}{c + \lambda - 1}, \tag{38}$$

and then we substitute $x_i = x_j = \dfrac{\lambda}{c + \lambda - 1}$ and $s_i = s_j = 1$ on Eq. (1) and Eq. (2) to get the equilibrium material payoff and subjective utility, respectively:

$$\pi_c(\lambda, \lambda) = \left(\frac{\lambda}{c + \lambda - 1}\right)^2 \left(1 - \frac{c}{2}\right), \quad U_c(\lambda, \lambda) = \left(\frac{\lambda}{c + \lambda - 1}\right)^2 \left(1 - \frac{c}{2}\right) - \frac{\lambda}{2}\left(\frac{c - 1}{c + \lambda - 1}\right)^2. \tag{39}$$

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

It is immediate that the material payoff is increasing in λ, and, therefore, that the maximal material payoff is obtained at $(\lambda_c^+, \lambda_c^+)$, which is the maximal symmetric point in $Cl(\Lambda_{max}^c)$ (i.e., $\pi\left(\lambda_c^+, \lambda_c^+\right) > \pi\left(\lambda', \lambda'\right)$ for any $\lambda' \neq \lambda_c^+$). We are left to show that the maximal subjective utility is obtained at $(\lambda_c^+, \lambda_c^+)$. Taking the derivative of the subjective utility yields[18]:

$$\frac{\partial U_c(\lambda, \lambda)}{\partial \lambda} = \frac{(c-1)[(3-c)\lambda - (c-1)^2]}{2(c+\lambda-1)^3}. \tag{40}$$

Observe that the denominator is always positive and that the numerator is positive for each $c \in (1,3)$ and $\lambda > \frac{(c-1)^2}{3-c}$. Recall that the minimal value of λ in Λ_{max}^c is $\lambda_c^- \equiv \frac{1+2c-2c^2}{2(c-1)} - \frac{\sqrt{5-4c}}{2(c-1)}$. Thus, we are left with showing that

$$\lambda_c^- \equiv \frac{1+2c-2c^2}{2(c-1)} - \frac{\sqrt{5-4c}}{2(c-1)} > \frac{(c-1)^2}{3-c}.$$

We use Mathematica to show that this inequality is satisfied for each $c \in (1, 1.25)$, which completes the proof (the code is available in the supplementary appendix).

2. Recall that

$$\lambda_c^+ = \frac{1+2c-2c^2}{2(c-1)} + \frac{\sqrt{5-4c}}{2(c-1)} = \frac{1+2c(1-c)+\sqrt{5-4c}}{2(c-1)}. \tag{41}$$

Note that as $c \to 1$, the numerator of Eq. (41) is increasing and the denominator of Eq. (41) converges to zero. Hence $\lim_{c\to 1} \lambda_c^+ = \infty$. To find the limit of the players' material payoff in the game $(\lambda_c^+, \lambda_c^+)$ as $c \to 1$, we substitute the expression for effort in a maximum-message equilibrium (Eq. 38) into that for the payoffs in a symmetric equilibrium (Eq. (39) and Eq. (2)) when $\lambda = \lambda_c^+$:

$$\pi_c(\lambda_c^+, \lambda_c^+) = \left(\frac{\lambda_c^+}{c+\lambda_c^+-1}\right)^2 \left(1 - \frac{c}{2}\right), \quad U_c(\lambda_c^+, \lambda_c^+) = \left(\frac{\lambda_c^+}{c+\lambda_c^+-1}\right)^2 \left(1 - \frac{c}{2}\right) - \frac{\lambda}{2}\left(\frac{c-1}{c+\lambda-1}\right)^2 \tag{42}$$

As $c \to 1$, $\lambda_c^+ \to \infty$ and therefore the limits of Eq. (42) are given by

$$\lim_{c\to 1} \pi_c(\lambda_c^+, \lambda_c^+) = \lim_{c\to 1} \left(\frac{\lambda_c^+}{c+\lambda_c^+-1}\right)^2 \left(1 - \frac{c}{2}\right) = (1 - \frac{1}{2}) = \frac{1}{2}. \tag{43}$$

$$\lim_{c\to 1} U_c(\lambda_c^+, \lambda_c^+) = \lim_{c\to 1} \left(\left(\frac{\lambda_c^+}{c+\lambda_c^+-1}\right)^2 \left(1 - \frac{c}{2}\right) - \frac{\lambda}{2}\left(\frac{c-1}{c+\lambda-1}\right)^2\right) = \frac{1}{2} - 0 = \frac{1}{2}. \tag{44}$$

3. We first show that any unilateral deviation from the candidate equilibrium to a lower level of reneging aversion yields a strictly lower payoff, i.e., $\pi_c\left(\lambda', \lambda_c^+\right) < \pi_c(\lambda_c^+, \lambda_c^+)$ for $\lambda' \in [0, \lambda_c^+)$. Point (2) of Lemma 6 implies that for all $\lambda' \in [0, \lambda_c^+)$, $\left(\lambda', \lambda_c^+\right) \notin Cl(\Lambda_{max}^c)$. Therefore for all such deviations, $\left(\lambda', \lambda_c^+\right) \in Cl\left(\Lambda_{2-msg}^c\right)$ or $\left(\lambda', \lambda_c^+\right) \in Cl\left(\Lambda_{0ef}^c\right)$. Suppose first that $\left(\lambda', \lambda_c^+\right) \in \Lambda_{0ef}^c$. Then the effort levels of both players are zero and so we have $\pi_c\left(\lambda_c^+, \lambda_c^+\right) > \pi_c(\lambda', \lambda_c^+) = 0$. Suppose instead that $\left(\lambda', \lambda_c^+\right) \in \Lambda_{2ms}^c$; the payoff to the deviating player is obtained by substituting the

[18] This derivative is obtained using Mathematica. The code is available in the supplementary appendix of this paper.

expression for equilibrium effort (Eq. 5) 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 deviating player):

$$\pi_c(\lambda_i, \lambda_c^+) = \frac{[(c+\lambda_c^+)\lambda_i \frac{\lambda_j}{\Theta_i} + \lambda_c^+][(c+\lambda_i)\lambda_c^+ + \lambda_i \frac{\lambda_j}{\Theta_i}]}{[(c+\lambda_i)(c+\lambda_c^+)-1]^2} - \frac{c[(c+\lambda_c^+)\lambda_i \frac{\lambda_j}{\Theta_i} + \lambda_c^+]^2}{2[(c+\lambda_i)(c+\lambda_c^+)-1]^2}. \tag{45}$$

The derivative of this expression with respect to λ_i is[19]

$$\frac{[\lambda_c^+]^2(c(\lambda_c^+ + c) - 1)^2}{[1 + (c+\lambda_c^+)(c+\lambda_i)(c(\lambda_c^+ + c) - 2)]^3}. \tag{46}$$

Clearly, the numerator of Eq. (46) is always positive. A *sufficient* condition for the denominator, and hence for the whole expression, to be strictly positive is that

$$c(\lambda_c^+ + c) - 2 > 0 \iff \lambda_c^+ > \frac{2}{c} - c. \tag{47}$$

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 $c < 1.33$. Therefore, $\pi\left(\lambda_c^+, \lambda_c^+\right) > \pi(\lambda', \lambda_c^+)$ for any $\lambda' \in [0, \lambda_c^+)$ such that $\left(\lambda', \lambda_c^+\right) \in \Lambda_{2-msg}^c$. Finally, suppose that $\left(\lambda', \lambda_c^+\right) \in \{Cl(\Lambda_{2ms}) \setminus \Lambda_{2ms}\} = \{(\lambda_i, \lambda_c^+) \subseteq [0,\infty)^2 \colon R_i \cdot R_j = 1 > R_i\}$. By Lemma 1 we have that any equilibrium will satisfy $s_i = \frac{\lambda_j}{\Theta_i} s_j$ and $s_j = min\left\{\frac{\lambda_j}{\Theta_i} s_i, 1\right\}$. Substituting the expression for equilibrium effort (Eq. 5) into the expression for material payoff (Eq. 1) and imposing this form of best reply yields utility to player i (the deviating player):

$$\pi_c(\lambda_i, \lambda_c^+) = \left(\frac{[(c+\lambda_c^+)\lambda_i \frac{\lambda_j}{\Theta_i} + \lambda_c^+][(c+\lambda_i)\lambda_c^+ + \lambda_i \frac{\lambda_j}{\Theta_i}]}{[(c+\lambda_i)(c+\lambda_c^+)-1]^2} - \frac{c[(c+\lambda_c^+)\lambda_i \frac{\lambda_j}{\Theta_i} + \lambda_c^+]^2}{2[(c+\lambda_i)(c+\lambda_c^+)-1]^2}\right) s_j^2. \tag{48}$$

Clearly, the highest possible payoff to player i in any possible equilibrium is that where $s_j = 1$. In this case, the equilibrium payoff is of the same form as Eq. (45) and, by the above arguments, it cannot represent a profitable deviation. We therefore have that $\pi\left(\lambda', \lambda_c^+\right) < \pi(\lambda_c^+, \lambda_c^+)$ for $\lambda' \in [0, \lambda_c^+)$. We now 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 Lemma 6, $\lambda' > \lambda_c^+$ implies that $\left(\lambda', \lambda_c^+\right) \notin Cl\left(\Lambda_{max}^c\right)$. Suppose first that $\left(\lambda', \lambda_c^+\right) \in \Lambda_{0ef}^c$. In this case,

[19] This derivative was calculated using Mathematica. The code is available in the supplementary appendix of this paper.

the effort levels of both players are zero and so we have $\pi_c\left(\lambda_c^+,\lambda_c^+\right) > \pi_c(\lambda',\lambda_c^+) = 0$. Suppose instead that $\left(\lambda',\lambda_c^+\right) \in \Lambda^c_{2-msg}$. In this case, the payoff to the deviating player is obtained by substituting the expression for equilibrium effort (Eq. 5) 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 deviating player):

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

In the supplementary appendix of this paper, we present the explicit formula for the 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 $(\lambda_i,\lambda_c^+) \in \Lambda^c_{2ms}$, we have that $\pi_c\left(\lambda_i,\lambda_c^+\right) < \pi_c(\lambda_c^+,\lambda_c^+)$. Finally, suppose that $\left(\lambda',\lambda_c^+\right) \in \{Cl(\Lambda_{2ms})\setminus\Lambda_{2ms}\} = \{\left(\lambda_i,\lambda_c^+\right) \subseteq [0,\infty)^2 \colon R_i\cdot R_j = 1 > R_j\}$. By arguments analogous to the case where $\lambda_i < \lambda_c^+$, we have that the maximum possible payoff from deviating in this case is of the form given by Eq. (49) and therefore not profitable. Hence for any $\lambda_i > \lambda_c^+$, $\pi_c\left(\lambda_i,\lambda_c^+\right) < \pi_c(\lambda_c^+,\lambda_c^+)$.
Therefore, we have shown that any possible deviation from the pure strategy equilibrium $(\lambda_c^+,\lambda_c^+)$ yields the deviating player a strictly lower payoff and hence this equilibrium is strict.

4. In the sequential game (with no reneging costs) let player i make his effort choice first with player j best-replying to this. Then, in equilibrium, $x_j = \underset{x_j}{argmax}\left\{x_i x_j - \frac{c\cdot x_j^2}{2}\right\} = \frac{x_j}{c}$ and hence $x_i = \underset{x_i}{argmax}\left\{\frac{x_i^2}{c} - \frac{c\cdot x_i^2}{2}\right\} = \underset{x_i}{argmax}\left\{\frac{(2-c^2)x_i}{2c}\right\} = 1$, where the last equality follows from the fact that $c < 1.25$. Therefore, in an equilibrium with sequential effort choices, $x_i = 1$, $x_j = \frac{1}{c}$, and the mean payoff is $\frac{1}{c} - \frac{c}{4}(1+\frac{1}{c^2}) = \frac{3-c^2}{4c}$. The payoff to either player in the equilibrium induced by $(\lambda_c^+,\lambda_c^+)$ is given by Eq. (42). We then have that

$$\pi_c(\lambda_c^+,\lambda_c^+) > \frac{1}{2}\cdot\left(\pi_i^s+\pi_j^s\right) \iff [\frac{\lambda_c^+}{c+\lambda_c^+-1}]^2[1-\frac{c}{2}] > \frac{3-c^2}{4c} \iff c < 1.22.$$

The final step is proven using Mathematica (code is available in the supplementary appendix).

5. Recall from part 1 that $\lambda_i = \lambda_j \implies (\lambda_i,\lambda_j) \in Cl(\Lambda^c_{max}) \cup \Lambda^c_{0ef}$. We consider these two sets of symmetric strategy profiles in turn and show that no candidate equilibria of the population game survive other than $(\lambda_c^+,\lambda_c^+)$ when $c \in (1, 1.24)$. For any λ such that the unique equilibrium in the corresponding partnership game $(\lambda,\lambda) \in \Lambda_{0ef}$, we have that $\pi\left(\lambda,\lambda\right) = 0$. Lemma 5 shows that for $c < 1.25$ and for $\lambda \geq 0$, there exists $\lambda' \geq 0$ such that $\pi\left(\lambda',\lambda\right) > 0$. Therefore, for all λ such that $\pi\left(\lambda,\lambda\right) = 0$, (λ,λ) cannot be a Nash equilibrium of the population game. For any λ such that $(\lambda,\lambda) \in Cl(\Lambda^c_{max})$, we say that such an equilibrium "admits an upward deviation within Λ^c_{max}" if there exists $\delta_\lambda > 0$ such that for all $\lambda' \in (\lambda,\lambda+\delta_\lambda)$, $(\lambda',\lambda) \in \Lambda^c_{max}$. For all λ such that $(\lambda,\lambda) \in Cl(\Lambda^c_{max})$, the equilibrium payoff to both players is obtained by substituting $s_i = s_j = 1$

into Eq. (11):

$$\pi_c(\lambda_i, \lambda_j) = \frac{[(c+\lambda_j)\lambda_i + \lambda_j][(c+\lambda_i)\lambda_j + \lambda_i]}{[(c+\lambda_i)(c+\lambda_j)-1]^2} - \frac{c[(c+\lambda_j)\lambda_i + \lambda_j]^2}{2[(c+\lambda_i)(c+\lambda_j)-1]^2}. \tag{50}$$

The first derivative of this function with respect to λ_i is

$$\frac{(c-1)(1+c+\lambda_j)[\lambda_i c^3 + 2c^2\lambda_i\lambda_j + c\lambda_i(\lambda_j^2 - \lambda_j - 2) - \lambda_j(1+\lambda_i(2+\lambda_j))]}{[c^2 - 1 + \lambda_i\lambda_j + c(\lambda_i + \lambda_j)]^3}. \tag{51}$$

Imposing the condition $\lambda_i = \lambda_j = \lambda$, we can simplify this expression to[20]

$$\frac{(c-1)[c(c+\lambda-1) - 1 - \lambda]\lambda}{[c+1+\lambda][c-1+\lambda]^3}. \tag{52}$$

This expression is strictly positive if and only if

$$c(c+\lambda-1) - 1 - \lambda > 0 \iff \lambda < \frac{1+c-c^2}{c-1}. \tag{53}$$

Recall from Theorem 1 and Corollary 2 that a maximum-message equilibrium exists only if $min(R_i, R_j) \geq 1$ and that this requires that either $\Theta_i \leq 0$ or $\frac{\lambda_j}{\Theta_i} \geq 1$ (and that the analogous conditions hold for j). By Lemma 2 and Lemma 3 each of these conditions implies that

$$\lambda_j < \frac{2-c^2}{c-1}. \tag{54}$$

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

$$\lambda < \frac{2-c^2}{c-1} < \frac{1+c-c^2}{c-1}, \tag{55}$$

where the second inequality clearly follows when $c > 1$. We can see that this yields the second inequality in Eq. (53) and hence Eq. (52) is always positive in a maximum-message equilibrium. Therefore, for any λ such that (λ, λ) "admits an upward deviation within Λ^c_{max}," there exists some $\lambda' > \lambda$ such that $\pi_c\left(\lambda', \lambda\right) > \pi_c(\lambda, \lambda)$ 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 that admit a maximum-message equilibrium and do not "admit an upward deviation within Λ^c_{max}." Lemma 6 implies that there is a unique pair $(\lambda^+_c, \lambda^+_c)$ that fulfills these conditions when $c \in (1, 1.24)$.

□

[20]The derivative given by Eq. (51) and its simplification when $\lambda_i = \lambda_j$ is obtained using Mathematica. The code available in the supplementary appendix of this paper.

E.7 Proof of Proposition 2

Proof. We solve for the subgame-perfect equilibria of this game using backwards induction. Best replies and equilibrium choices of effort in the last stage are the same function of prior-stage messages as in the games with simultaneous communication and are given by Eq. (4) and Eq. (5). Utility as a function of messages is therefore given by Eq. (6). We first note that Section 3.2 demonstrated that if $\Theta_i \leq 0$ then (other than in the "knife edge" case where $\Theta_k = 0$ and $\lambda_l \cdot s_l = 0$ for $k = i$ and $l = j$ or for $k = j$ and $l = i$), regardless of player j's choice of message, player i's level of utility is always increasing in his message, and his optimal choice is $s_i = 1$ for any message sent by j. In the "knife edge" cases where $\Theta_i = 0$ (i.e., where player i is the first player to make a promise) and $R_j \neq 0$ and $\lambda_j \neq 0$, then player j will respond to any positive promise with $s_j = min\{R_j s_i, 1\} > 0$, meaning that i's utility is convex and increasing in his message and he chooses $s_i = 1$. In the "knife edge" cases where $\Theta_j = 0$ and $R_i \neq 0$, player i knows that playing $s_i > 0$ will induce $s_j = 1$. Given that playing $s_i = min\{R_i s_j, 1\}$ is a best reply when taking $s_j = 1$ as given in the simultaneous game, it must also be a best reply in the sequential game. Therefore, if either $\Theta_i \leq 0$ or $\Theta_j \leq 0$ or both of these conditions hold, equilibrium messages and effort levels will be the same under sequential communication as under simultaneous communication.[21]

In the case where $\Theta_i, \Theta_j > 0$, the second-stage best reply of player j (the second player to make a promise) is derived in the same way as the first-stage best reply under simultaneous communication, except that instead of the expectation of player i's promise, we derive the best reply as a function of his actual promise. From the analysis in Section 3.2 we therefore know that player j will choose $s_j = min\{R_j s_i, 1\}$.

Next, we analyse the choice of player i taking j's second-stage best reply function as given. First, we show that when $\Theta_i, \Theta_j > 0$, there exists no equilibrium in which $s_i \in (0, min\{\frac{1}{R_j}, 1\})$. Note that when $\Theta_i, \Theta_j > 0$, then $R_j < \infty$ and so, for $s_i \in (0, min\{\frac{1}{R_j}, 1\})$, inserting the best reply $s_j = min\{R_j s_i, 1\} = R_j s_i$ into Eq. (5) and substituting this into player i's utility function yields

$$\begin{aligned} U_i(s_i, c) =& \frac{[(c+\lambda_j)\lambda_i s_i + \lambda_j R_j s_i][(c+\lambda_i)\lambda_j R_j s_i + \lambda_i s_i]}{[(c+\lambda_i)(c+\lambda_j)-1]^2} \\ & - \frac{c[(c+\lambda_j)\lambda_i s_i + \lambda_j R_j s_i]^2}{2[(c+\lambda_i)(c+\lambda_j)-1]^2} - \frac{\lambda_i}{2}\left[s_i - \frac{(c+\lambda_j)\lambda_i s_i + \lambda_j R_j s_i}{(c+\lambda_i)(c+\lambda_j)-1}\right]^2 \\ =& \Psi(\lambda_i, \lambda_j, c)\, s_i^2. \end{aligned} \tag{56}$$

Here, $\Psi(\lambda_i, \lambda_j, c)$ is a function of the parameters λ_i, λ_j, c only. Therefore, if there exists $s_i^{'} \in (0, min\{\frac{1}{R_j}, 1\})$ such that $U_i(s_i^{'}, c) > 0$, then $U_i(s_i, c) < U_i(min\{\frac{1}{R_j}, 1\}, c)$ for all $s_i \in (0, min\{\frac{1}{R_j}, 1\})$. Conversely, if there exists $s_i^{'} \in (0, min\{\frac{1}{R_j}, 1\})$ such that $U_i(s_i^{'}, c) < 0$, then $U_i(s_i^{'}, c) < U_i(0, c)$ for all $s_i \in (0, min\{\frac{1}{R_j}, 1\})$.

Consider first the case where $R_j \leq 1$. The fact that there exists no equilibrium in which $s_i \in (0, min\{\frac{1}{R_j}, 1\})$ implies that if $R_j \leq 1$, then player i's optimal choice is $s_i = 1$ if his utility following the

[21] In the case where $\Theta_k = 0$ and $[\lambda_l = 0$ or $R_l = 0]$ for $k = i$ or $k = j$, multiple equilibria are possible. Our results are invariant to what is assumed about equilibrium selection in this case. For ease of exposition, we assume that players in this case play $s_i = 1$ and $s_j = 0$.

subsequent equilibrium play is positive (i.e., if $\Psi > 0$) and the optimal choice is $s_i = 0$ otherwise (as this message guarantees a utility level of zero). We know from the analysis of simultaneous communication that if $R_i R_j \geq 1 \geq R_j$, then there exists an equilibrium in which $s_i = 1$ and $s_j = R_j$ and hence the utility of player i is positive in this case and in the corresponding candidate equilibrium under sequential communication (as subsequent effort levels are identical following simultaneous or sequential communication of the same pair of messages). Therefore if $R_i R_j \geq 1 \geq R_j$, then $s_i = 1$ is a unique best reply and there is a unique equilibrium in which $s_i = 1$ and $s_j = R_j$. This equilibrium under sequential communication yields the same utility levels and payoffs to both players as that under simultaneous communication. If $R_i R_j < 1$ then there exists either a unique equilibrium in which $s_i = 1$ and $s_j = R_j$ or a unique equilibrium in which $s_i = 0$ and $s_j = 0$. In the latter case, the utility levels and payoffs are the same as under simultaneous communication. In the former case, they are strictly greater for both players.

Consider next the case in which $R_j > 1$. We have thus far shown that i's best reply is either 0 or in $[\frac{1}{R_j}, 1]$. We first consider the optimal choice of message from the interval $[\frac{1}{R_j}, 1]$. For all $s_i \in [\frac{1}{R_j}, 1]$, j's best reply is fixed, $s_j = 1$. Player i chooses his message taking j's choice as given, which is the same optimisation problem as under simultaneous communication. From Section 3.2 we know that i's best reply from this interval is therefore $s_i = min\left\{max\{R_i, \frac{1}{R_j}\}, 1\right\}$. From Section 3.2 we know that if $\min(R_i, R_j) \geq 1$, then there exists an equilibrium under simultaneous communication in which $s_i = s_j = 1$ and both players achieve positive utility. This implies that if $R_i \geq 1$ (and hence $\min(R_i, R_j) \geq 1$) then there is a unique subgame-perfect equilibrium under sequential communication in which $s_i = s_j = 1$. If $R_i < 1$ and $R_i R_j \geq 1 > R_i$ then i's optimal choice is R_i so long as this yields positive utility. We know from Section 3.2 that if $R_i R_j \geq 1 > R_i$ then there exists an equilibrium under simultaneous communication in which $s_i = R_i$ and $s_j = 1$ and hence the same messages form part of the unique subgame-perfect equilibrium with sequential communication. This equilibrium yields the same utility levels and payoffs to both players as under simultaneous communication. If $R_i R_j < 1$ then $R_i < \frac{1}{R_j}$ and there exists either a unique equilibrium in which $s_i = \frac{1}{R_j}$ and $s_j = 1$ or a unique equilibrium in which $s_i = 0$ and $s_j = 0$. In the latter case, the utility levels and payoffs are the same as under simultaneous communication. In the former case, they are strictly greater for both players.

We have therefore seen that if either ($\Theta_i \leq 0$ or $\Theta_j \leq 0$ [or both]) or ($\Theta_i > 0$ and $\Theta_j > 0$ and $R_i R_j \geq 1$), then there is a unique equilibrium in which (1) players' payoffs are invariant to whether they send their message first or second and hence to the method by which nature selects the first mover, and (2) both players' messages, efforts, and payoffs are the same as under simultaneous communication. If $\Theta_i > 0$ and $\Theta_j > 0$ and $R_i R_j < 1$ and $R_j < 1$, then in equilibrium either ($s_i = 1$ and $s_j = R_j$) or ($s_i = 0$ and $s_j = 0$). If $\Theta_i > 0$ and $\Theta_j > 0$ and $R_j \geq 1 > R_i R_j$, then in equilibrium either ($s_i = \frac{1}{R_j}$ and $s_j = 1$) or ($s_i = 0$ and $s_j = 0$); i.e., when $\Theta_i > 0$ and $\Theta_j > 0$ and $R_i R_j < 1$ we have that (1) equilibrium messages, efforts and payoffs may depend on which player is selected to send their message first, and (2) payoffs may be strictly greater under sequential communication than under simultaneous communication.

We can now see that under sequential communication, $(\lambda_c^+, \lambda_c^+)$ induces the same messages, effort levels, and payoffs as under simultaneous communication (point 1 of the proposition) by noting that in the partnership game induced by the pair $(\lambda_c^+, \lambda_c^+)$, by the definition of λ_c^+ we have $\min(R_i, R_j) \geq 1 \Longrightarrow R_i R_j \geq 1$. We next establish that $(\lambda_c^+, \lambda_c^+)$ remains a strict Nash equilibrium of the population game (point 4 of the proposition). Consider a deviation to $\lambda' \neq \lambda_c^+$. Recall from Lemma 6 that $\lambda' \neq \lambda_c^+ \Longrightarrow \left(\lambda', \lambda_c^+\right) \notin Cl(\Lambda_{max})$ and hence $min\{R_i, R_j\} < 1$. First, consider a deviation to $\lambda' < \lambda_c^+$. We show that this implies that for player k, with $\lambda_k = \lambda_c^+$, we have that $R_k > 1$ (with $k = i$ or $k = j$, depending on which player is selected to send his message first). To see this, observe that as established in the proof of Theorem 2 (specifically in Eq. (27)), the set of points satisfying Eq. (26) is convex. If we can establish that in the game induced by $(0, \lambda_c^+)$ we have $R_k > 1$, then for $\lambda' \in [0, \lambda_c^+]$ we have that in the game induced by $\left(\lambda', \lambda_c^+\right)$, $R_k > 1$. Reproducing Eq. (26) but imposing $\lambda_i = \lambda_k = \lambda_c^+$ and $\lambda_j = \lambda' = 0$, we have that $R_k > 1$ if

$$
\begin{aligned}
& \lambda_c^+ > \frac{1}{(\lambda')^2(1-c) + \lambda'(2 - 2c^2 + c) + c(2 - c^2)} - c \quad \text{and} \quad \lambda' < \frac{2 - c^2}{c - 1} \\
\Longleftrightarrow\ & \lambda_c^+ > \frac{1}{c(2 - c^2)} - c \quad \text{and} \quad 0 < \frac{2 - c^2}{c - 1} \\
\Longleftrightarrow\ & \frac{1 + 2c - 2c^2}{2(c-1)} + \frac{\sqrt{5 - 4c}}{2(c-1)} > \frac{1}{c(2 - c^2)} - c,
\end{aligned}
$$

which holds for all $c \in [1, 1.25]$. Hence we have that $R_k > 1$. Given that $min\{R_i, R_j\} < 1$, we have that either $R_i > 1 > R_j$ or $R_j > 1 > R_i$. If $R_i > 1 > R_j$ then as shown above either (1) the payoff in equilibrium is the same as under simultaneous communication, or (2) there is an equilibrium in which $s_i = 1$ and $s_j = R_j$. If $R_j > 1 > R_i$ then as shown above either (3) the payoff in equilibrium is zero to both players, or (4) there is an equilibrium in which $s_i = \frac{1}{R_j}$ and $s_j = 1$. The proof of part 3 of Theorem 2 shows that for any $\left(\lambda', \lambda_c^+\right)$ such that $\lambda' < \lambda_c^+$, $\pi_c\left(\lambda', \lambda_c^+\right) < \pi_c\left(\lambda_c^+, \lambda_c^+\right)$ in the simultaneous communication setup and therefore, in cases (1) and (3) any deviation yields a strictly lower payoff also in the sequential setup. The proof of part 3 of Theorem 2 also shows that for any (λ_i, λ_c^+) such that $\lambda_i \leq \lambda_c^+$, if $s_j = 1$ and $s_i = R_i$, then $\pi_c\left(\lambda_i, \lambda_c^+\right) < \pi_c\left(\lambda_c^+, \lambda_c^+\right)$. Given that play following any given pair of messages is the same in both the sequential and simultaneous setups, it is therefore the case that in case (2) deviation also yields a strictly lower payoff. In case (4), the equilibrium payoff to player i is given by substituting $s_j = 1$, $s_i = \frac{1}{R_j}$, $\lambda_i = \lambda'$, and $\lambda_j = \lambda_c^+$ into the expressions for the second-stage effort choices given by Eq. (5) and substituting the resulting expression into Eq. (1). Simplifying the resulting expression yields [22]

$$
\frac{[(c(c+\lambda') - 1][c - 3c\sqrt{5-4c} + c^3(1+\sqrt{5-4c}) + c^2\lambda'(1+\sqrt{5-4c}) - 2(2 + \lambda' + \sqrt{5-4c}\lambda')]}{2(1+\sqrt{5-4c})(c+\lambda')^2}. \tag{57}
$$

The payoff $\pi\left(\lambda_c^+, \lambda_c^+\right)$ is obtained by substituting $\lambda_i = \lambda_j = \lambda_c^+$ and $s_i = s_j = 1$ into Eq. (5) and substituting the resulting expression into Eq. (1). Simplifying the resulting expression yields[23]

$$
\frac{(c-2)(1+\sqrt{5-4c} - 2(c-1)c)^2}{2(3+\sqrt{5-4c} - 2c)^2}. \tag{58}
$$

[22] This simplification was obtained using Mathematica. The code is available in the supplementary appendix of this paper.

[23] This simplification was obtained using Mathematica. The code is available in the supplementary appendix of this paper.

The value of Eq. (58) is strictly greater than that of Eq. (57) for all $\lambda^{'} < \lambda_c^+$ for all[24] $c < 1.25$. Therefore, in all possible cases (1), (2), (3) and (4), if $\lambda_k = \lambda_c^+$ then a deviation by player l from λ_c^+ to $\lambda^{'} < \lambda_c^+$ yields a strictly lower payoff and hence for $\lambda^{'} < \lambda_c^+$, we have that $\pi\left(\lambda^{'}, \lambda_c^+\right) < \pi\left(\lambda_c^+, \lambda_c^+\right)$.

Next, consider a deviation to $\lambda_l = \lambda^{'} > \lambda_c^+ = \lambda_k$. This implies that $R_l > 1$ as, given that the pair $(\lambda_c^+, \lambda_c^+)$ satisfies Eq. (26), if $\lambda_l > \lambda_c^+$ then the pair $\left(\lambda_c^{'}, \lambda_c^+\right)$ must also satisfy Eq. (26) with $l = i$ and $k = j$, as λ_i is increased while λ_j stays constant. Hence we have that $R_i > 1 > R_j$ or $R_j > 1 > R_i$ (depending on whether player k or l is selected to play first), and again either (1) the payoffs in equilibrium are the same as under simultaneous communication, or (2) there is an equilibrium in which $s_i = 1$ and $s_j = R_j$, or (3) the payoff in equilibrium is zero to both players, or (4) there is an equilibrium in which $s_i = \frac{1}{R_j}$ and $s_j = 1$. Part 3 of Theorem 2 shows that for any $\left(\lambda^{'}, \lambda_c^+\right)$ such that $\lambda^{'} > \lambda_c^+$, it is the case that $\pi_c\left(\lambda^{'}, \lambda_c^+\right) < \pi_c\left(\lambda_c^+, \lambda_c^+\right)$ in the simultaneous communication setup, and that for $\lambda_i > \lambda_c^+$ and following messages of $s_i = 1$ and $s_j = R_j$, the payoff to player i is strictly lower than $\pi_c\left(\lambda_c^+, \lambda_c^+\right)$. Therefore, in cases (1), (2), and (3), deviation to $\lambda^{'} > \lambda_c^+$ leads to a strictly lower payoff. In case (4), the equilibrium payoff to player j (the deviator) is given by substituting $s_i = \frac{1}{R_j}$ and $s_j = 1$ and $\lambda_i = \lambda_c^+$ into the expressions for the second-stage effort choices given by Eq. (5) and substituting the resulting expression into Eq. (1). Simplifying the resulting expression yields[25]

$$\frac{\sqrt{5-4c}+c-2}{2} + \frac{(3-\sqrt{5-4c})(2\lambda^{'}+c)+2(\lambda^{'})^2}{4(c+\lambda^{'})^2}. \tag{59}$$

The value of Eq. (58) is strictly greater than that of Eq. (59) for all $\lambda^{'} < \lambda_c^+$ for all[26] $c < 1.25$. Therefore in all possible cases (1), (2), (3) and (4), if $\lambda_k = \lambda_c^+$ then deviation by player l from λ_c^+ to $\lambda^{'} < \lambda_c^+$ yields a strictly lower payoff and hence for $\lambda^{'} \neq \lambda_c^+$ we have that $\pi_c\left(\lambda^{'}, \lambda_c^+\right) < \pi_c\left(\lambda_c^+, \lambda_c^+\right)$ and $\left(\lambda_c^+, \lambda_c^+\right)$ is a strict equilibrium of the game with sequential communication.

That points 3 and 5 of Theorem 2 also hold in the sequential setup follows from the fact that $\pi_c\left(\lambda_c^+, \lambda_c^+\right)$ is the same under both forms of communication (these results are points 3 and 5 of the proposition).

Finally, we prove that if $c < 1.18$, then $\pi_c\left(\lambda_c^+, \lambda_c^+\right) > \pi_c\left(\lambda^{'}, \lambda^{'}\right)$ and $U_c\left(\lambda_c^+, \lambda_c^+\right) > U_c\left(\lambda^{'}, \lambda^{'}\right)$ for any $\lambda' \neq \lambda_c^+$ (point 2 of the theorem). If $\lambda_i = \lambda_j = \lambda^{'}$, this implies $min\{R_i, R_j\} \geq 1$ or $R_i \cdot R_j < 1$. In the former case, material payoffs and subjective utilities are the same as in the game with simultaneous communication, which by point 1 of Theorem 2 are less than $\pi\left(\lambda_c^+, \lambda_c^+\right)$. In the latter case, either the effort choices are the same as in the game with simultaneous communication (and both players' material payoffs and subjective utilities are zero and therefore less than $\pi_c\left(\lambda_c^+, \lambda_c^+\right)$ and $U_c\left(\lambda_c^+, \lambda_c^+\right)$, respectively) or there is an equilibrium in which $s_i = 1$ and $s_j = R_j$. In the latter of these two cases, material payoffs to players i and j can be obtained by substituting $\lambda_i = \lambda_j = \lambda^{'}$ and $s_i = 1$ and $s_j = R_j$ into Eq. (5) and substituting the resulting expression into Eq. (1) and the corresponding equation for player j. Letting

[24]This result is obtained using Mathematica. The code is available in the supplementary appendix of this paper.
[25]This simplification was obtained using Mathematica. The code is available in the supplementary appendix of this paper.
[26]This result is obtained using Mathematica. The code is available in the supplementary appendix of this paper.

$\pi_i\left(\lambda^{'},\lambda^{'}\right)$ denote the payoff to the player selected to send his message first, we have that

$$\pi_i\left(\lambda^{'},\lambda^{'}\right)=\frac{(\lambda^{'})^2(c+\lambda^{'})[1-c(c+\lambda^{'})][2+(c+\lambda^{'})[-2\lambda^{'}+c(c^2+c\lambda^{'}-3)]]}{2[1+(c+\lambda^{'})^2(-2+c(c+\lambda^{'}))]^2}.$$

Player j's payoff is

$$\pi_j\left(\lambda^{'},\lambda^{'}\right)=\frac{(\lambda^{'})^2(c+\lambda^{'}-1)(1+c+\lambda^{'})[c(c+\lambda^{'}-1)(1+c+\lambda^{'})-2\lambda^{'}]}{2[1+(c+\lambda^{'})^2(-2+c(c+\lambda^{'}))]^2}.$$

By using Mathematica (the code is available in the supplementary appendix) we have verified that for all $c<1.18$, $\frac{\pi_i\left(\lambda^{'},\lambda^{'}\right)+\pi_j\left(\lambda^{'},\lambda^{'}\right)}{2}=\pi\left(\lambda^{'},\lambda^{'}\right)<\pi\left(\lambda_c^+,\lambda_c^+\right)$; i.e., the average payoff to the two players in any candidate equilibrium in which $\lambda_i=\lambda_j=\lambda^{'}$ and $R_i\cdot R_j<1$ and $s_i=1$ and $s_j=R_j$ is strictly less than $\pi\left(\lambda_c^+,\lambda_c^+\right)$. We derive the analogous expressions for $U_i\left(\lambda^{'},\lambda^{'}\right)$ and $U_j\left(\lambda^{'},\lambda^{'}\right)$ in the supplementary Mathematica appendix and we show there that for all $c<1.18$, $\frac{U_i\left(\lambda^{'},\lambda^{'}\right)+U_j\left(\lambda^{'},\lambda^{'}\right)}{2}=U_c\left(\lambda^{'},\lambda^{'}\right)<\pi_c\left(\lambda_c^+,\lambda_c^+\right)$, which completes the proof. □

E.8 Proof of Proposition 3

Proof. In the exposition of this proof, let $\pi_c^1(\lambda_i,\lambda_j)$ denote the payoff to player i in the unique equilibrium of the partnership game with one-sided reneging costs (in cases where there are multiple equilibria that may be selected with positive probability by the equilibrium selection function, $\pi_c^1(\lambda_i,\lambda_j)$ denotes the expected payoff) and use $\pi_c(\lambda_i,\lambda_j)$ to denote the payoff in the corresponding two-sided case. We first derive the second-stage best-reply function under one-sided reneging costs. Individuals have an expectation of their partner's effort choice, denoted by μ_{χ_j}. Their expected utility function is

$$U_i(x_i,\mu_{\chi_j},s_i,c)=x_i\mu_{\chi_j}-\frac{cx_i^2}{2}-\mathbf{1}_{s_i>x_i}\frac{\lambda_i}{2}(s_i-x_i)^2. \tag{60}$$

Suppose first that $s_i\le\frac{\mu_{\chi_j}}{c}$. As the sum of the first two terms of the expected utility function, $x_i\mu_{\chi_j}-\frac{cx_i^2}{2}$, is maximised when $x_i=\frac{\mu_{\chi_j}}{c}$ and the intrinsic cost term, $\mathbf{1}_{s_i>x_i}\frac{\lambda_i}{2}(s_i-x_i)^2$, is minimised for any $x_i>s_i$, we have that the best reply is $x_i=\frac{\mu_{\chi_j}}{c}$. Suppose instead that $s_i>\frac{\mu_{\chi_j}}{c}$. There is no intrinsic cost from playing $x_i>s_i$ but, due to the concavity of utility in x_i, there is a loss induced to the material payoff and so $x_i=s_i$ dominates all $x_i>s_i$. When a player optimises over $x_i\in[0,s_i]$, his optimal choice is characterised by the same first-order condition and so we have the same best-reply function as in the case of two-sided reneging costs (given by Eq. (4)). Players therefore always choose pure strategies. The second-stage best-reply function is therefore:

$$x_i^*(x_j,s_i,s_j,\lambda_i,\lambda_j,c)=\begin{cases}\frac{x_j}{c} & if\ s_i\le\frac{x_j}{c}\\ \dfrac{x_j+\lambda_i s_i}{c+\lambda_i} & if\ s_i>\frac{x_j}{c}\end{cases}. \tag{61}$$

For expositional convenience and without loss of generality, in writing this best-reply function we have imposed that players choose pure strategies. We can deduce from the best-reply function the following facts. (1) In any equilibrium either $s_i > x_i$ or $s_j > x_j$ i.e., at most one player reneges upwards in equilibrium. To see why (1) is true, suppose that $s_i \leq x_i$ and $s_j \leq x_j$. If, for some i, $s_i > \frac{x_j}{c}$, then $x_i^* = \dfrac{x_j + \lambda_i s_i}{c + \lambda_i} < \dfrac{c \cdot s_i + \lambda_i s_i}{c + \lambda_i} = s_i$ which is a contradiction. If instead $s_i \leq \frac{x_j}{c}$ and $s_j \leq \frac{x_i}{c}$, then $x_i = \frac{x_j}{c} > x_j$ and $x_j = \frac{x_i}{c} > x_i$, which is also a contradiction. (2) By comparing the best-reply function to Eq. (4), we see that in any equilibrium in which both players renege downwards, effort choices are the same function of the player's message and the opponent's effort choice as in the model with two-sided reneging costs and hence equilibrium effort choices when both players renege downwards are the same function of first-stage messages as in the two-sided model. (3) In any equilibrium in which a player, i, reneges upwards, we have that $s_i < x_i = \frac{x_j}{c} < x_j = \dfrac{x_i + \lambda_j s_j}{c + \lambda_j} < s_j$, which implies the following effort choices in equilibrium:

$$x_i^e(x_j, s_i, s_j, \lambda_i, \lambda_j, c) = \frac{\lambda_j}{c^2 + c\lambda_j - 1} s_j \equiv \alpha_i s_j \tag{62}$$

$$x_j^e(x_i, s_i, s_j, \lambda_i, \lambda_j, c) = \frac{\lambda_j}{c + \lambda_j - \frac{1}{c}} s_j \equiv c\alpha_i s_j. \tag{63}$$

Note that $\alpha_i < 1$. Therefore, a pair of first-stage messages (s_i, s_j) induce a second-stage equilibrium in which one player, i, reneges upward only if $s_i \leq \alpha_i s_j$. In what follows we show that for any (λ_i, λ_j), $\alpha_i < R_i$. Either $\Theta_i \leq 0$, and so by definition $R_i = \infty > 1 > \alpha_i$, or $\Theta_i > 0$, in which case

$$\begin{aligned}
&\alpha_i < R_i \iff \alpha_i < \frac{\lambda_j}{\Theta_i} \\
\iff &\frac{\lambda_j}{c^2 + c\lambda_j - 1} < \frac{\lambda_j}{c(c + \lambda_j) + \frac{1}{(c+\lambda_i)(c+\lambda_j)} - 2} \\
\iff &c(c + \lambda_j) + \frac{1}{(c + \lambda_i)(c + \lambda_j)} - 2 < c^2 + c\lambda_j - 1 \\
\iff &c^2 + c\lambda_j + \frac{1}{(c + \lambda_i)(c + \lambda_j)} - 2 < c^2 + c\lambda_j - 1 \\
\iff &\frac{1}{(c + \lambda_i)(c + \lambda_j)} < 1,
\end{aligned} \tag{64}$$

which always holds as $c > 1$.

We can now show that for any (λ_i, λ_j) the only candidate equilibrium in which both players renege downward induces the same effort levels as in the unique equilibrium with two-sided reneging costs. Consider a candidate equilibrium in which both players renege downward, i.e., $\alpha_i s_j < s_i$ and $\alpha_j s_i < s_j$. This implies that $s_i \in (\alpha_i s_j, min\{\frac{s_j}{\alpha_j}, 1\}]$ and $s_j \in (\alpha_j s_i, min\{\frac{s_i}{\alpha_i}, 1\}]$. Lemma 1 implies that in the two-sided game, conditional on reneging downward, each player i's best reply is $s_i = min\{R_i s_j, 1\}$. Given that for both players, i, $R_i s_j > \alpha_i s_j$ and $\alpha_j s_i < 1$, their optimal choice of message in the game with one-sided reneging costs, conditional on reneging downward, must satisfy $s_i^* = min\{R_i s_j, \frac{s_j}{\alpha_j}, 1\}$. If $\min\{R_i, R_j\} > 1$ then this implies $s_i^* > s_j$ or $s_i^* = 1$ for both players i, and this is jointly satisfied only if $s_i^* = s_j^* = 1$, which is the same choice of messages as in the corresponding game with two-sided costs.

If $R_i \cdot R_j > 1 > R_j$, then we have that $s_j^* = R_j s_i$. Rearranging $s_j^* = R_j s_i$ we obtain

$$s_i = \frac{s_j}{R_j} < \frac{s_j}{\alpha_j} \tag{65}$$

and

$$s_i = \frac{s_j}{R_j} < \frac{s_j}{R_j} R_i \cdot R_j = R_i s_j, \tag{66}$$

where the inequality in Eq. (66) follows given that $R_i \cdot R_j > 1$. Eq. (65) and Eq. (66) are consistent with i's optimal choice only if $s_i^* = 1$. We therefore have that if $R_i \cdot R_j > 1 > R_j$ then $s_i^* = 1 > R_j s_i = s_j^*$, which implies the same messages and effort choices as in the game with two-sided reneging costs. If $R_i \cdot R_j = 1$, a continuum of candidate equilibria survive in which messages satisfy $s_i^* = s_j^*$ and effort levels correspond to those in the equilibria of the game with two-sided reneging costs. In the case where $R_i = R_j = 1$ we make the assumption, made also in the model with two-sided costs, that the equilibrium selected is that in which $s_i = s_j = 1$. Finally, consider the case where $R_i \cdot R_j < 1$. Assume without loss of generality that $R_i < 1$. Given that $R_i > \alpha_i$ for both players i, we have that $s_i^* = R_i s_j$ and $s_j^* = min\{R_j s_i, 1\}$. Suppose that $s_j^* = 1$; then $s_i^* = R_i$ and $s_j^* = R_i \cdot R_j < 1$, which is a contradiction. Suppose instead that $s_j^* = R_j s_i < 1$; then $s_i^* = R_i s_j = R_i \cdot R_j \cdot s_i < s_i$, which is a contradiction. Therefore, the only equilibrium candidate in which players renege (weakly) downward is that where $s_i = s_j = x_i = x_j = 0$; i.e., the messages and effort levels are the same as in the unique equilibrium of the game with two-sided costs. We have therefore seen that in all possible cases, the messages and effort choices in the unique candidate equilibrium where both players renege downwards are the same as those in the unique equilibrium of the corresponding game with two-sided reneging costs.

We next show that for any (λ_i, λ_j) there are two candidate continua of equilibria (one for each player), where the effort choices are the same within each continuum, and one player reneges upwards. Suppose that there is an equilibrium in which player i reneges upward and $s_j < 1$. Player j must achieve positive utility in equilibrium (otherwise he could do better by playing $s_j = x_j = 0$). Since both players' effort choices are linear functions of s_j, for choices of message that satisfy $s_i \leq \alpha_i s_j$, substituting these linear functions into the utility function implies that j's utility is a linear function of s_j^2 whenever $s_i \leq \alpha_i s_j$. This implies that deviating to $s_j = 1$ yields higher utility for player j. Thus, we are left with two candidate continua of equilibria in which one player reneges upward. These are strategies that satisfy (for each player i) $s_i \leq \alpha_i s_j = \alpha_i$, with second-stage best replies as given in Eq. (62) and Eq. (63). Note that the effort levels of the player in any of these candidate equilibria are independent of s_i (conditional on its being less than α_i) and, therefore, they are the same in all candidate equilibria in the same continuum. Note by comparing to Eq. (5) that the effort levels of both players are the same as in all equilibria of the partnership game with two-sided reneging costs where $\lambda_i = 0$.

We now analyse the game with levels of reneging aversion $(\lambda_c^+, \lambda_c^+)$ and show that there is a unique equilibrium in which the messages and effort levels are the same as in the two-sided reneging cost game. We know that the only candidate equilibrium in which both players renege downward must involve the same choice of messages and effort levels as in the two-sided game. This candidate equilibrium is the one where $s_i = s_j = 1$. We now show that this candidate equilibrium is a subgame-perfect equilibrium by showing that neither player would wish to deviate to any message that would induce them to renege upwards in the second stage (i.e., a player, i, would not want to deviate to $s_i \leq \alpha_i s_j$). We obtain the utility of player i as a function of c, in the candidate equilibrium by substituting $\lambda_i = \lambda_j = \lambda_c^+$ and $s_i = s_j = 1$ and the equation for second-stage effort levels as a function of first-stage messages (Eq. (5)) into the utility function. Simplifying the resulting expression yields[27]

$$U_i(c) = \frac{c(\sqrt{5-4c}-1)+2}{4}. \tag{67}$$

We obtain the utility of player i as a function of c in the case where he deviates to $s_i < \alpha_i$ by substituting $s_j = 1$ and the equations for second-stage effort levels as a function of s_j (Eq. (62) and Eq. (63)) into the utility function. Simplifying the resulting expression yields[28]

$$U_i(c) = \frac{c(\sqrt{5-4c}-2c+3)}{4}. \tag{68}$$

Therefore, the deviation is not profitable if

$$\frac{c(\sqrt{5-4c}-1)+2}{4} > \frac{c(\sqrt{5-4c}-2c+3)}{4}$$
$$\Longleftrightarrow 2-c > 3c-2c^2 \Longleftrightarrow 2+2c^2 > 4c \Longleftrightarrow 1+c^2 > 2c.$$

This holds for all $c \in (1, 1.25)$, and thus with reneging aversion $(\lambda_c^+, \lambda_c^+)$ there exists a subgame-perfect equilibrium with one-sided reneging costs in which $s_i = s_j = 1$ and $s_i > x_i$ and $s_j > x_j$; i.e., there exists an equilibrium with messages and effort levels that are the same as those in the standard game with two-sided costs. The foregoing reasoning also implies that none of the candidate equilibria in which one player reneges upward are subgame-perfect equilibria. To see this, note first that, for any candidate equilibrium with (s_i, s_j) such that $s_i < \alpha_i s_j < s_j \in (0, 1]$, the utility is equal to the expression in Eq. (68) multiplied by s_j^2. Note also that utility from deviating to $s_i = s_j$ is equal to the expression in Eq. (67) multiplied by s_j^2. Therefore, player i will deviate from such an equilibrium. Therefore the subgame-perfect equilibrium with levels of reneging aversion $(\lambda_c^+, \lambda_c^+)$ is unique and is such that $s_i = s_j = 1$ and $s_i > x_i$ and $s_j > x_j$ and the material payoffs are the same as in the standard game with two-sided reneging costs and hence $\pi_c^1(\lambda_c^+, \lambda_c^+) = \pi_c(\lambda_c^+, \lambda_c^+)$. This completes the proof of parts 1, 3, and 5 of the proposition.

We now demonstrate that $(\lambda_c^+, \lambda_c^+)$ remains a strict Nash equilibrium of the population game under one-sided reneging costs (part 4 of the proposition). We established that for any (λ', λ_c^+) with $\lambda' \neq \lambda_c^+$ the only possible equilibria of the one-sided partnership game involve either both players reneging downward, or exactly one player reneging upward. In the former case, the effort levels are the same as in the corresponding equilibria of the two-sided game; thus, since $\pi(\lambda_c^+, \lambda_c^+) > \pi(\lambda', \lambda_c^+)$, we have that $\pi_c^1(\lambda_c^+, \lambda_c^+) > \pi_c^1(\lambda', \lambda_c^+)$. In the case where one player, i, reneges upward, the equilibrium effort levels are of the form given by Eq. (62) and Eq. (63). Note that if any candidate equilibrium yields positive payoff to a player then the highest possible payoff for that player is obtained in the candidate equilibrium where $s_j = 1$. By inspection of Eq. (62) and Eq. (63) we see that the equilibrium efforts when $s_j = 1$

[27] This simplification was obtained using Mathematica. Code available in the supplementary appendix.

[28] This simplification was obtained using Mathematica. The code is available in the supplementary appendix of this paper.

are the same as those in the equilibrium of the two-sided game where $\lambda_i = 0$; i.e., if (λ_i, λ_j) induces an equilibrium in which player i reneges upward, then $\pi_c^1(\lambda_i, \lambda_j) = \pi_c(0, \lambda_j)$. We know from Theorem 2 that $\pi_c(\lambda_c^+, \lambda_c^+) > \pi_c(0, \lambda_c^+)$. Therefore, for any (λ_i, λ_c^+) that induces an equilibrium in which player i reneges upward in the one-sided game, we have that $\pi_c^1(\lambda_i, \lambda_c^+) = \pi_c(0, \lambda_c^+) < \pi_c(\lambda_c^+, \lambda_c^+) = \pi_c^1(\lambda_c^+, \lambda_c^+)$. Finally, consider any (λ_i, λ_c^+) that induces player j to renege upward. In this case, i's payoff in the most profitable possible equilibrium is obtained by imposing $s_j = 1$ on Eq. (62) and Eq. (63) and substituting the resulting expressions into the expression for the material payoff (Eq. (1) yields

$$\left[\frac{\lambda_i}{c + \lambda_i - \frac{1}{c}}\right]^2 (\frac{1}{c} - \frac{c}{2}) = \frac{(c^2 - 2)c\lambda_i^2}{2\left(c(c + \lambda_i) - 1\right)^2}. \tag{69}$$

A deviation from $(\lambda_c^+, \lambda_c^+)$ that induces such an equilibrium gains a weakly greater payoff only if

$$\frac{(c^2 - 2)c\lambda_i^2}{2\left(c(c + \lambda_i) - 1\right)^2} \geq \frac{c(\sqrt{5 - 4c} - 1) + 2}{4},$$

which never holds for $c \in (1, 1.25)$ and[29] $\lambda_i \geq 0$. Therefore, $\pi^1(\lambda_c^+, \lambda_c^+) > \pi^1(\lambda^{'}, \lambda_c^+)$ for all $\lambda^{'} \neq \lambda_c^+$ such that $(\lambda^{'}, \lambda_c^+)$ induces an equilibrium in which one player reneges upward. We have therefore seen that for all $(\lambda^{'}, \lambda_c^+)$, in all possible equilibria of the induced partnership game with one-sided reneging costs, the payoff achieved by a player with reneging aversion $\lambda^{'}$ is strictly lower than the payoff in the unique equilibrium under $(\lambda_c^+, \lambda_c^+)$. Therefore, regardless of the equilibrium selection function underlying $\pi_c^1(\lambda^{'}, \lambda_c^+)$, we have that $\pi_c^1(\lambda_c^+, \lambda_c^+) > \pi_c^1(\lambda^{'}, \lambda_c^+)$ for all $\lambda^{'} \neq \lambda_c^+$.

Finally, we show that if $c < 1.22$, then $\pi_c^1\left(\lambda_c^+, \lambda_c^+\right) > \pi_c^1\left(\lambda^{'}, \lambda^{'}\right)$ for all $\lambda^{'} \neq \lambda_c^+$ (point 2 of the proposition). If $\lambda_i = \lambda_j = \lambda^{'}$ then $min\{R_i, R_j\} \geq 1$ or $R_i \cdot R_j < 1$. In the former case, the payoffs are the same as in the game with two-sided reneging costs, which by point 1 of Theorem 2 are less than $\pi_c\left(\lambda_c^+, \lambda_c^+\right)$. In the latter case, there are three possibilities: (1) the payoffs are the same as in the game with two-sided reneging costs (and both players' payoffs are zero and therefore less than $\pi_c\left(\lambda_c^+, \lambda_c^+\right)$); (2) there is a symmetric pair of continua of equilibria in which one player reneges upward. In one $s_i = 1 > \alpha_j \geq s_j$ and in the other $s_j = 1 > \alpha_i \geq s_i$; (3) there exist two continua of equilibria of the same form as in (2), plus a third equilibrium in which effort levels are zero. If equilibria in which one player reneges upward exist, these must yield positive payoffs to both players. We show that, whatever equilibrium selection function is assumed, $\pi_c^1\left(\lambda_c^+, \lambda_c^+\right) > \pi_c^1\left(\lambda^{'}, \lambda^{'}\right)$. We do this by considering the equilibrium selection functions that yield the highest possible expected payoff. This is any function putting full weight on the two equilibria where one player reneges upward (as the payoffs in the third possible equilibrium are zero for both players). In Section 7 we make the assumption that in cases with symmetric levels of reneging aversion, if an asymmetric equilibrium is selected, $\pi_c^1\left(\lambda^{'}, \lambda^{'}\right)$ is the average of the equilibrium payoffs of players in the two roles (denoted by i and j). To obtain the payoff function $\pi_c^1\left(\lambda^{'}, \lambda^{'}\right)$ we therefore impose $s_j = 1$ on Eq. (62) and Eq. (63) and substitute the resulting expressions into Eq. (1), and its equivalent for player j, to give the payoffs for the two roles. We then take the

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

average to yield the payoff:

$$\pi_c^1\left(\lambda', \lambda'\right) = \frac{c\alpha_i^2}{2} + \frac{c\alpha_i^2}{2} - \frac{c\alpha_i^2}{4} - \frac{c^3\alpha_i^2}{4} = c\alpha_i^2 - \frac{(c+c^3)\alpha_i^2}{4} = \frac{(3c-c^3)\alpha_i^2}{4}.$$

We next note that α_i is an increasing function of λ' and hence the payoff function is increasing in λ'. We therefore consider the limit of the payoff as $\lambda' \to \infty$. This is given by

$$\lim_{\lambda' \to \infty} \frac{(3c-c^3)\alpha_i^2}{4} = \frac{(3c-c^3)}{4} \lim_{\lambda' \to \infty} \left[\frac{\lambda'}{c^2 + c\lambda' - 1}\right]^2 = \frac{(3c-c^3)}{4}[\frac{1}{c}]^2 = \frac{3-c^2}{4c}.$$

This is the same payoff that obtained under sequential effort choices (with no reneging costs), as derived in point 4 of Theorem 2. As shown in Theorem 2, this is strictly less than the payoff $\pi_c\left(\lambda_c^+, \lambda_c^+\right)$ if $c < 1.22$, yielding the result. □

E.9 Proof of Proposition 4

Fix $c \in (1,2)$. Let $0 < \beta_c^+ \equiv \frac{1}{2\cdot c} + \frac{c}{2} - 1 = \frac{1}{2} \cdot \left(\frac{1}{c} + c\right) - 1$. Consider the partnership game with fixed reneging costs $\beta_i = \beta_j = \beta_c^+$. For each player i, let $x_i^* : [0,1]^2 \to [0,1]$ be a (pure) second-stage strategy that satisfies: (1) $x_i^*(1,1) = 1$ (i.e., a player exerts maximal effort if both players promise maximal effort), and (2) for each $(s_i, s_j) \neq (1,1)$, define x^* in an arbitrary way such that for each pair of messages (s_i, s_j), the effort $x_i^*(s_i, s_j)$ is a best reply to the effort $x_i^*(s_j, s_i)$.

In what follows we show that $\left(1, 1, x_i^*, x_j^*\right)$ is a trembling-hand perfect equilibrium. We begin by showing that both players exerting maximal effort constitutes a second-stage Nash equilibrium of the subgame following the promises $s_i = s_j = 1$. Assume that the opponent exerts maximal effort in this subgame. If the player exerts maximal effort his payoff is equal to $U_i(1,1,1,c) = 1 - 0.5 \cdot c$. Conditional on exerting a nonmaximal effort (and reneging on the agent's promise), the payoff of the agent is maximised when exerting an effort of $\frac{1}{c}$ (by analogous arguments to those presented in Section 3.2), and it is equal to

$$U_i\left(\frac{1}{c}, 1, 1, c\right) = \frac{1}{c} - \frac{1}{2\cdot c} - \beta_c^+ = \frac{1}{2\cdot c} - \left(\frac{1}{2\cdot c} + \frac{c}{2} - 1\right) = 1 - \frac{c}{2} = U_i(1,1,1,c).$$

Thus, the agent obtains his maximal payoff by exerting maximal effort in this subgame.

Next, we show that in any subgame in which the agent (player i) has promised maximal effort, while the opponent (player j) has promised less than maximal effort, $s_j < 1$, the agent's exerted effort is non-maximal and equal to $\frac{1}{c}$ times the opponent's effort in any second-stage Nash equilibrium of the induced subgame. In order to see this, observe first that the opponent (player j) will never exert effort x_j strictly higher than $\max\left(s_j, \frac{1}{c}\right) < 1$ in any Nash equilibrium of this subgame because a strictly higher effort $x_j > \max\left(s_j, \frac{1}{c}\right) < 1$ yields the agent a suboptimal subjective payoff which is equal to a non-optimal material payoff minus the reneging cost. Thus, $x_j < 1$ in any Nash equilibrium of the subgame following messages $(1, s_j < 1)$. If the agent keeps his promise and exerts a maximal effort his payoff is equal to $U_i(1, x_j, 1, c) = x_j - 0.5 \cdot c$. Conditional on reneging on his promise, the agent's best reply is to exert an effort of $\frac{1}{c} \cdot x_j$, which yields a payoff of

$$U_i\left(\frac{1}{c}\cdot x_j, x_j, 1, c\right) = \frac{1}{c}\cdot x_j^2 - \frac{x_j^2}{2\cdot c} - \beta_c^+ = \frac{x_j^2}{2\cdot c} - \left(\frac{1}{2\cdot c} + \frac{c}{2} - 1\right) = 1 - \frac{c}{2} - \frac{1-x_j^2}{2\cdot c}.$$

Observe that the difference in the payoffs, $U_i\left(\frac{1}{c}\cdot x_j, x_j, 1, c\right) - U_i\left(1, x_j, 1, c\right)$, is equal to

$$\begin{aligned} U_i\left(\frac{1}{c}\cdot x_j, x_j, 1, c\right) - U_i\left(1, x_j, 1, c\right) &= 1 - \frac{c}{2} - \frac{1-x_j^2}{2\cdot c} - \left(x_j - 0.5\cdot c\right) \\ &= 1 - x_j - \frac{\left(1-x_j\right)\left(1+x_j\right)}{2\cdot c} = \left(1-x_j\right)\cdot\left(1 - \frac{1+x_j}{2\cdot c}\right) > 0, \end{aligned}$$

where the latter inequality is due to $1 + x_j < 1 + 1 < 2 < 2 \cdot c$. This implies that the agent exerts an effort of $\frac{1}{c} \cdot x_j < x_j$ in any Nash equilibrium an induced subgame following a promise of less than maximal effort by the opponent.

Next, observe that the opponent's (player j's) payoff in any Nash equilibrium of the induced subgame following a promise of less than maximal effort by the opponent and a promise of maximal effort by the player (player i) is equal to

$$U_j\left(x_j, \frac{1}{c}\cdot x_j, s_j, c\right) \leq \pi_i\left(x_j, \frac{1}{c}\cdot x_j, c\right) = \frac{1}{c}\cdot x_j^2 - \frac{c\cdot x_j^2}{2} < 1 - 0.5\cdot c = U_j\left(1,1,1,c\right), \tag{70}$$

which implies that the first-stage best reply of the opponent to the agent's promise of maximal effort is to promise maximal effort as well. This shows that $\left(\left(1,1\right), \left(x_i^*, x_j^*\right)\right)$ is a subgame-perfect equilibrium of the partnership game with fixed reneging costs $\beta_i = \beta_j = \beta_c^+$. Moreover, observe that Eq. (70) implies that promising maximal effort is the unique best reply to an agent who promises maximal effort with a sufficiently high probability (yet, strictly below one), which implies that promising maximal effort remains the unique best reply also to an agent who plays a slightly perturbed strategy by playing a full-support strategy that assigns a high probability to the maximal message in the first stage, which implies that $\left(1, x_i^*, 1, x_j^*\right)$ is a subgame-perfect equilibrium of the partnership game with fixed reneging costs $\beta_i = \beta_j = \beta_c^+$.

E.10 Proof of Proposition 5

We first present two lemmas used in the proof of Proposition 5 before presenting the main proof itself.

Lemma 7. *Define* $\max\left(BR_\pi\left(x_j\right)\right) \equiv max\left(argmax_{x_i\in[a,b]}\left(\pi(x_i, x_j)\right)\right)$. *Then (1)* $\max\left(BR_\pi\left(x_j\right)\right) < x_j$ *for each* $x_j > \bar{x}$ *and (2)* $\max\left(BR_\pi\left(x_j\right)\right) \leq \bar{x}$ *for each* $x_j \leq \bar{x}$.

Proof. Let $x_j \in [0,1]$. We have to show that (1) $\max\left(BR_\pi\left(x_j\right)\right) < x_j$ if $x_j > \bar{x}$, and (2) $\max\left(BR_\pi\left(x_j\right)\right) \leq \bar{x}$ if $x_j \leq \bar{x}$. Assume first that $x_j \in [0, \bar{x}]$. The fact that $\bar{x} \in BR_\pi\left(\bar{x}\right)$ and the strategic complementarity imply that $\max\left(BR_\pi\left(x_j\right)\right) \leq \bar{x}$. We are left with the case $x_j > \bar{x}$. Assume to the contrary that $\max\left(BR_\pi\left(x_j\right)\right) \geq x_j$. Consider the restricted game in which each agent is restricted to choose a strategy in $[x_j, 1]$. This restricted game admits a symmetric Nash equilibrium (x', x'). The strategic complementarity and $\max\left(BR_\pi\left(x_j\right)\right) \geq x_j$ imply that (x', x') is also a Nash equilibrium of the unrestricted game, and we get a contradiction to $\bar{x}$ being the highest equilibrium strategy $\square$

Lemma 8. *For each $\hat{x} > \bar{x}$, there exists $\epsilon > 0$ such that for each $x_i \geq \hat{x}$ and for each $x_j \leq x_i$, there exists $x_i' \leq x_i$ such that $\pi(x_i', x_j) > \pi(x_i, x_j) + \epsilon$.*

Proof. Fix $\hat{x} > \bar{x}$. For each $x_i \geq \hat{x}$ and each $x_j \leq x_i$ define

$$f(x_i, x_j) = \max_{x_i' \in [0, x_i]} (\pi_i(x_i', x_j) - \pi_i(x_i, x_j)).$$

The fact that $\pi(x_i, x_j)$ is continuously differentiable implies that $f(x_i, x_j)$ is continuous in both parameters. Lemma 7 implies that $\max(BR_\pi(x_j)) \leq \bar{x}$ if $x_j \leq c$, and $\max(BR_\pi(x_j)) < x_j$ if $x_j > \bar{x}$. These inequalities, in turn, imply that $f(x_i, x_j) > 0$ for each $x_i \geq \hat{x}$ and each $x_j \leq x_i$. Define

$$\tilde{\epsilon} = \min_{x_i \in [\hat{x}, 1],\, x_j \in [0, x_i]} f(x_i, x_j).$$

The compactness of the set $\left\{(x_i, x_j) \in [0,1]^2 \,|x_i \in [\hat{x}, 1], x_j \in [a, x_i]\right\}$ and the continuity of $f(x_i, x_j)$ imply that $\tilde{\epsilon} > 0$. Fix $x_i \in [\hat{x}, 1]$ and $x_j \in [0, x_i]$. Let $x_i' \in BR_\pi(x_j)$. Let $\epsilon = \frac{\tilde{\epsilon}}{2}$. Then the definition of $\tilde{\epsilon}$ implies that

$$\pi(x_i', x_j) - \pi(x_i, x_j) \geq \tilde{\epsilon} > \epsilon,$$

which proves the lemma. $\square$

E.10.1 Proof of Proposition 5

Proof. We first prove point (1) (namely, that the equilibrium efforts are at most $\bar{x} + \epsilon$ if $\lambda_i, \lambda_j < \underline{\lambda}_\epsilon$ for some $\underline{\lambda}_\epsilon > 0$). Fix $\epsilon > 0$. Lemma 8 implies that there exists $\delta > 0$ such that for each $x_i \geq \bar{x} + \epsilon$ and each $x_j \leq x_i$ there exists $x_i' \leq x_i$ such that $\pi(x_i', x_j) > \pi(x_i, x_j) + \delta$. Let $\underline{\lambda}_\epsilon$ be sufficiently small such that $\underline{\lambda}_\epsilon \cdot D < \frac{\delta}{2}$. Assume that there is a pure subgame-perfect equilibrium $(\overrightarrow{s}^*, \overrightarrow{x}^*)$ of the partnership game with levels of reneging aversion $\lambda_i, \lambda_j \leq \underline{\lambda}_\epsilon$ in which agent i exerts effort of at least $\bar{x} + \epsilon$, i.e., $x_i^*(\overrightarrow{s}^*) \geq \bar{x} + \epsilon$. Assume without loss of generality that $x_i^*(\overrightarrow{s}^*) \geq x_j^*(\overrightarrow{s}^*)$. Lemma 8 implies that there exists x_i' satisfying

$$\pi(x_i', x_j^*(\overrightarrow{s}^*)) > \pi(x_i^*(\overrightarrow{s}^*), x_j^*(\overrightarrow{s}^*)) + \delta,$$

which implies that

$$U(x_i', x_j^*(\overrightarrow{s}^*), s_i^*, \lambda_i) > \pi(x_i^*(\overrightarrow{s}^*), x_j^*(\overrightarrow{s}^*)) + \delta - \frac{\delta}{2} > U(x_i^*(\overrightarrow{s}^*), x_j^*(\overrightarrow{s}^*), s_i^*, \lambda_i),$$

where the first inequality is due to $\underline{\lambda}_\epsilon \cdot D < \frac{\delta}{2}$. Thus, we get a contradiction to $x_i^*(\overrightarrow{s}^*)$ being a second-stage best-reply against $x_j^*(\overrightarrow{s}^*)$.

We now prove point (2) (namely, that the equilibrium efforts are at most $\bar{x} + \epsilon$ if $\lambda_i, \lambda_j > \overline{\lambda}_\epsilon$). The fact that π is twice continuously differentiable implies that it is Lipschitz continues, i.e., that there exists $M > 0$ such that for each $(x_i, x_j), \left(x_i', x_j'\right) \in [0,1]^2$,

$$\left|\pi(x_i, x_j) - \pi\left(x_i', x_j'\right)\right| < M \cdot \left(|x_i - x_i'| + \left|x_j - x_j'\right|\right).$$

Let $\epsilon' > 0$ be sufficiently small such that $M \cdot \epsilon' < \frac{\delta}{4}$ and $\epsilon' < \frac{\epsilon}{8}$. Let $\overline{\lambda}_\epsilon > 0$ be sufficiently large such that $\overline{\lambda}_\epsilon \cdot D \cdot \epsilon' > 2 \cdot M$. Assume that there is a pure subgame-perfect equilibrium $(\overrightarrow{s}^*, \overrightarrow{x}^*(\overrightarrow{s}))$ of the partnership game with levels of reneging aversion $\lambda_i, \lambda_j \geq \overline{\lambda}_\epsilon$ in which agent i exerts effort of at least $\bar{x} + \epsilon$, i.e., $x_i^*(\overrightarrow{s}^*) \geq \bar{x} + \epsilon$. Assume without loss of generality that $x_i^*(\overrightarrow{s}^*) \geq x_j^*(\overrightarrow{s}^*)$. Lemma 8 c implies that there exists $\delta > 0$ such that for each $x_i^* \geq \bar{x} + \epsilon$ and each $x_j^* \leq x_i^*$ there exists $x_i' \leq x_i^*$ such that $\pi\left(x_i', x_j^*\right) > \pi\left(x_i^*, x_j^*\right) + \delta$. The fact that $\overline{\lambda}_\epsilon \cdot D \cdot \epsilon' > 2 \cdot M$ implies that $x_j^*\left(s_i', s_j^*\right) \geq s_j^* - \epsilon'$ for each $s_i' \in [0,1]$. This, in turn, implies that $\left|x_j^*\left(s_i', s_j^*\right) - x_j^*(s^*)\right| \leq 2 \cdot \epsilon'$ for each $s_i' \in [0,1]$.

Consider the deviation of player i to promising x_i' in the first round and exerting effort x_i' in the second round. We complete the proof by showing that this deviation induces a higher payoff to the deviator relative to the equilibrium behaviour (which contradicts $(\overrightarrow{s}^*, \overrightarrow{x}^*(\overrightarrow{s}))$ being a subgame-perfect equilibrium):

$$U\left((x_i', x_j^*\left(s_i = x_i', s_j^*\right), s_i^*, \lambda_i)\right) = \pi(x_i', x_j^*\left(s_i = x_i', s_j^*\right)) \geq \pi(x_i', x_j^*(s^*)) - M \cdot 2 \cdot \epsilon' \geq \pi(x_i', x_j^*(s^*)) - \frac{\delta}{2}$$

$$\geq \pi(x_i^*, x_j^*(s^*)) - \frac{\delta}{2} + \delta \geq U\left((x_i^*, x_j^*(s^*), s_i^*, \lambda_i)\right) + \frac{\delta}{2}.$$

□