



Munich Personal RePEc Archive

# **Promises and Endogenous Reneging Costs**

Heller, Yuval and Sturrock, David

Bar Ilan University, Institute for Fiscal Studies

April 2017

Online at <https://mpra.ub.uni-muenchen.de/90249/>  
MPRA Paper No. 90249, posted 29 Nov 2018 08:08 UTC

# Promises and Endogenous Reneging Costs\*

Yuval Heller<sup>†</sup>

David Sturrock<sup>‡</sup>

November 25, 2018

## Abstract

We present a novel theoretical 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 renegeing on their promises. We demonstrate a strong tendency for evolutionary processes to select agents who incur intermediate costs of renegeing, and show that these intermediate 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 (renegeing). Second, we demonstrate that these endogenously determined intermediate psychological costs yield second-best optimal outcomes

---

\*We thank two anonymous referees, Vincent Crawford, Rachel Griffith, Nick Netzer, Alon Raviv, Nils Rochowicz, Peyton Young, 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.

<sup>†</sup>Department of Economics, Bar Ilan University. [yuval.heller@biu.ac.il](mailto: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).

<sup>‡</sup>University College London and Institute for Fiscal Studies, [david.sturrock.17@ucl.ac.uk](mailto: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.

## General Partnership Game

We study a general class of partnership games (also known as input games; see, e.g., [Holmstrom, 1982](#); [Cooper & John, 1988](#)) with cheap talk preplay communication. In the setting we examine, agents simultaneously communicate promised levels of effort, and, following this, they simultaneously choose their levels of effort. We make two mild assumptions about the material payoffs of the agents: (1) the payoff of each agent is weakly increasing in his partner’s effort, and (2) the payoff function encourages “shirking” above some effort level  $c$ : the best reply to his partner exerting effort  $x > c$  is for the agent to exert less effort than his partner. Lemma 1 shows that any game with strategic complements encourages shirking above the highest symmetric Nash equilibrium effort level (see, e.g., Example 1 of price competition with differentiated goods).

We explore the impact of introducing into this setting a direct cost of renegeing on promises. Specifically, we consider an arbitrary renegeing cost function that satisfies the following two mild assumptions: (1) the renegeing cost is weakly increasing in the difference between the promised level of effort in the first stage and the exerted level of effort in the second stage, and (2) any nonzero difference between promised and exerted effort induces a positive cost. Next, we endow each agent  $i$  with a level of renegeing aversion  $\lambda_i \geq 0$  (which is observed by his partner). The subjective payoff of each agent  $i$  is equal to the material payoff (which depends on the levels of exerted effort) minus  $\lambda_i$  times the renegeing cost.

## Nonintermediate Costs Induce Zero Effort

Our first result (Proposition 1) shows that agents exert low effort of at most  $c$  in any equilibrium of the partnership game whenever the renegeing costs are either too low or too high. The intuition is as follows. Too low renegeing costs induce too little commitment power and, as a result, each agent undercuts his partner’s effort in the second round, regardless of the promise. Too high renegeing costs do not leave enough flexibility for the second round, making agents unwilling to promise effort. Specifically, each agent in the second round exerts a level of effort very close to his own promise, regardless of the partner’s promise. This, in turn, implies that each agent undercuts his partner’s promise in the first round, which implies that the agents promise effort of at most  $c$  in the first round, and exert effort of at most  $c$  in the second round. Thus, only when both agents have intermediate levels of renegeing costs it is possible to induce effort above  $c$  and to have a successful partnership.

## Partnership Game with Quadratic Payoffs

For tractability, we focus in the subsequent analysis on a specific family of quadratic payoff functions that satisfy: (I) a convex material cost of exerting effort, (II) a convex intrinsic cost of promising an effort different from the exerted effort, and (III) strategic complementarity between the effort levels of the two players. Specifically, we assume that (1) the material payoff of each agent is equal to the product of the two efforts minus a cost that is proportional to the square of the agent's exerted effort, and (2) the renegeing cost is proportional to the square of the difference between the promised effort and the exerted effort. We further assume that the set of feasible effort levels is a bounded interval.

In Theorem 1 we fully characterise the set of perfect equilibria of the partnership game. It turns out that the partnership game, essentially, admits a unique perfect equilibrium, and that the properties of this unique equilibrium depend on which of three regions the pair of the players' levels of renegeing aversion belongs to (as demonstrated in Figure 2 in Section 4.4).

1. There is a convex symmetric region where both players have an intermediate level of renegeing aversion and the unique equilibrium is for both players to promise maximal effort (*maximum-message equilibrium*); in the second stage both players exert positive levels of effort. The intuition is as follows. The indirect benefit of promising a higher level of effort than his partner does ("overcutting," which induces his partner to exert more effort in the second stage) is increasing in the player's renegeing aversion (as his promise is more credible), and decreasing in the partner's renegeing aversion (as a higher level of renegeing aversion gives the partner less flexibility to respond to the agent's promise). Thus, the indirect benefit is sufficiently large to induce an agent to overcut his partner's promise if and only if (1) the agent's level of renegeing aversion is sufficiently high, and (2) the partner's level of renegeing aversion is sufficiently low. This implies that both agents are induced to overcut each other (which implies that both promise the maximal effort) if and only if both players' levels of renegeing aversion are neither too low nor too high.
2. There are two disjoint areas in which the partnership game admits the *no-effort equilibrium*: (I) an area in which both players have sufficiently high levels of renegeing aversion, and (II) an area in which both players have sufficiently low levels of renegeing aversion. The intuition is the same as for Proposition 1, discussed above.
3. The remaining region is divided into two disjoint areas in which (I) one player has a sufficiently high level of renegeing aversion and he promises the maximal level of effort, and (II) his partner has a sufficiently low level of renegeing aversion and promises a positive nonmaximal level of effort (*two-message equilibrium*). The intuition is that only the player with the high level of renegeing aversion has substantial commitment power, while his partner's promise has a very small impact on either player's choice of effort. As a result, the agent with high renegeing aversion 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.

## Appealing Properties of Intermediate Reneging Aversion

Let  $\lambda_c^+$  be the maximal level of reneging aversion for which the perfect equilibrium of the game between two players with this level of reneging aversion is a maximum-message equilibrium. Our next result (Theorem 2) shows that the equilibrium induced by both players having this “intermediate” level of reneging aversion  $\lambda_c^+$  has three appealing properties:

1. *“Second-best” outcome:* This equilibrium induces the best equilibrium outcome among all equilibrium outcomes of symmetric partnership games.
2. *“First-best” outcome in the limit of small effort costs:* When the cost of exerting effort converges to zero, the equilibrium payoff converges to the maximal feasible payoff induced by both agents exerting maximal efforts.
3. *Better outcome than Stackelberg equilibrium outcome:* The equilibrium payoff is larger than the mean payoff 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.

## Evolutionary Stability of Intermediate Reneging Aversion

We study the endogenous determination of players’ levels of reneging aversion in an evolutionary framework. 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. We assume that in each such partnership game, the players play the unique perfect equilibrium of the game.

Our final main result shows that the homogeneous population state in which all agents have the same intermediate level of reneging aversion  $\lambda_c^+$  is evolutionarily stable. Moreover, there does not exist any other homogeneous stable state. This demonstrates the strong tendency of evolutionary processes to select for agents who incur intermediate psychological reneging costs.

## Variants and Extensions

We demonstrate the robustness of our results by showing that our main results hold also when some of the key assumptions of our model are relaxed. Specifically, we show that the appealing properties of the equilibrium induced by  $\lambda_c^+$ , and the evolutionary stability of the homogeneous population state in which agents have a reneging aversion of  $\lambda_c^+$ , hold in the following three variants/extensions of our baseline model:

1. *Sequential communication:* A variant of the model in which agents make promises sequentially, rather than simultaneously. That is, nature chooses one of the players at random, and this player sends his promise first.
2. *One-sided reneging costs:* A variant of the model in which an agent suffers a reneging cost only when his promise is higher than the level of effort he exerts in the second stage. Unlike the baseline model, an agent does not suffer a cost when his promise is lower than his exerted effort.

3. *Partial observability*: We show that our results hold also in a setup in which the partner’s reneging cost is observed with a sufficiently high probability (which is strictly less than one). Moreover, we show that a weaker version of this result holds also when players can observe their partner’s level of reneging aversion with a low, yet positive, probability. In this latter case, we show that in any stable state players must have positive reneging aversion and exert positive effort in equilibrium.

In addition, we show that our qualitative results hold in a variant of the model in which an agent incurs a *fixed reneging cost* whenever the exerted effort is different from that promised, regardless of the size of the difference.

## Structure

The paper is organised as follows. Section 2 discusses the related literature and the contributions made by this paper. Section 3 studies general partnership games. Section 4 analyses partnership games with quadratic payoffs. Section 5 shows the appealing properties and evolutionary stability of intermediate reneging aversion  $\lambda_c^+$ . Section 6 shows the robustness of our main results to the relaxation of various assumptions in our model. We conclude in Section 7. The formal definition of a trembling-hand perfect equilibrium with a continuum of strategies is relegated to Appendix A. We discuss a few technical aspects of our evolutionary interpretation in Appendix B. Additional illustrative figures are presented in online Appendix C. Formal proofs appear in online Appendix D.

## 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 subgame-perfect equilibrium involves no such cooperation (Charness & Dufwenberg, 2006; Vanberg, 2008; Ederer & Stremitzer, 2017; Di Bartolomeo *et al.*, 2018). We advance the theoretical analysis of preplay communication by presenting a novel mechanism (intermediate reneging costs) by which 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 in which the informed agent has an incentive to distort the receiver’s belief, and incurs a convex cost of sending a false message. Matsushima (2008) and Kartik *et al.* (2014) introduce arbitrarily small lying costs into settings of mechanism design and implementation. The present paper moves beyond the existing literature 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 renegeing costs, and allow them to be determined as part of a stable population state.<sup>1</sup>

We contribute to the literature on partnerships with strategic complementarities by introducing renegeing costs into a general class of partnership games. Games in which  $n$  players experience a common outcome, which is increasing in a privately costly action, are examined from a mechanism design perspective in [Holmstrom \(1982\)](#). [Radner et al. \(1986\)](#) analyse a two-player partnership game in which a project succeeds with a probability equal to the minimum of the players’ effort choices, which are made at quadratic cost, and show the capacity of repeated interaction to sustain effort when such an outcome is efficient but is not an equilibrium of the one-shot game (see also related models of partnership games in [Cooper & John, 1988](#); [?; Cahuc & Kempf, 1997](#); [Marx & Matthews, 2000](#)). We demonstrate that renegeing 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 is able to rationalise and ground the main stylised facts of the related experimental literature. Intrinsic costs of lying or renegeing 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. , Forthcoming](#)). Experimental evidence suggests that subjects do not always lie to gain money, even when their doing so cannot be detected.<sup>2</sup> In promising experiments, 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 main stylised facts from these experiments suggest that the intrinsic costs of lying/renegeing are intermediate, and are increasing (potentially convexly) with one or more of the following factors: (I) the difference between the reported/exerted outcome and the true/promised outcome, (II) the damage induced to the partner by an agent lying/renegeing, and (III) others’ perceptions of the agent’s behaviour. In our model the intrinsic cost of renegeing is proportional to the difference between the promised effort and the exerted effort, which directly captures factor (I). In a richer model, in which others observe the exerted effort with some random noise, this difference can also capture factor (III). In Section 6.2, we study a variant of our model in which an intrinsic cost of renegeing is incurred only if the promised effort is smaller than the exerted effort. This captures factor (II), as in this variant

---

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

<sup>2</sup>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.

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 renegeing aversion to be endogenously determined by an evolutionary process, providing a theoretical foundation and explanation for these stylised experimental facts.

In our theoretical exploration of the potential evolutionary determinants of renegeing 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 renegeing costs. Second, our main result is qualitatively different from the stylised result in the existing literature, according to which if preferences are observed with high probability, then the Pareto-efficient outcome is played in any stable population state. We show that in the setup in which the set of feasible preferences is the set of levels of renegeing aversion, evolutionary forces take the population into stable states in which agents have intermediate renegeing aversion and the agents achieve partial, rather than full, efficiency.

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

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 renegeing 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 a broad class of partnership games with renegeing costs, and show that if both agents have either too low or too high renegeing costs, essentially no efforts are exerted in the game. By contrast, in the next section we will show how intermediate renegeing costs induce the second-best outcome.

#### 3.1 Stages and Strategies

There are two players ( $i$  and  $j$ ) and two stages (or rounds) of the game. In the first stage, both players simultaneously send a message  $s_k \in [a, b]$  to their partner (where  $k = i, j$ , and  $a < b$ ). 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,  $x_k \in [a, b]$  (after observing the promises made in the first stage).

A (pure) strategy of an agent is a pair  $(s_i^*, x_i^*)$ , where  $s_i^* \in [a, b]$  is the agent's promise in the first round, and  $x_i^* : [0, 1]^2 \rightarrow [0, 1]$  is a function describing the agent's effort as a function of the pair of promises sent in the first round by the agent and his partner.

#### 3.2 The Material Payoffs

The "material" payoffs of the agents depend only on the levels of effort they exert in the second round. Let  $\pi(x_i, x_j)$  be the material payoff of an agent exerting effort  $x_i \in [a, b]$ , given that his partner exerted effort  $x_j \in [a, b]$ . We make three assumptions about the function  $\pi$ :

1. *Lipschitz continuity*: We assume that  $\pi(x_i, x_j)$  is Lipschitz continuous in both variables, i.e., that there exists  $M > 0$  such that for each  $(x_i, x_j), (x'_i, x'_j) \in [a, b]^2$ ,

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

2. *Positive externalities*: We assume that  $\pi(x_i, x_j)$  is weakly increasing in its second argument, i.e., that  $x'_j \geq x_j$  implies that  $\pi(x_i, x'_j) \geq \pi(x_i, x_j)$ . That is, an agent weakly gains from his partner exerting more effort.
3. *Encouraging shirking* above some effort level  $c$ : Our last assumption regarding  $\pi$  is that there exists an effort level  $c \in [a, b]$ , such that for each strategy of the agent  $x_i > c$ , and for each strategy of the partner  $x_j \leq x_i$ , the agent strictly gains by exerting less effort than  $x_i$ , and, moreover, this increase in payoff can be bounded away from zero. Formally:

**Definition 1.** Let  $\pi : [a, b]^2 \rightarrow \mathbb{R}$  be a payoff function and let  $c \in [a, b]$ . We say that the function  $\pi$  *encourages shirking above  $c$*  if for each  $\underline{x} > c$ , there exists  $\epsilon > 0$  such that for each  $x_i \geq \underline{x}$  and for each  $x_j \leq x_i$ , there exists  $x'_i \leq x_i$  such that  $\pi(x'_i, x_j) > \pi(x_i, x_j) + \epsilon$ .

The following lemma shows that when the payoff function is differentiable, then the property of encouraging shirking above  $c$  is implied by having the maximal best reply of each strategy  $x_j$  smaller than

$\max(c, x_j)$ . Formally, define

$$\max(BR_\pi(x_j)) \equiv \operatorname{argmax}_{x_i \in [a, b]} (\pi(x_i, x_j))$$

as the maximal value of  $x_i$  that maximises the material payoff of player  $i$  when player  $j$  plays  $x_j$ . Then (the simple technical proof is presented in Appendix D.1):

**Lemma 1.** *Assume that  $\pi(x_i, x_j)$  is continuously twice differentiable and that (1)  $\max(BR_\pi(x_j)) < x_j$  for each  $x_j > c$ , and (2)  $\max(BR_\pi(x_j)) \leq c$  for each  $x_j \leq c$ . Then  $\pi(x_i, x_j)$  encourages shirking above  $c$ .*

Recall that a game with a continuously twice differentiable payoff function  $\pi(x_i, x_j)$  has *strategic complements* if  $\frac{\partial^2 \pi_i(x_i, x_j)}{\partial x_i \partial x_j} > 0$  for each  $x_i, x_j$ . Given a game with strategic complements, let  $\bar{x}$  be the highest rationalizable strategy. Recall (see, e.g., [Milgrom & Roberts, 1990](#); [Levin, 2003](#)) that  $(\bar{x}, \bar{x})$  is a Nash equilibrium of the game. The following lemma shows that any game with strategic complements encourages shirking above  $\bar{x}$ . Formally:

**Lemma 2.** *Let  $\pi : [a, b]^2 \rightarrow \mathbb{R}$  be a continuously twice differentiable payoff function that satisfies strategic complements. Let  $\bar{x}$  be the highest rationalizable strategy. Then  $\pi$  encourages shirking above  $\bar{x}$ .*

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

Thus, the three mild assumptions presented above are, essentially, satisfied in any game with strategic complements (and positive externalities), such as price competition with differentiated goods, which is presented in Example 1.

**Example 1** (*Price competition with differentiated goods; see a textbook analysis in [Mas-Colell et al., 1995, Section 12.C](#)*). Consider a mass one of consumers equally distributed in the interval  $[0, 1]$ . Consider two firms that produce widgets, located at the two extreme locations: 0 and 1. Every consumer wants at most one widget. Producing a widget has a constant marginal cost, which we normalise to zero. Each firm  $i$  chooses a price  $x_i \in [0, M]$  for its widgets. The total cost of buying a widget from firm  $i$  is equal to its price,  $x_i$ , plus  $t$  times the consumer's distance from the firm, where  $t \in [0, M]$ . Each consumer buys a widget from the firm with the lower total buying cost. This implies that the total demand for good  $i$  is given by function  $q_i(x_i, x_j)$ :

$$q_i(x_i, x_j) = \begin{cases} 0 & x_i > x_j + t \\ \frac{x_j - x_i + t}{2t} & x_i \in [x_j - t, x_j + t] \\ 1 & x_i < x_j - t, \end{cases}$$

The payoff (profit) of firm  $i$  is given by  $\pi_i(x_i, x_j) = x_i \cdot q_i(x_i, x_j)$ . Observe that the payoff function is Lipschitz continuous in both variables, and that the payoff of each agent is weakly increasing in the opponent firm's price (as a larger opponent firm's price increases the demand for the firm's widgets, which in turn, weakly increases the firm's profit). One can show that the best reply function of each player is

$$BR(x_j) = \begin{cases} \frac{x_j+t}{2} & x_j < 3 \cdot t \\ x_j - t & x_j \geq 3 \cdot t. \end{cases}$$

Thus we have that  $BR(x_j) < \max(x_j, t)$ , which implies that the payoff function encourages shirking above  $t$  (which, one can show, is the unique rationalizable strategy).

### 3.3 The Reneging Costs

We assume that each player  $i$  is endowed with a level of reneging aversion  $\lambda_i$ . 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  $D : [0, b - a] \rightarrow R^+$  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  $D(0) = 0$ .

### 3.4 Nonintermediate Costs Induce Zero Effort

Our first result shows that agents do not exert effort above  $c$  in any pure subgame-perfect equilibrium of the partnership game whenever the reneging costs are either too low or too high. The intuition is as follows:

1. Too low reneging costs induce too little commitment power. As a result, no promise to exert effort greater than  $c$  is sufficiently "credible" to induce effort from the agent's partner and each agent will undercut any level of their partner's effort above  $c$  in the second round, regardless of the promises made. As a result, both agents exert effort of at most  $c$ .
2. Too high reneging costs leave too little flexibility for the second round, making agents unwilling to promise effort. Specifically, each agent in the second round exerts a level of effort very close to his own promise, regardless of his partner's promise. This, in turn, implies that each agent undercuts his partner's promise in the first round, which implies that both agents promise effort of at most  $c$  in the first round and exert effort of at most  $c$  in the second round.

**Proposition 1.** *Assume that the payoff function  $\pi : [a, b]^2 \rightarrow \mathbb{R}$  is Lipschitz continuous, weakly increasing in the partner’s effort, and encourages shirking above  $c$ . Then, for any  $\epsilon > 0$ , there exist  $\bar{\lambda}_\epsilon > \underline{\lambda}_\epsilon > 0$ , such that the effort level exerted by any agent in any pure subgame-perfect equilibrium of the partnership game is at most  $c + \epsilon$  if either (1)  $\lambda_i, \lambda_j < \underline{\lambda}_\epsilon$  or (2)  $\lambda_i, \lambda_j > \bar{\lambda}_\epsilon$ .*

*Remark 1.* The result that too high reneging costs induce efforts of at most  $c$  depends on the combination of the following two assumptions: (1) simultaneous communication and (2) “two-sided” reneging costs, in the sense that an agent incurs a reneging cost also when exerting more effort than promised.

In Section 6 we demonstrate (partial) robustness of this result to the relaxation of these assumptions. Specifically, we study games with (1) sequential communication and (2) “one-sided” reneging costs (while focusing on quadratic payoffs), and we demonstrate that when both agents have high reneging costs they may exert efforts above  $c$  in equilibrium, but these efforts are smaller than those induced by intermediate reneging costs.

## 4 Partnership Game with Quadratic Payoffs

### 4.1 Quadratic Payoffs

For tractability, we focus in the remaining analysis 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 (which is homogeneous of degree 2):

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

The interpretation of the quadratic 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.<sup>3</sup>

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

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

The utility loss from reneging on a promise is proportional to the square of the difference between the promised effort and the exerted effort, multiplied by the agent’s level of reneging aversion  $\lambda_i \geq 0$ .

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 3 and Footnote 15.)

---

<sup>3</sup>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  $x_i = x_j = 1$ .

## 4.2 Unique Second-Stage Equilibrium

In the second stage of the game, player  $i$ 's first-order condition for his choice of  $x_i$  is given by<sup>4</sup>

$$x_j - cx_i + \lambda_i(s_i - x_i) = 0. \quad (3)$$

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

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

This equation embodies a player's (possibly conflicting) desires to undercut (exert less effort than) his partner and to minimise his renegeing.

**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 renegeing 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  $s_j$ :

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

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 renegeing. Due to the convex cost of renegeing 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<sup>5</sup>  $s_i, s_j$ , and 0. As a player's level of renegeing aversion increases, he will exert effort closer to his own promise.

## 4.3 First-Stage Best-Reply Functions

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

---

<sup>4</sup>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.

<sup>5</sup>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  $s_j$  in the numerator.

Nash equilibrium in the second stage of the game).

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

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,<sup>6</sup> which implies that, without loss of generality, we can focus on pure strategies (formal details for this argument are presented in the proof of Prop. 2 in Appendix D.4).

Our next 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 cost  $\lambda_i$ ) to a partner's message of  $s_j$ , where the cost of effort is  $c$ . First, we show that if  $c \geq \sqrt{2}$ , then each agent always wants to undercut his partner's message, i.e.,  $s_i^*(s_j|\lambda_i, \lambda_j, c) < s_j$  for each  $s_j > 0$ . The intuition is that such a high value of  $c$  provides too large incentives to undercut the partner. By contrast, if  $c < \sqrt{2}$ , then there exists  $\bar{\lambda}_c > 0$ , such that:

1. If  $\lambda_j > \bar{\lambda}_c$ , then the agent's best reply is to undercut his partner's message. The intuition is that if  $\lambda_j$  is sufficiently high, the partner is going to exert a level of effort close to  $s_j$  regardless of the level of  $s_i$ , and this implies that agent  $i$  is better off if he undercuts his partner's message, as the direct benefit to the agent of being able to undercut his partner in the second stage outweighs the small indirect cost of inducing his partner to exert a little bit less effort in the second stage.
2. If  $\lambda_j < \bar{\lambda}_c$ , then the agent's best reply is to overcut his partner (i.e.,  $s_i^*(s_j|\lambda_i, \lambda_j, c) > s_j$  for each  $s_j < 1$ ), provided that  $\lambda_i$  is sufficiently large (and to undercut his partner otherwise). The intuition is that when the partner's reneging cost is not too large, the indirect benefit of overcutting the partner's message (which induces the partner to exert more effort in the second stage) is increasing in the agent's reneging cost (as his promise is more credible), and for a sufficiently large  $\lambda_i$ , this benefit outweighs the direct cost of restricting the agent's ability to shirk in the second stage.

### Proposition 2.

1. If  $c \geq \sqrt{2}$ , then  $s_i^*(s_j|\lambda_i, \lambda_j, c) < s_j$  for each  $s_j > 0$ ,  $\lambda_i > 0$ , and  $\lambda_j \geq 0$  (undercutting).
2. Otherwise, for each  $c < \sqrt{2}$ , let  $\bar{\lambda}_c \equiv \frac{2-c^2}{c-1} > 0$ .
  - (a) If  $\lambda_j \geq \bar{\lambda}_c$ , then  $s_i^*(s_j|\lambda_i, \lambda_j, c) < s_j$  for each  $s_j > 0$  and  $\lambda_i > 0$  (undercutting).
  - (b) If  $\lambda_j < \bar{\lambda}_c$ , then, there exist  $x(\lambda_j, c) > 0$ , such that:
    - i.  $s_i^*(s_j|\lambda_i, \lambda_j, c) < s_j$  for each  $0 < \lambda_i < x(\lambda_j, c)$  and  $s_j > 0$  (undercutting).
    - ii.  $s_i^*(s_j|\lambda_i, \lambda_j, c) > s_j$  for each  $\lambda_i > x(\lambda_j, c)$  and  $0 < s_j < 1$  (overcutting).

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  $c = 1.2$  (additional effort costs are illustrated in Appendix C).

---

<sup>6</sup>The choice of best reply in these measure-zero cases plays no role in our analysis.

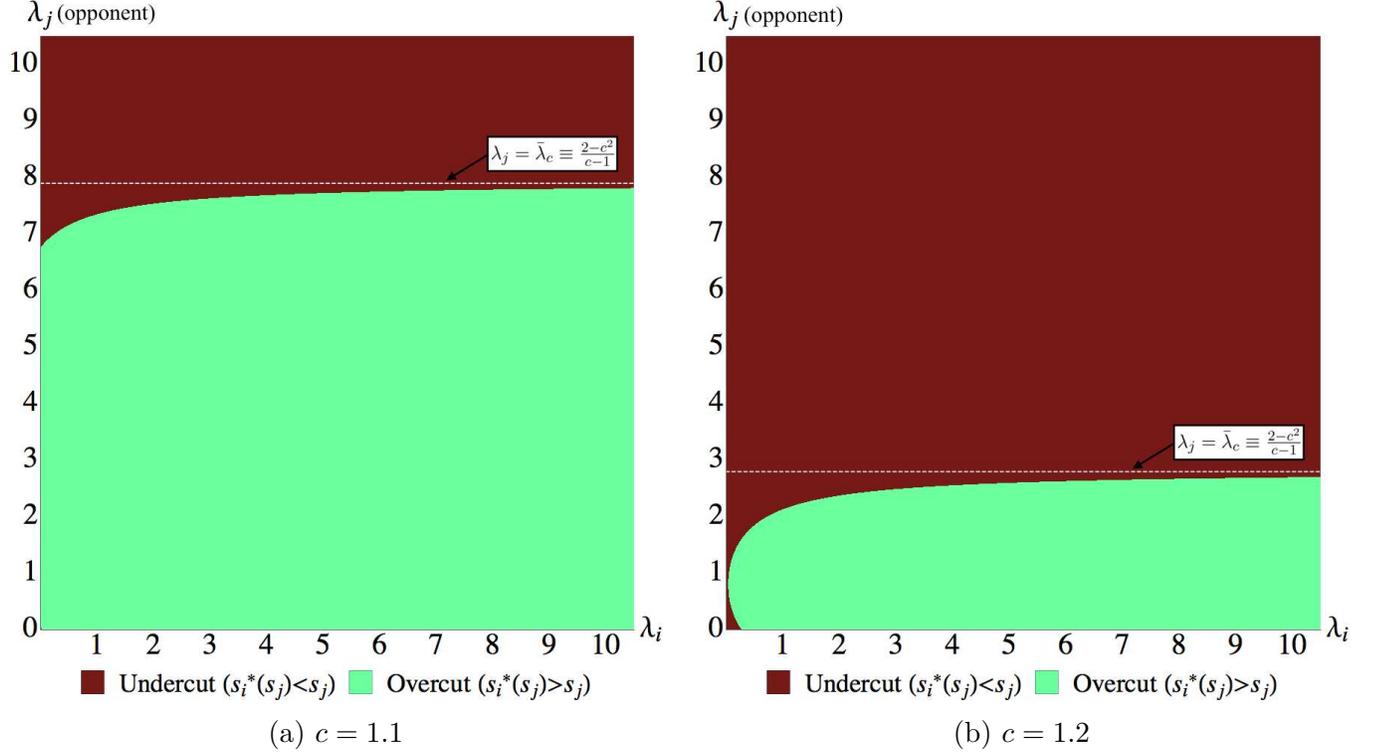


Figure 1: **Best-Reply Types for Player  $i$  in a Reneging Aversion Parameter Space.** The  $x$  axis in each figure presents the player's level of reneging aversion ( $\lambda_i$ ) and the  $y$  axis presents the partner's level of reneging aversion ( $\lambda_j$ ). The left panel deals with a cost 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 2, above which player  $i$ 's best reply is to undercut his partner regardless of the value of  $\lambda_i$ .

#### 4.4 Unique Perfect Equilibrium

We now characterise the subgame-perfect equilibria of the partnership game. Recall that a strategy profile  $\left( (s_i^*, s_j^*), (x_i^*(\vec{s}), x_j^*(\vec{s})) \right)$  is a subgame-perfect equilibrium if for each player  $i$  (1)  $x_i^*(\vec{s}) = x_i^e(s_i, s_j, \lambda_i, \lambda_j, c)$  (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.,  $x_i^*(\vec{s}^*) = x_j^*(\vec{s}^*) = 0$ . In this equilibrium any agent with a positive reneging cost promises nothing, i.e.,  $\lambda_i > 0 \Rightarrow s_i^* = 0$ .
3. *Two-message equilibrium*. In this equilibrium one of the agents 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,  $\epsilon$ , each player trembles and chooses his promise uniformly) induces at least one of the agents to overcut his partner, leading to this equilibrium being eliminated from the perturbed game.

Theorem 1 (below) shows that:

1. If  $c \geq \sqrt{2}$ , then the partnership game admits only a no-effort equilibrium.
2. If  $c \in (1.25, \sqrt{2})$ , then:
  - (a) There is a connected symmetric region in which the partnership games admit the no-effort equilibrium. This region includes all games in which (1) both agents have the same levels of reneging aversion (i.e.,  $\lambda_i = \lambda_j$ ), (2) both agents have sufficiently high levels of reneging aversion, and (3) both agents have sufficiently low levels of reneging aversion.
  - (b) The remaining region is divided into 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 the two-message equilibrium.
3. If  $c \in (1, 1.25)$ , then:
  - (a) There is a convex symmetric region of intermediate levels of reneging aversion in which the game admits a maximum message equilibrium.
  - (b) There are two disjoint areas in which the partnership game admits the no-effort equilibrium: (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.
  - (c) The remaining region is divided into 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 the two-message equilibrium.

Figure 2 illustrates the division of the reneging aversion parameter space into the three classes of unique equilibria for the effort costs of  $c = 1.1$  and  $c = 1.2$  (additional effort costs, namely, 1.05, 1.15, 1.25, and 1.35) are illustrated in Appendix C.

Formally (the definition of trembling-hand perfection is presented in Appendix A):

**Theorem 1.** *For each  $c > 1$ , there exist pairwise disjoint symmetric sets  $\Lambda_{0-eff}^c, \Lambda_{max}^c, \Lambda_{2-msg}^c \subseteq [0, \infty)^2$  that satisfy the following properties:*

1. *The union of the closures is exhaustive:<sup>7</sup> (i.e.,  $Cl(\Lambda_{0-eff}^c) \cup Cl(\Lambda_{max}^c) \cup Cl(\Lambda_{2-msg}^c) = [0, \infty)^2$ ).*

---

<sup>7</sup> $Cl(\Lambda)$  is the closure of the set  $\Lambda$ , i.e., the set  $\Lambda$  together with all its limit points.

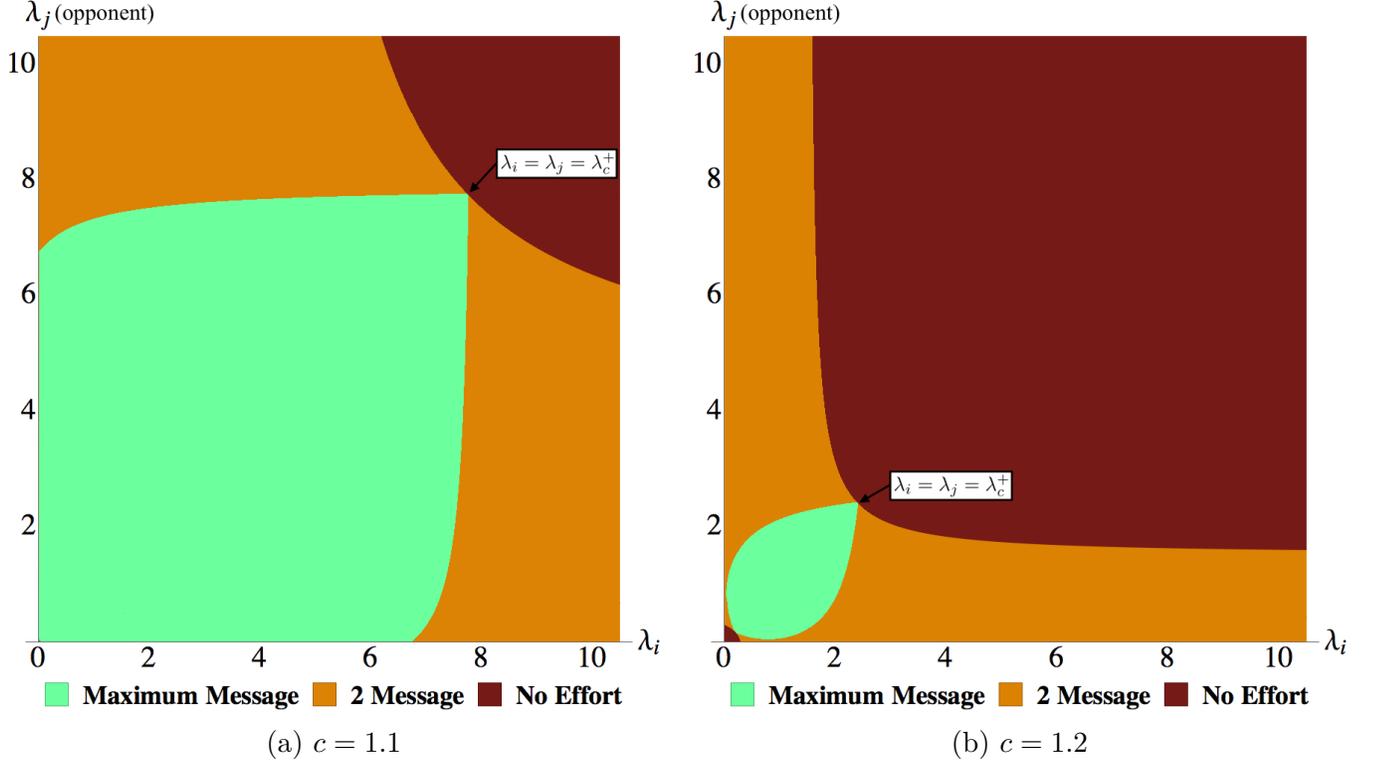


Figure 2: **Unique Perfect Equilibrium Types in a Reneging Aversion Parameter Space.** The  $x$  axis in each figure presents the player's level of reneging aversion ( $\lambda_i$ ) and the  $y$  axis presents the partner's level of reneging aversion ( $\lambda_j$ ). The left panel deals with a cost 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").

2. Let  $\left( (s_i^*, s_j^*), (x_i^*(\vec{s}), x_j^*(\vec{s})) \right)$  be a trembling-hand perfect equilibrium of the partnership game with effort cost  $c$  and reneging costs of  $\lambda_i$  and  $\lambda_j$ .
  - (a) No-effort equilibrium: If  $(\lambda_i, \lambda_j) \in \Lambda_{0-eff}^c$  then  $x_i^*(\vec{s}^*) = x_j^*(\vec{s}^*) = 0$  and, for each agent  $k \in \{i, j\}$ ,  $\lambda_k > 0$  implies that  $s_k^* = 0$ .
  - (b) Two-message equilibrium: If  $(\lambda_i, \lambda_j) \in \Lambda_{2-msg}^c$  and  $\lambda_i > \lambda_j$  then (I)  $s_i^* = 1$ , and (II)  $\lambda_j > 0$  implies that  $s_j^* < 1$ .
  - (c) Maximum message equilibrium: If  $(\lambda_i, \lambda_j) \in \Lambda_{max}^c$  then  $s_i^* = s_j^* = 1$ .
3. If  $c \geq \sqrt{2}$  then  $\Lambda_{0-eff}^c = [0, \infty)^2$  (i.e., only the no-effort equilibrium exists).
4. If  $c \in (1, \sqrt{2})$  and  $\lambda_j < \frac{2}{c} - c$ , then there is  $x_{\lambda_i}^c$ , such that  $(\lambda_i, \lambda_j) \in \Lambda_{2-msg}^c \forall \lambda_i > x_{\lambda_i}^c$ .
5. If  $c \in (1.25, \sqrt{2})$ , then: (I)  $\Lambda_{max}^c = \emptyset$ , (II)  $(\lambda, \lambda) \in \Lambda_{0-eff}^c$  for each  $\lambda \geq 0$ , and (III) there exist  $0 < \underline{\lambda}_c < \bar{\lambda}_c$ , such that either  $\lambda_i, \lambda_j \leq \underline{\lambda}_c$  or  $\lambda_i, \lambda_j \geq \bar{\lambda}_c$  implies that  $(\lambda_i, \lambda_j) \in \Lambda_{0-eff}^c$ .

6. If  $c \in (1, 1.25)$ , then there exist  $0 < \lambda_c^- < \lambda_c^+$ , s.t. (I)  $\Lambda_{max}^c$  is convex, (II)  $\lambda \in (\lambda_c^-, \lambda_c^+)$  implies that  $(\lambda, \lambda) \in \Lambda_{max}^c$ , (III)  $(\lambda, \lambda') \in \Lambda_{max}^c$  implies that  $\lambda_c^- \leq \max(\lambda, \lambda') < \lambda_c^+$ , (IV) there exist  $\underline{\lambda}_c \leq \lambda_c^-$  and  $\bar{\lambda}_c \geq \lambda_c^+$ , such that if either  $\lambda_i, \lambda_j < \underline{\lambda}_c$  or  $\lambda_i, \lambda_j > \bar{\lambda}_c$ , then  $(\lambda_i, \lambda_j) \in \Lambda_{0-eff}^c$ .

*Sketch of proof for the case of  $c \in (1, 1.25)$ .* When both  $\lambda_i$  and  $\lambda_j$  are low or when both are high, the unique equilibrium is a no-effort equilibrium. The intuition is similar to the one presented before Proposition 1: (1) 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 and (2) 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, 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 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.

Finally, if both players' levels of reneging aversion are intermediate (and sufficiently similar) then we have the maximum message equilibrium. The intuition is similar to that presented in case (2) of Proposition 2. If the partner's level of reneging aversion is not too high, the indirect benefit of overcutting the partner's message (which induces the partner to exert more effort in the second stage) is increasing in the agent's level of reneging aversion (as his promise is 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.  $\square$

## 5 Appealing Properties of Intermediate Reneging Aversion, $\lambda_c^+$

### 5.1 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  $\lambda_i, \lambda_j$  that are located on the boundaries between the open sets  $\Lambda_{0-eff}^c$ ,  $\Lambda_{max}^c$ , and  $\Lambda_{2-msg}^c$ . In this "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  $\lambda_i$  who is matched with a partner with reneging aversion  $\lambda_j$ .

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  $\Lambda_{max}^c$  of intermediate levels of reneging aversion that

induce maximal messages. In Corollary 2 of Theorem 1, 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 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  $(\lambda_i, \lambda_j)$  with multiple equilibria, we can apply any arbitrary equilibrium selection function (without affecting our results), and we let  $\pi(\lambda, \lambda')$  be the material payoff induced by the arbitrarily selected equilibrium.

The payoff function  $\pi_c : \mathbb{R}^+ \times \mathbb{R}^+ \rightarrow \mathbb{R}^+$  defined above induces a symmetric two-player population game  $\Gamma = (\mathbb{R}^+, \pi)$ . 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 perfect equilibrium of the partnership game (applying the equilibrium selection function mentioned above when multiple perfect equilibria exist). In Section 5.3 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  $(\lambda, \lambda)$  is a symmetric (strict) pure Nash equilibrium of the population game if  $\pi(\lambda, \lambda) \geq \pi(\lambda', \lambda)$  ( $\pi(\lambda, \lambda) > \pi(\lambda', \lambda)$ ) for each  $\lambda' \neq \lambda$ .

## 5.2 Appealing Properties of $\lambda_c^+$

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 equilibrium induced by  $(\lambda_c^+, \lambda_c^+)$  induces the best equilibrium outcome among all equilibrium outcomes of symmetric partnership games, i.e.,  $\pi(\lambda_c^+, \lambda_c^+) > \pi(\lambda', \lambda')$  for any  $\lambda' \neq \lambda_c^+$ .
2. As  $c$  converges to 1, the equilibrium payoff converges to the maximum feasible payoff, achieved by both agents exerting the maximum effort of one. This maximum feasible payoff is equal to  $1 - \frac{1}{c}$ , and it converges to 0.5 as  $c$  converges to one.
3. It is a strict equilibrium of the population game (i.e.,  $\pi(\lambda_c^+, \lambda_c^+) > \pi(\lambda', \lambda_c^+)$ ).
4. The equilibrium payoff  $\pi(\lambda_c^+, \lambda_c^+)$  is larger than the mean payoff 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.

Formally:

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

1. “Second-best” symmetric outcome:  $\pi(\lambda_c^+, \lambda_c^+) > \pi(\lambda', \lambda')$  for any  $\lambda' \neq \lambda_c^+$ .
2. Convergence to “first-best” outcome:  $\lim_{c \rightarrow 1} \pi(\lambda_c^+, \lambda_c^+) = \frac{1}{2}$  (which is the best symmetric feasible material payoff).
3. Strict equilibrium of the population game:  $\pi(\lambda_c^+, \lambda_c^+) > \pi(\lambda', \lambda_c^+)$  for each  $\lambda' \neq \lambda_c^+$ .
4. Better outcome than the sequential-game equilibrium outcome when effort costs are low: Let  $\pi_i^s$  be the payoff to player  $i$  in the unique equilibrium of the game where efforts are chosen sequentially (and there are no renegeing costs). Then if  $c < 1.22$ ,  $\pi(\lambda_c^+, \lambda_c^+) > \frac{1}{2} \cdot (\pi_i^s + \pi_j^s)$ .
5. Unique pure symmetric equilibrium: If  $c < 1.24$  and there is  $\lambda^*$  such that  $\pi(\lambda^*, \lambda^*) \geq \pi(\lambda', \lambda^*)$  for each  $\lambda'$ , then<sup>8</sup>  $\lambda^* = \lambda_c^+$ .

Sketch of proof. □

1. “Second-best” symmetric outcome: Recall that  $\lambda_c^+$  is the highest level of renegeing aversion that induces a maximum-message equilibrium. Theorem 2 implies that any higher symmetric level of renegeing aversion  $\lambda > \lambda_c^+$  induces the no-effort equilibrium with the lowest possible payoff. One can show that the weaker commitment power induced by lower symmetric levels of renegeing aversion  $\lambda < \lambda_c^+$  induces agents to exert less effort relative to the symmetric equilibrium induced by  $\lambda_c^+$ , and thus to achieve a lower payoff.
2. Convergence to “first-best” outcome: Recall that agents promise maximal efforts in the equilibrium induced by  $(\lambda_c^+, \lambda_c^+)$ , and that they somewhat shirk in the second round due to the fact that the material payoffs are maximised when exerting  $\frac{1}{c}$  times the partner’s effort choice. As  $c$  converges to one, the effort level that maximises the material payoff converges to the partner’s effort choice, the incentives to shirk are diminished, and, as a result, the equilibrium effort levels exerted by the players converge to one.
3. Strict equilibrium of the population game: For any  $\lambda > \lambda_c^+$  the game induced by  $(\lambda, \lambda_c^+)$  admits only the no-effort equilibrium, which yields each player a zero payoff. When  $\lambda < \lambda_c^+$ , the lower commitment power of the player with renegeing aversion  $\lambda$  implies that the players exert less effort in the unique perfect equilibrium, and that the payoff of both players is strictly worse than the equilibrium payoff of the game induced by  $(\lambda_c^+, \lambda_c^+)$ . This implies that  $(\lambda_c^+, \lambda_c^+)$  is a strict equilibrium of the population game.
4. Better outcome than the sequential-game equilibrium outcome: In the sequential effort setting (with no renegeing costs) the “Stackelberg leader” will choose effort level 1 and the follower will choose

---

<sup>8</sup>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  $c \in (1.24, 1.25)$ .

effort level  $\frac{1}{c}$ . In the equilibrium outcome under  $\lambda_c^+$ , 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.<sup>9</sup> 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 choose to put in more effort himself, leading his partner to want 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 cost  $\lambda < \lambda_c^+$ , which implies that  $(\lambda, \lambda)$  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, in turn, induces the partner to exert more effort in the unique equilibrium) outweighs the loss induced by the smaller flexibility in the choice of effort in the second stage. Observe that Theorem 1 implies that for any  $\lambda > \lambda_c^+$  the game induced by  $(\lambda, \lambda)$  admits the no-effort equilibrium, which yields each player a payoff of zero. 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. This implies that  $(\lambda, \lambda)$  is not a Nash equilibrium of the population game for any  $\lambda > \lambda_c^+$ .

### 5.3 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 such 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  $\lambda$ s 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:

1. Any symmetric strict equilibrium corresponds to a stable population state in which all the incumbents have the same level of reneging aversion. Any agent who is endowed with a different

---

<sup>9</sup>It turns out that the higher payoff in the equilibrium induced by  $(\lambda_c^+, \lambda_c^+)$  holds for any  $c < 1.22$ , but it does not hold for  $c \in (1.22, 1.25)$ .

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 implies that the homogeneous population state in which all agents have the same intermediate level of reneging aversion  $\lambda_c^+$  is dynamically stable. Part (3) of Theorem 2 implies that this state is the unique homogeneous stable state. This suggests a tendency of evolutionary processes to select the level of reneging aversion  $\lambda_c^+$  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 \(2001\)](#), 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 Section 7 we discuss the relevant notions of continuous stability proposed by these authors, and explain why imposing these more restrictive solution concepts does not affect our results.

*Remark 2* (Alternative equilibrium selection). Theorem 2 depends on the equilibrium election function choosing the most efficient equilibrium in the game in which both agents have reneging cost  $\lambda_c^+$ . In what follows, we discuss two possible arguments suggesting that the result can hold also without this 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). In such discrete environments one can state a result that is essentially the same as Theorem 2, in which  $\lambda_c^+$  is replaced with the highest feasible discrete reneging cost that is smaller than  $\lambda_c^+$ .
2. **Focality of the efficient equilibrium:** We do not formalise the dynamic process leading a population to the homogeneous state in which every individual has the level of reneging aversion  $\lambda_c^+$ . Intuitively, one plausible way in which the population can converge to  $\lambda_c^+$  is from 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  $\lambda$  in which the messages are maximal. As argued in the proof of part (4), the state  $\lambda$  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  $\lambda_c^+$ , and along this dynamic sequence the agents play the unique perfect equilibrium, which has maximal messages. Arguably, it is plausible that also after the population converges to every agent

having reneging aversion  $\lambda_c^+$  (and multiple equilibria exist) the agents will continue to play the the “focal” equilibrium, which is similar to the unique maximum-message equilibrium played against the previous incumbents with  $\lambda < \lambda_c^+$ .

## 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) reneging costs are perfectly observed in the population game. In this section we relax each of these assumptions, and show that our main results are robust to these changes. Specifically, we show that the equilibrium induced by the symmetric pair of intermediate levels of reneging aversion  $(\lambda_c^+, \lambda_c^+)$  still satisfies all the appealing properties of Theorem 2 except uniqueness, namely: (1) it is a strict Nash equilibrium of the population game, (2) it induces the second-best symmetric outcome, (3) its outcome converges to the first-best outcome in the limit, as  $c$  converges to one, and (4) its outcome is better than the equilibrium outcome in the sequential effort choice game without reneging costs.

### 6.1 Sequential Communication

In this subsection, we examine a variant of the partnership game where promises are sequential rather than simultaneous. We show that the equilibrium induced by the symmetric pair of intermediate levels of reneging aversion  $(\lambda_c^+, \lambda_c^+)$  still satisfies the appealing properties of Theorem 2.

#### 6.1.1 Adaptations to the Model

The partnership game with sequential communication proceeds as follows. In stage 0, nature chooses at random which player (denoted by  $i$ ) will be the first to communicate (where each player has a probability of 50% to be 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.

#### 6.1.2 Robustness of Main Results

We now show that the equilibrium induced by the partnership game with sequential communication in which both players have reneging aversion  $\lambda_c^+$  induces a unique perfect equilibrium that satisfies the same appealing properties as in the baseline model:<sup>10</sup> (1) “second-best” symmetric outcome, (2) convergence to “first-best” outcome, (3) strict equilibrium of the population game, and (4) better outcome than that of the sequential-game equilibrium without reneging costs. Formally:

---

<sup>10</sup>We leave for future research the question of whether the population game with sequential communication admits additional symmetric pure equilibria.

**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 simultaneous communication game (as defined in Theorem 1). The unique subgame-perfect equilibrium induced by  $(\lambda_c^+, \lambda_c^+)$  under sequential communication 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  $c < 1.2$  then  $\pi(\lambda_c^+, \lambda_c^+) > \pi(\lambda', \lambda')$  for any  $\lambda' \neq \lambda_c^+$ .
3. Convergence to “first-best” outcome:  $\lim_{c \rightarrow 1} \pi(\lambda_c^+, \lambda_c^+) = \frac{1}{2}$  (which is the best feasible material payoff).
4. Strict equilibrium of population game:  $\pi(\lambda_c^+, \lambda_c^+) > \pi(\lambda', \lambda_c^+)$  for each  $\lambda' \neq \lambda_c^+$ .
5. Better outcome than that of the sequential-game equilibrium without reneging costs: Let  $\pi_i^s$  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  $\pi(\lambda_c^+, \lambda_c^+) > \frac{1}{2} \cdot (\pi_i^s + \pi_j^s)$ .

*Sketch of proof.* It turns out that for the combinations of levels of reneging aversion that induce maximum-message equilibria or two-message equilibria in the game with simultaneous communication, under sequential communication a unique subgame-perfect equilibrium (which is therefore also trembling-hand perfect) is induced in which promises (and therefore effort levels and payoffs) are the same as under simultaneous communication. Therefore, the function  $\pi(\lambda_i, \lambda_j)$  takes the the same values in these regions as in the simultaneous game, and the results in relation to these regions carry over to the sequential communication setting. For combinations of levels of reneging aversion that induce a no-effort equilibrium under simultaneous communication, promises, effort levels, and payoffs are weakly greater in the unique subgame-perfect equilibrium under sequential communication. Yet, it is shown that, despite these somewhat higher payoffs under these combinations of reneging aversion, it is still not profitable for any player to deviate from  $(\lambda_c^+, \lambda_c^+)$  in the population game. Finally, we show that even though with some symmetric pairs of reneging aversion higher than  $\lambda_c^+$  there exist some asymmetric equilibria (analogous to the “two-message” equilibria in the simultaneous communication game) that yield one player a higher realised payoff than  $\pi(\lambda_c^+, \lambda_c^+)$ , nevertheless, when the cost of effort is low enough (i.e., when  $c < 1.2$ ) the average payoff across the two players in such equilibria (and hence the expected payoff for both players) is lower than  $\pi(\lambda_c^+, \lambda_c^+)$ , which therefore remains the “second-best” outcome.  $\square$

## 6.2 One-Sided Reneging Costs

In this subsection, we examine a variant of the model in which an agent suffers a reneging cost only when he exerts less effort in the second stage than he promised in the first. Unlike the baseline model, an agent does not suffer a cost when he exerts more effort than he promised. This “one-sided” reneging cost may reflect “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 he promised, there is no damage to the partner, and thus no guilt. When the agent’s exerted effort

( $x_i$ ) is less than his promise ( $s_i$ ), then the loss to the partner is  $x_j \cdot (s_i - x_i)$ , which is proportional to the difference between the promised and exerted effort.

Our main result in this subsection shows that the equilibrium induced by the symmetric pair of intermediate levels of renegeing aversion ( $\lambda_c^+, \lambda_c^+$ ) still satisfies most of the appealing properties of Theorem 2. Specifically, we show that ( $\lambda_c^+, \lambda_c^+$ ) is a strict Nash equilibrium of the population game, 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 renegeing costs.

### 6.2.1 Adaptations to the Model

The material payoffs remain the same as in the baseline model. The renegeing cost term in the subjective utility function of each player  $i$  is redefined as follows:

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

That is, an agent incurs an intrinsic cost of renegeing only if his promise is higher than his exerted effort. In this case, he incurs a quadratic cost analogous to that in the baseline model. All other aspects of the partnership game remain the same as in the baseline model. 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 renegeing costs coincides with the set of equilibria in the baseline model with two-sided renegeing 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 renegeing costs, some symmetric partnership games ( $\lambda, \lambda$ ) (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  $\pi(\lambda, \lambda)$  as the mean payoff of the two players' roles. This corresponds to a homogeneous population of agents with renegeing aversion  $\lambda$ , in which each agent has equal probability of being assigned to each role in the selected asymmetric equilibrium.

### 6.2.2 Robustness of Main Results

We now show that the equilibrium induced by  $\lambda_c^+$  satisfies the same appealing properties as in the baseline model: (1) “second-best” symmetric outcome, (2) convergence to “first-best” outcome, (3) strict equilibrium of the population game, and (4) better outcome than that of the sequential-game equilibrium without renegeing costs. Formally:

**Proposition 4.** *Fix  $c \in (1, 1.25)$ . Let  $(\lambda_c^+, \lambda_c^+)$  be the highest symmetric pair of levels of renegeing aversion inducing a maximum-message equilibrium in the partnership game with two-sided renegeing costs (as defined in Theorem 1). The equilibrium induced by  $(\lambda_c^+, \lambda_c^+)$  with one-sided renegeing 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  $\pi(\lambda_c^+, \lambda_c^+) > \pi(\lambda', \lambda')$  for any  $\lambda' \neq \lambda_c^+$ .
3. Convergence to “first-best” outcome:  $\lim_{c \rightarrow 1} \pi(\lambda_c^+, \lambda_c^+) = \frac{1}{2}$  (which is the best feasible material payoff).
4. Strict equilibrium of the population game:  $\pi(\lambda_c^+, \lambda_c^+) > \pi(\lambda', \lambda_c^+)$  for each  $\lambda' \neq \lambda_c^+$ .
5. Better outcome than that of the sequential-game equilibrium without renegeing costs: Let  $\pi_i^s$  be the payoff to player  $i$  in the unique equilibrium of the game where efforts are chosen sequentially (and there are no renegeing costs). Then, if  $c < 1.22$  then  $\pi(\lambda_c^+, \lambda_c^+) > \frac{1}{2} \cdot (\pi_i^s + \pi_j^s)$ .

*Sketch of proof.* We show that in the setup with one-sided renegeing costs, a player’s best reply function either leads him to “renege downward” (i.e., exert effort lower than his promise) in which case the trade-off he faces is essentially the same as in the game with two-sided costs, or leads him to promise low effort and then “renege upward” and perfectly undercut his partner by playing  $x_i = \frac{x_j}{c}$ . In the game induced by  $(\lambda_c^+, \lambda_c^+)$ , a maximum-message equilibrium in which players renege downward and exert effort levels equal to those in the two-sided case exists and remains unique. For pairs of levels of renegeing aversion that induce a “two-message” equilibrium in the two-sided game, there may exist equilibria in which one player promises maximum effort and his partner makes a promise of low effort and reneges upward (and the equilibria with promises and effort levels equal to those in the two-sided case may or may not exist, depending on the parameters of the game). While the player who makes a promise of low effort undercuts “perfectly” in such an equilibrium, the effort levels induced are so low that any player deviating from  $\lambda_c^+$  to a level of renegeing aversion that induces them to renege upward in equilibrium achieves a lower payoff than  $\pi(\lambda_c^+, \lambda_c^+)$ . For this reason, when the cost of effort is low (i.e., when  $c < 1.22$ ),  $(\lambda_c^+, \lambda_c^+)$  still induces the second-best equilibrium outcome.  $\square$

### 6.3 Fixed Renegeing Costs

In our baseline model we assumed that the renegeing cost is proportional to the difference between the promised and exerted effort. In this subsection, we examine a variant of the model in which an agent incurs a fixed renegeing 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 renegeing on a promise, regardless of the size of the renegeing, incurs the same intrinsic cost to the agent. In what follows we show that our main results on the appealing properties of intermediate renegeing aversion can be extended to this setup.

#### 6.3.1 Adaptations to the Model

Fix  $c \in (1, 2)$ . For each  $\beta_i, \beta_j \geq 0$  we define *the partnership game with fixed renegeing costs*  $\beta_i, \beta_j$  in the same way as the partnership game defined in Section 4, except that we change the renegeing cost term such that the subjective utility function of each player  $i$  is redefined as follows:

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

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).

### 6.3.2 Robustness of Main Results

Observe that the payoff function defined in (8) satisfies all the assumptions of Proposition 1, which implies that both agents essentially exert no effort in any pure subgame-perfect equilibrium whenever the fixed reneging costs are either too low or too high.

Our next result shows that for any  $c \in (1, 2)$ , there exists an intermediate level of fixed reneging aversion,  $\beta_c^+$ , that induces the players to promise and exert the maximal level of effort as part of a trembling-hand perfect equilibrium. In particular, this equilibrium induces the first-best outcome (i.e., it yields the best feasible symmetric payoff, which maximises the sum of payoffs of the players).

**Proposition 5.** *For any  $c \in (1, 2)$ , there exists an intermediate level of reneging aversion  $\beta_c^+$ , 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.*

*Sketch of proof.* Let  $\beta_c^+$  be the level of fixed reneging aversion, for which an agent who has promised maximal effort, and believes his partner be exerting maximal effort, is indifferent between exerting maximal effort (and keeping his promise) and breaking his promise and exerting an effort of  $\frac{1}{c}$  (which is the effort that maximises the agent's payoff, conditional on breaking his promise). The definition of  $\beta_c^+$  implies that following a pair of maximal promises, exerting maximal effort by both players is a second-stage equilibrium in the induced subgame. Next, consider a deviation of player  $i$  to promising effort  $s_i < 1$ . The fact that agent  $i$  promises nonmaximal effort implies that he will exert nonmaximal effort in any second-stage equilibrium of the induced subgame. One can show that this implies that player  $j$  (who anticipates that her partner will exert nonmaximal effort) strictly prefers exerting nonmaximal effort (and breaking her promise) to exerting maximal effort (and keeping her promise). Player  $i$  therefore cannot successfully undercut player  $j$  at the promising stage, and so this deviation is not profitable.  $\square$

*Remark 3.* 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.

## 6.4 Partial Observability of Reneging Aversion

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

### 6.4.1 Population Game with Partial Observability

In what follows, we describe the adaptations to the model of the population game presented in Section 5 (and to its evolutionary interpretation presented in Section 5.3) that are required to accommodate partial observability. Let  $q \in [0, 1]$  denote the fraction of matches in which both players observe their partner's

level of reneging aversion. That is, we assume that when the agents are randomly matched into pairs, in a share  $q$  of the pairs both agents observe their partner’s level of reneging aversion, while in the remaining share  $1 - q$  of the pairs the partners are “strangers,” and neither of them observes any information about their partner’s reneging aversion. One may interpret the observation of reneging aversion to be the result of obtaining information about a partner’s past behaviour (either through direct observation or by communicating with agents who interacted with the partner in the past). Under this interpretation,  $q$  may represent the likelihood that agents who are matched together have prior information about each other.

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

Consider a setup in which the incumbent agents have reneging aversion  $\lambda \in \mathbb{R}^+$ , while occasionally one of the agents is endowed with a different level of reneging aversion (henceforth, a *mutant*). Let  $\pi_{no}(\lambda', \lambda|\lambda)$  be the material payoff of a mutant (he) with a reneging aversion of  $\lambda'$  who faces an incumbent partner (she) with a reneging aversion of  $\lambda$  who believes with probability one that her partner has a reneging aversion of  $\lambda$ . Note that this belief is consistent with a situation in which a single mutant experiments with a different level of reneging aversion within an infinite population of agents. The partner plays her part of the unique perfect equilibrium of the game with observability, denoted by  $G(\lambda, \lambda)$ , while the mutant plays his best reply to her strategy.

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

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

Observe that when  $q = 1$  the current model coincides with the perfect observability described in Section 5, whereas when  $q = 0$  it corresponds to the nonobservability of the partner’s reneging aversion.

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

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

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

### 6.4.2 Robustness of Theorem 2

The following result demonstrates the robustness of Theorem 2 to almost perfect observability. It is immediate that the equilibrium outcome induced by  $(\lambda_c^+, \lambda_c^+)$  is still a second-best symmetric outcome, that it converges to the first-best outcome, and that it is better than the sequential-game equilibrium, for any  $q \in [0, 1]$  (as the payoff in a homogeneous population is independent of  $q$ ). In what follows we show that the remaining results of Theorem 2 hold also for any observability level  $q < 1$  that is sufficiently close to one. Formally:

**Proposition 6.** *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 (as defined in Theorem 1). Then, there exists  $\bar{q} \in (0, 1)$  such that for each  $q \in [\bar{q}, 1]$ , the equilibrium induced by  $(\lambda_c^+, \lambda_c^+)$  has the following properties:*

1. *Second-best outcome:  $\pi(\lambda_c^+, \lambda_c^+) > \pi(\lambda', \lambda')$  for any  $\lambda' \neq \lambda_c^+$ .*
2. *Convergence to first-best outcome:  $\lim_{c \rightarrow 1} \pi_q(\lambda_c^+, \lambda_c^+) = \frac{1}{2}$ .*
3. *Strict equilibrium of the population game:  $\pi(\lambda_c^+, \lambda_c^+) > \pi_q(\lambda', \lambda_c^+ | \lambda_c^+)$  for each  $\lambda' \neq \lambda_c^+$ .*
4. *Better outcome than the sequential-game equilibrium outcome: let  $\pi_i^s$  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  $\pi_q(\lambda_c^+, \lambda_c^+) > \frac{1}{2} \cdot (\pi_i^s + \pi_j^s)$ .*

### 6.4.3 Nonrobustness of No-Effort Equilibrium

We recover a central result from the evolutionary literature on the stability of payoff-maximising preferences under anonymity but show that it is not robust to *any* positive probability of correlated observation of preferences in our model. The following simple result shows that when there is no observability (i.e.,  $q = 0$ ) no effort is exerted in any equilibrium of the population game. Formally:

**Proposition 7.** *Fix  $c \in (1, 1.25)$  and  $q = 0$ . In any symmetric pure Nash equilibrium of the population game, all agents exert an effort of zero on the equilibrium path, and any agent  $i$  with  $\lambda_i > 0$  sends a message of zero.*

This result is similar to those in the existing literature that show that when agents are matched uniformly and anonymously (i.e., no observability or assortativity) and the selection dynamics are payoff monotone, then players maximise their material payoffs in any stable population state (see, e.g., [Ok & Vega-Redondo, 2001](#); [Dekel et al., 2007](#)).<sup>11</sup>

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

<sup>11</sup>A notable exception is [Frenkel et al. \(2018\)](#) who present a plausible model of evolutionary dynamics that are not payoff-monotone due to sexual inheritance in a biological process, or due to combining traits from more than one mentor in a social learning process. They show that in such processes, stable population states do not correspond to Nash equilibria of the underlying material payoff game.

**Proposition 8.** *Fix  $c \in (1, 1.25)$  and  $q > 0$ . Then, in any symmetric pure Nash equilibrium of the population game, players exert positive levels of effort on the equilibrium path.*

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

## 7 Conclusion

We have demonstrated that an intermediate level of renegeing 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. While our baseline model assumes a specific family of quadratic payoff functions, we have shown in the previous section that our main results are robust to various relaxations of the model’s assumptions of sequential communication, one-sided renegeing costs, discontinuous renegeing costs, and partial observability. Moreover, the general model of Section 3 shows that our main qualitative insight, namely, that nonintermediate renegeing costs induce zero effort and only intermediate renegeing costs have the potential to obtain efficient outcomes in which agents exert effort, holds under mild assumptions on the payoff function.

These results demonstrate a strong tendency of evolution to select preferences for the partial keeping of promises. In stable populations, we see players making slightly “overoptimistic” promises and, while these are not fully realised, the outcome is welfare-maximising among symmetric equilibria of the game. This outcome stands in sharp contrast to the cheap talk prediction of no effort ever being exerted in these partnerships.

We have here developed the first evolutionary analysis of a direct concern for keeping one’s word. In doing so, we give an evolutionary explanation of several key observations in the related empirical literature. In our model, a population of players with the stable level of renegeing aversion will exert no effort if they are not allowed to communicate before choosing their actions, but the opportunity to send messages will lead to promises being made and higher levels of effort being exerted. This replicates the 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 have the ability to communicate before playing, and that players communicating promises are particularly likely to cooperate.<sup>12</sup> Secondly, in the presence of communication, the degree of cooperation in our model is both incomplete (some renegeing 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,

---

<sup>12</sup>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.

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 renegeing, consistent with our modeling of convex renegeing 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*).<sup>13</sup> We lay the theoretical foundations for both of these accounts of promise-keeping. In our baseline model, the renegeing costs can be interpreted as a cost of promise-breaking *per se*. In the one-sided variant of [Section 6.2](#), individuals suffer a cost of renegeing 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 renegeing aversion.

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

## A Trembling-Hand Perfection

In this section we formally define the refinement of trembling-hand perfection in our setup. This refinement requires that the equilibrium behaviour should be a limit of equilibria of perturbed environments in which the players occasionally make mistakes (“tremble”), where the limit is taken when the error probability converges to zero.

It turns out that none of our results depend on whether or not players occasionally tremble in the second stage. Thus, in order to simplify the notation, we present a simpler definition in which players only tremble in the first 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](#)).

---

<sup>13</sup>[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.

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  $BR_i^c(\sigma_{-i}) \subseteq [0, 1]$  be the set of distributions over promises (mixed promises) that are best replies to  $\sigma_{-i}$  (where the players are assumed to follow the unique Nash equilibrium when choosing their effort levels in the second-stage of the game), i.e.,

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

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

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

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

**Definition 3.** 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  $\epsilon_n \rightarrow 0$  of a sequence of  $\epsilon_n$ -perfect equilibria.

Simon & Stinchcombe (1995, Theorem 2.1) show that the set of perfect equilibria is a closed, nonempty subset of the set of Nash equilibria of the promise game. The arguments presented in the proof of Lemma 3 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  $\left( (s_i^*, s_j^*), (x_i^*(\vec{s}), x_j^*(\vec{s})) \right)$  of the partnership game is trembling-hand perfect if the pair of promises  $(s_i^*, s_j^*)$  is a trembling-hand perfect equilibrium of the induced one-shot promise game.

*Remark 4.* In our analysis 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 Definition 2. Weak perfection has no bite in our setup: any subgame-perfect equilibrium of the partnership game satisfies weak perfection. Specifically, consider the region  $\Lambda_{max}^c$  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  $\epsilon$  away from the unique best reply message  $\epsilon$ , 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.

### B.1 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).<sup>14</sup> We leave the analytic analysis of this conjecture (which, we believe, holds also for partial observability with a sufficiently high  $q$ ) for future research.

It is relatively straightforward to extend Propositions 7 and 8 to mixed equilibria and to asymmetric equilibria. We refrain from doing so in order to simplify the notation of Section 6.4 (the formal definition of symmetric equilibria requires a somewhat more complicated notation). The arguments presented in the proofs of both propositions hold with minor changes also for mixed and asymmetric equilibria, and it can be shown that: (1) if  $q = 0$ , then all incumbents exert zero effort in any equilibrium of the population game, and (2) for any  $q > 0$  in any equilibrium of the population game, a positive share of incumbent agents exert positive effort with positive probability (and hence make positive promises, and are endowed with positive renegeing aversion). Thus, the endowment of players with positive levels of renegeing aversion in stable population states is a robust property that holds for any positive level of partial observability (at least with the simplifying assumption of perfect correlation between the observations of the two matched agents).

### B.2 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 renegeing aversion. Eshel (1983) argues that in some setups one should also require stability against perturbations in which many (or all) agents slightly change their renegeing aversion. Eshel presents the notion of a *continuous stable strategy* to capture stability also against the latter class of perturbations, and Oechssler & Riedel (2001) further refine it by presenting the notion of evolutionary robustness, which requires stability against all small perturbations consistent with the weak topology (see also the related notions of stability in Milchtaich, 2016). Population state  $\lambda^*$  is *evolutionarily robust* if an agent with cost  $\lambda^*$  outperforms other agents (on average) in any sufficiently

---

<sup>14</sup>The extension to asymmetric equilibria is especially interesting in setups in which the partnership game is played between agents from two different populations of complementary skills, and a stable state of the two populations corresponds to a possibly asymmetric Nash equilibrium of the two-population game (see the related setup studied in Ritzberger & Weibull, 1995).

close perturbed population state  $\mu \in \Delta(\mathbb{R}^+)$ , i.e.,

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

One can show that the population state  $(\lambda_c^+, \lambda_c^+)$  satisfies a slightly 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  $\mu$  assigns positive mass to agents having a renegeing aversion of at most  $\lambda_c^+$ . The intuition is that agents with a slightly higher renegeing aversion (i.e., strictly above  $\lambda_c^+$ ) play a no-effort equilibrium against all agents in the perturbed state  $\mu$ . Thus, they are trivially weakly outperformed by a level of aversion  $\lambda_c^+$ , and strictly outperformed as long as  $\mu$  includes some agents with a renegeing aversion of at most  $\lambda_c^+$  (against whom an agent with renegeing aversion  $\lambda_c^+$  achieves strictly positive payoffs). Finally, minor modifications to the arguments presented in the proof of Theorem 2 show that agents with a renegeing aversion strictly below  $\lambda_c^+$  are strictly outperformed by agents with a renegeing aversion of  $\lambda_c^+$ .

## References

- Abeler, Johannes, Raymond, Collin, & Nosenzo, Daniele. Forthcoming. Preferences for truth-telling. *Econometrica*.
- Alger, Ingela, & Weibull, Jörgen W. 2010. Kinship, incentives, and evolution. *The American Economic Review*, **100**(4), 1725–1758.
- Alger, Ingela, & Weibull, Jörgen W. 2012. A generalization of Hamilton’s rule: Love others how much? *Journal of Theoretical Biology*, **299**, 42–54.
- Alger, Ingela, & Weibull, Jörgen W. 2013. Homo Moralis: Preference evolution under incomplete information and assortative matching. *Econometrica*, **81**(6), 2269–2302.
- Bicchieri, Cristina, & Lev-On, Azi. 2011. Studying the ethical implications of e-trust in the lab. *Ethics and Information Technology*, **13**(1), 5–15.
- Cahuc, Pierre, & Kempf, Hubert. 1997. Alternative time patterns of decisions and dynamic strategic interactions. *The Economic Journal*, **107**(445), 1728–1741.
- Caruana, Guillermo, & Einav, Liran. 2008. A theory of endogenous commitment. *The Review of Economic Studies*, **75**(1), 99–116.
- Charness, Gary. 2000. Self-serving cheap talk: A test of Aumann’s conjecture. *Games and Economic Behavior*, **33**(2), 177–194.
- Charness, Gary, & Dufwenberg, Martin. 2006. Promises and partnership. *Econometrica*, **74**(6), 1579–1601.

- Cooper, Russell, & John, Andrew. 1988. Coordinating coordination failures in Keynesian models. *The Quarterly Journal of Economics*, **103**(3), 441–463.
- Crawford, Vincent. 1998. A survey of experiments on communication via cheap talk. *Journal of Economic Theory*, **78**(2), 286–298.
- Dekel, Eddie, Ely, Jeffrey C., & Yilankaya, Okan. 2007. Evolution of preferences. *The Review of Economic Studies*, **74**(3), 685–704.
- Demichelis, Stefano, & Weibull, Jörgen W. 2008. Language, meaning, and games: A model of communication, coordination, and evolution. *The American Economic Review*, **98**(4), 1292–1311.
- Di Bartolomeo, Giovanni, Dufwenberg, Martin, Papa, Stefano, & Passarelli, Francesco. 2018. Promises, expectations & causation. *Games and Economic Behavior*.
- 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.
- Fischbacher, Urs, & Föllmi-Heusi, Franziska. 2013. Lies in disguise: An experimental study on cheating. *Journal of the European Economic Association*, **11**(3), 525–547.
- Frenkel, Sivan, Heller, Yuval, & Teper, Roe. 2018. The endowment effect as blessing. *International Economic Review*, **59**(3), 1159–1186.
- 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.
- Mas-Colell, Andreu, Whinston, Michael Dennis, & Green, Jerry R. 1995. *Microeconomic Theory*. Vol. 1. Oxford University Press New York.
- 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. 2001. Evolutionary dynamics on infinite strategy spaces. *Economic Theory*, **17**(1), 141–162.
- Ok, Efe A, & Vega-Redondo, Fernando. 2001. On the evolution of individualistic preferences: An incomplete information scenario. *Journal of Economic Theory*, **97**(2), 231–254.
- Radner, Roy, Myerson, Roger, & Maskin, Eric S. 1986. Example of a repeated partnership game with discounting and with uniformly inefficient equilibria. *Review of Economic Design*, **53**(1), 59–69.
- Ritzberger, Klaus, & Weibull, Jörgen W. 1995. Evolutionary selection in normal-form games. *Econometrica*, **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.

## C Additional Figures (For Online Publication)

The appendix presents additional figures demonstrating: (1) how the value of  $\lambda_c^+$ , level of effort, fitness, and subjective payoff change as a function of the cost of effort  $c$ , and (2) the best-reply types and the equilibrium types for four additional values of cost of effort  $c$ : 1.05, 1.15, 1.24, and 1.26.

### C.1 Intermediate Reneging Aversion $\lambda_c^+$ and Equilibrium Values as a Function of $c$

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  $\lambda_c^+$  as a Function of the Cost of Effort  $c$

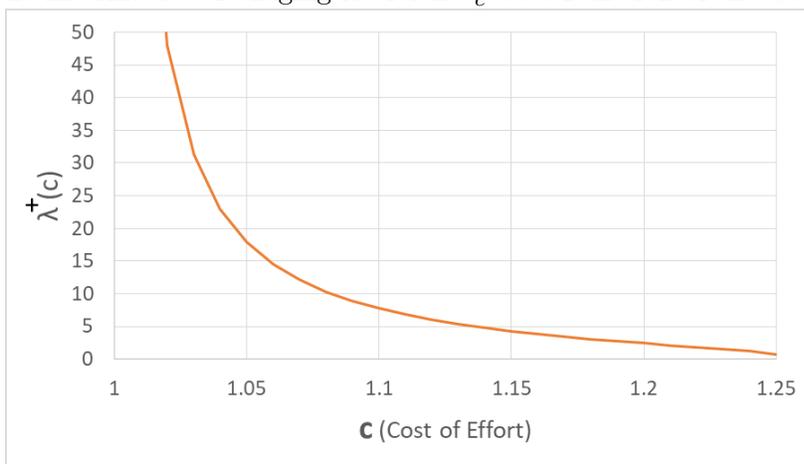
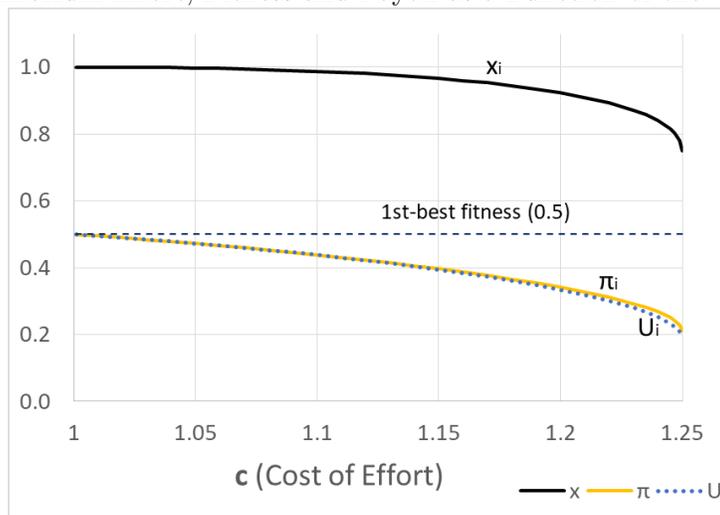


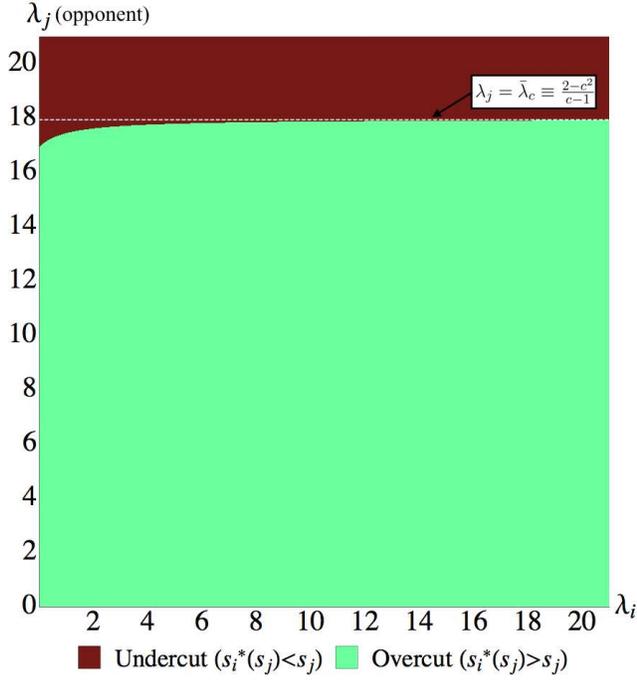
Figure 4 shows the level of effort,  $x_i$ , the material payoff of each player,  $\pi_i$ , and the subjective utility of each player,  $U_i$ , as a function of the cost of effort,  $c$ , in the unique equilibrium induced by the partnership game in which both players have the intermediate level of reneging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

Figure 4: Equilibrium Effort, Fitness and Payoff as a Function of the Cost of Effort  $c$

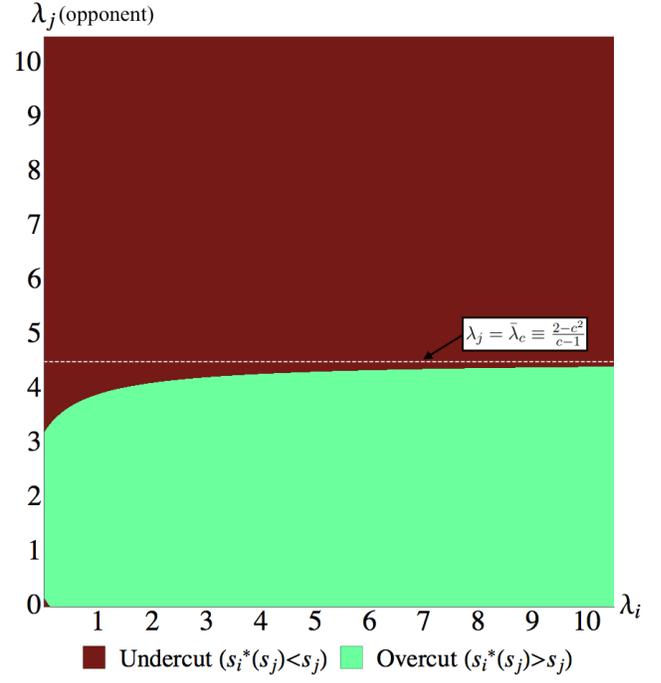


## C.2 Best-Reply Types and the Unique Perfect Equilibrium Types

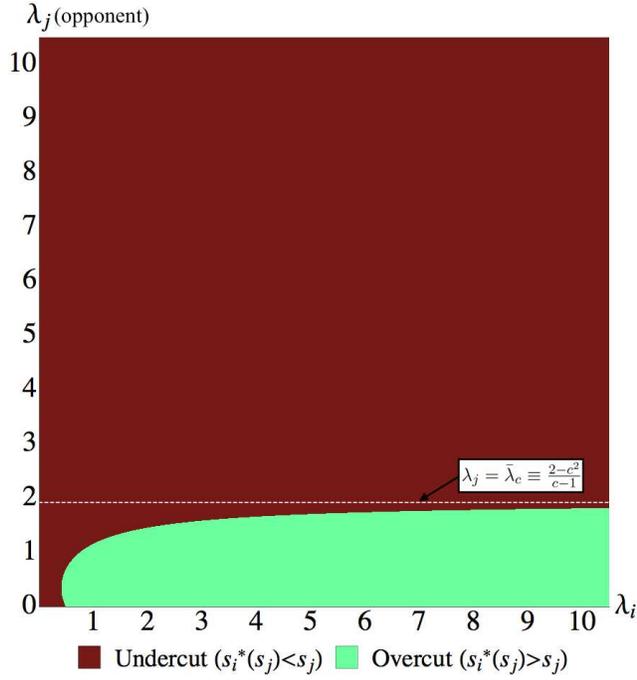
Figure 5 presents the best-reply types for player  $i$  in the reneging aversion parameter space for four costs of effort  $c$ : 1.05, 1.15, 1.24, and 1.26. Figure 6 presents the unique perfect equilibrium types in the reneging aversion parameter space for the same four costs of effort. (The values of 1.1 and 1.2 are presented in Figures 1 and 2 in Section 4.)



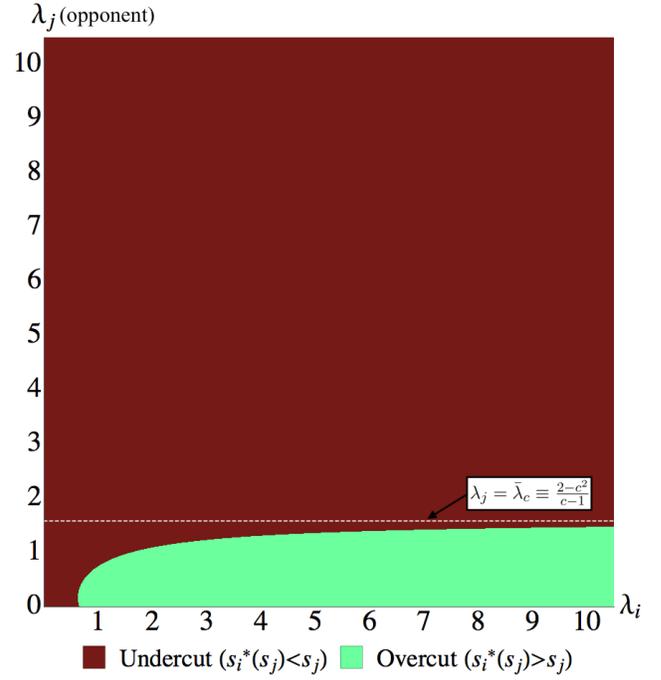
(a)  $c = 1.05$  (Scale of each axis is  $[0,20]$ )



(b)  $c = 1.15$



(c)  $c = 1.24$



(d)  $c = 1.26$

Figure 5: **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 ( $\lambda_i$ ) and the  $y$  axis presents the partner's level of reneging aversion ( $\lambda_j$ ). The dark area in each panel is the region in which player  $i$ 's best reply is to undercut the 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 the partner, i.e.,  $s_i^*(s_j) > s_j$ . The dashed line in each panel shows the value  $\lambda_j = \bar{\lambda}_c \equiv \frac{2-c^2}{c-1} > 0$  presented in Proposition 2, above which player  $i$ 's best reply is to undercut his partner regardless of the value of  $\lambda_i$ .

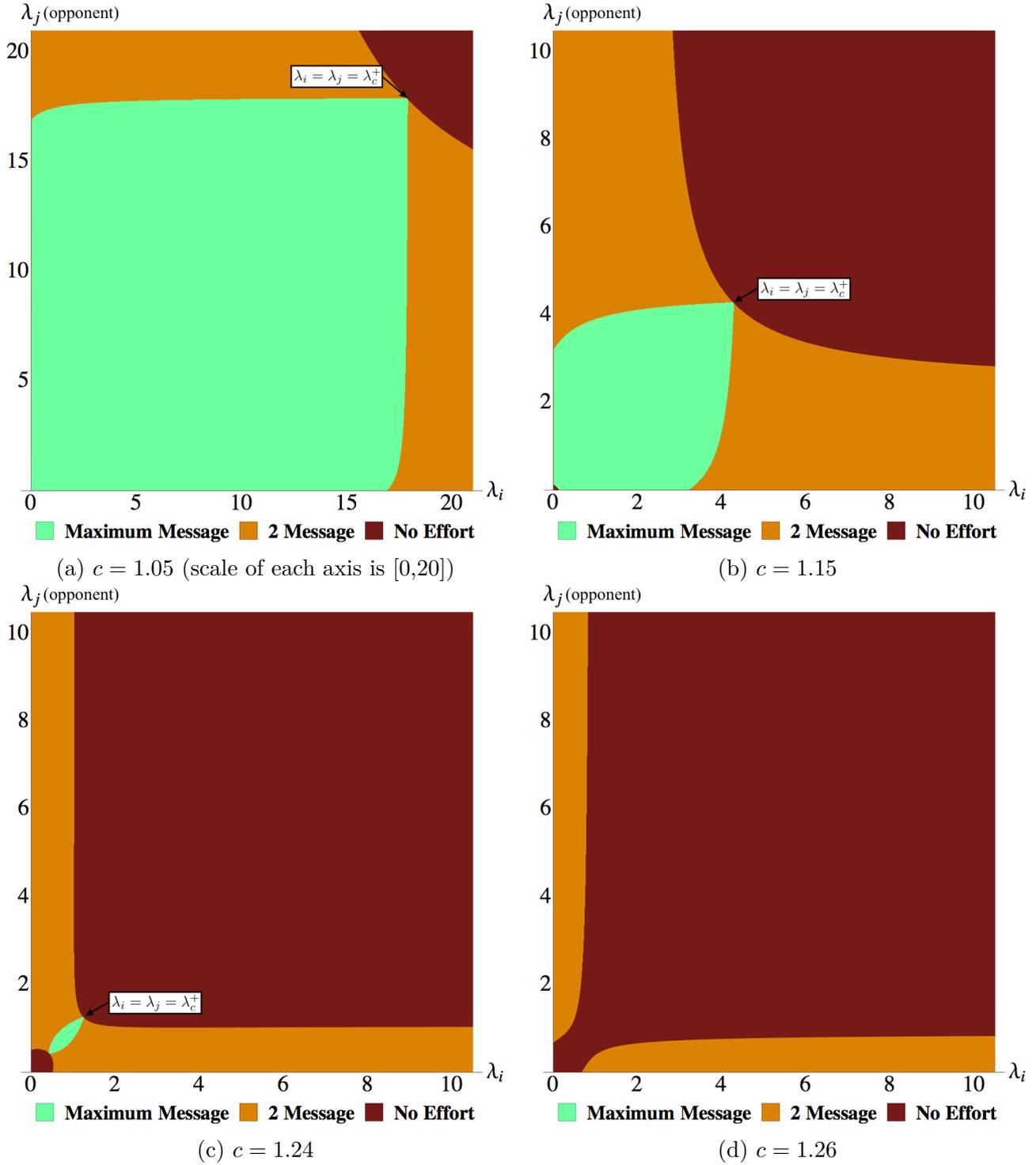


Figure 6: **Unique Perfect Equilibrium Types in a Reneging Aversion Parameter Space.** The  $x$  axis in each panel presents the player's level of reneging aversion ( $\lambda_i$ ) and the  $y$  axis presents the partner's level of reneging aversion ( $\lambda_j$ ). The dark areas in each panel are the regions in which both agents exert no effort in equilibrium ("No Effort"). The light area in each panel is the region in which both agents promise maximal efforts in the unique perfect equilibrium ("Maximum Message"); this region is empty in the case of  $c = 1.26$ . The remaining areas are the regions in which one of the agents sends a maximal promise ("2 Message").

## D Proofs (For Online Publication)

### D.1 Proof of Lemma 1

*Proof.* Fix  $\underline{x} > c$ . For each  $x_i \in [\underline{x}, b]$  and each  $x_j \in [a, x_i]$  define

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

The assumption that  $\pi(x_i, x_j)$  is continuously differentiable implies that  $f(x_i, x_j)$  is continuous in both parameters. The assumptions that  $x_i > c$ ,  $x_i \geq x_j$ ,  $\max(BR_\pi(x_j)) \leq c$  if  $x_j \leq c$ , and  $\max(BR_\pi(x_j)) < x_j$  if  $x_j > c$  imply that  $f(x_i, x_j) > 0$ . Define

$$\tilde{\epsilon} = \min_{x_i \in [\underline{x}, b], x_j \in [a, x_i]} f(x_i, x_j).$$

The compactness of the set  $\{(x_i, x_j) \in [a, b]^2 \mid x_i \in [\underline{x}, b], x_j \in [a, x_i]\}$  and the continuity of  $f(x_i, x_j)$  imply that  $\tilde{\epsilon} > 0$ . Fix  $x_i \in [\underline{x}, b]$  and  $x_j \in [a, x_i]$ . 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 that  $\pi$  encourages shirking above  $c$ .  $\square$

### D.2 Proof of Proposition 1

*Proof.* We first prove point (1) (namely, that the equilibrium efforts are at most  $c + \epsilon$  if  $\lambda_i, \lambda_j < \underline{\lambda}_\epsilon$ ). Fix  $\epsilon > 0$ . The fact that function  $\pi$  encourages shirking above  $c$  implies that there exists  $\delta > 0$  such that for each  $x_i \geq c + \epsilon$  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  $\underline{\lambda}_\epsilon \cdot D(b - a) < \frac{\delta}{2}$ . Assume that there is a pure subgame-perfect equilibrium  $(\vec{s}^*, \vec{x}^*(\vec{s}^*))$  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  $c + \epsilon$ , i.e.,  $x_i^*(\vec{s}^*) \geq c + \epsilon$ . Assume without loss of generality that  $x_i^*(\vec{s}^*) \geq x_j^*(\vec{s}^*)$ . The fact that  $\pi$  encourages shirking above  $c$  implies that there exists  $x'_i$  satisfying

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

which implies that

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

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

We now prove point (2) (namely, that the equilibrium efforts are at most  $c + \epsilon$  if  $\lambda_i, \lambda_j > \bar{\lambda}_\epsilon$ ). Let  $\epsilon' > 0$  be sufficiently small such that  $(b - a) \cdot M \cdot \epsilon' < \frac{\delta}{4}$  and  $\epsilon' < \frac{\epsilon}{8}$  (where  $M$  is the constant from the definition of Lipschitz continuity). Let  $\bar{\lambda}_\epsilon > 0$  be sufficiently large such that  $\bar{\lambda}_\epsilon \cdot D(\epsilon') > 2 \cdot M$ . Assume that there is a pure subgame-perfect equilibrium  $(\vec{s}^*, \vec{x}^*(\vec{s}^*))$  of the partnership game with levels of

reneging aversion  $\lambda_i, \lambda_j \geq \bar{\lambda}_\epsilon$  in which agent  $i$  exerts effort of at least  $c + \epsilon$ , i.e.,  $x_i^*(\vec{s}^*) \geq c + \epsilon$ . Assume without loss of generality that  $x_i^*(\vec{s}^*) \geq x_j^*(\vec{s}^*)$ . The fact that the function  $\pi$  encourages shirking above  $c$  implies that there exists  $\delta > 0$  such that for each  $x_i^* \geq c + \epsilon$  and each  $x_j^* \leq x_i^*$  there exists  $x'_i \leq x_i^*$  such that  $\pi(x'_i, x_j^*) > \pi(x_i^*, x_j^*) + \delta$ . The fact that  $\bar{\lambda}_\epsilon \cdot D(\epsilon') > 2 \cdot M$  implies that  $x_j^*(s'_i, s_j^*) \geq s_j^* - \epsilon'$  for each  $s'_i \in [0, 1]$ . This, in turn, implies that  $|x_j^*(s'_i, s_j^*) - x_j^*(s^*)| \leq 2 \cdot \epsilon'$  for each  $s'_i, s''_i \in [0, 1]$ .

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  $(\vec{s}^*, \vec{x}^*(\vec{s}))$  being a subgame-perfect equilibrium):

$$\begin{aligned} U\left((x'_i, x_j^*(s_i = x'_i, s_j^*)), s_i^*, \lambda_i\right) &= \pi(x'_i, x_j^*(s_i = x'_i, s_j^*)) \geq \pi(x'_i, x_j^*(s^*)) - (b - a) \cdot M \cdot 2 \cdot \epsilon' \geq \pi(x'_i, x_j^*(s^*)) - \frac{\delta}{2} \\ &\geq \pi(x_i^*, x_j^*(s^*)) - \frac{\delta}{2} + \delta \geq U\left((x_i^*, x_j^*(s^*)), s_i^*, \lambda_i\right) + \frac{\delta}{2}. \end{aligned}$$

□

### D.3 Definitions of Notation Used in the Remaining Proofs

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

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

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

### D.4 Proof of Proposition 2

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

#### D.4.1 Lemma Characterising the Best-Reply Correspondence

**Lemma 3.** *Let  $\mu_{\sigma_j}$  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<sup>15</sup>*

---

<sup>15</sup>The choice of the best reply in the latter “knife-edge” case, in which  $\Theta_i = \lambda_j \cdot \mu_{\sigma_j} = 0$ , does not play any role in our results. In all other cases, the unique best-reply function of both players always induces them to choose a pure message and, as a result, both players choose pure messages in all equilibria. This justifies the focus on pure strategies in the main text.

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

*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  $s_j$ :

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

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  $\mu_{\sigma_j}$  (and indeed any  $s_j$ ). When  $\lambda_i > 0$ , the first derivative of player  $i$ 's utility function with respect to  $s_i$ , taking  $\mu_{\sigma_j}$  as given, is a linear function of  $s_i$  and  $\mu_{\sigma_j}$ :

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

Given that  $\lambda_j$  and  $\mu_{\sigma_j}$  are constrained to be (weakly) positive, the second term in Eq. (12) is also (weakly) positive. Therefore, when  $\Theta_i > 0$  (and hence the term multiplying  $s_i$  in Eq. (12) 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:

$$s_i(\mu_{\sigma_j}, \lambda_i, \lambda_j, c) = \frac{\lambda_j}{\Theta_i} \cdot \mu_{\sigma_j} \quad (13)$$

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. (12) 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.  $\square$

#### D.4.2 Conditions for the Existence of Each Best-Reply ‘‘Type’’

**Lemma 4.**  $\Theta_i \leq 0$  (which implies that player  $i$ 's best reply is to send the maximum message) if and only if

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

*Proof.* By the definition of  $\Theta_i$ :

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

where the second  $\iff$  is obtained by multiplying by  $(c + \lambda_i)$  and the third and fourth by gathering terms in  $\lambda_i$  and rearranging. To solve for  $\lambda_i$  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:

$$\lambda_i \leq \frac{-1}{(c + \lambda_j)[c(c + \lambda_j) - 2]} - c < 0, \quad \text{and} \quad c(c + \lambda_j) - 2 > 0, \quad (14)$$

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

We can see that the solution given by Eq. (14) implies that  $\lambda_i < 0$ , which is ruled out by assumption. Therefore, we have that  $\Theta_i \leq 0 \iff$  Eq. (15) holds. Rearranging the second inequality in Eq. (15) to give a condition in terms of  $\lambda_j$  yields the lemma.  $\square$

**Lemma 5.**  $\frac{\lambda_j}{\Theta_i} > 1$  (which implies that player  $i$  sends a message that is some multiple (greater than 1) of player  $j$ 's message) if and only if

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

*Proof.* By the definition of  $\Theta_i$ ,

$$\frac{\lambda_j}{\Theta_i} > 1 \iff \frac{\lambda_j}{c(c + \lambda_j) + \frac{1}{(c + \lambda_i)(c + \lambda_j)} - 2} > 1. \quad (16)$$

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

$$\lambda_j > c(c + \lambda_j) + \frac{1}{(c + \lambda_i)(c + \lambda_j)} - 2 > 0. \quad (17)$$

The second of these inequalities is the requirement that  $\Theta_i > 0$ , which is the converse of the condition derived for Lemma 4, and this second inequality holds when

$$\lambda_i < \frac{1}{(c + \lambda_j)[2 - c(c + \lambda_j)]} - c \quad \text{or} \quad \lambda_j \geq \frac{2}{c} - c. \quad (18)$$

The first inequality in Eq. (17) holds if and only if

$$\begin{aligned} \lambda_j &> c(c + \lambda_j) + \frac{1}{(c + \lambda_i)(c + \lambda_j)} - 2 \\ \iff \lambda_j + 2 - c(c + \lambda_j) &> \frac{1}{(c + \lambda_i)(c + \lambda_j)} \iff (c + \lambda_i)(\lambda_j + 2 - c(c + \lambda_j)) > \frac{1}{(c + \lambda_j)} \\ \iff \lambda_i(\lambda_j + 2 - c(c + \lambda_j)) &> -c(\lambda_j + 2 - c(c + \lambda_j)) + \frac{1}{c + \lambda_j}. \end{aligned} \quad (19)$$

The second  $\iff$  is obtained by multiplying by  $(c + \lambda_i)$ , and the first and third by rearranging. To solve for  $\lambda_i$ , 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:

$$\lambda_i > \frac{1}{(\lambda_j + 2 - c(c + \lambda_j))(c + \lambda_j)} - c > 0 \quad \text{and} \quad \lambda_j + 2 - c(c + \lambda_j) > 0, \quad (20)$$

$$\lambda_i < \frac{1}{(\lambda_j + 2 - c(c + \lambda_j))(c + \lambda_j)} - c < 0 \quad \text{and} \quad \lambda_j + 2 - c(c + \lambda_j) < 0. \quad (21)$$

We can see that the solution given by Eq. (21) implies that  $\lambda_i < 0$ . This is ruled out by assumption, and so we have that the first inequality in Eq. (17)  $\iff$  Eq. (20) holds. Rearranging the inequalities in Eq. (20) yields

$$\lambda_i > \frac{1}{\lambda_j^2(1 - c) + \lambda_j(2 - 2c^2 + c) + c(2 - c^2)} - c > 0 \quad \text{and} \quad \lambda_j < \frac{2 - c^2}{c - 1}. \quad (22)$$

Combining the inequalities in Eqs. (22) and (18) and observing that  $\frac{2}{c} - c = \frac{2 - c^2}{c} < \frac{2 - c^2}{c - 1}$ , and that therefore  $0 < \lambda_i < \frac{1}{(c + \lambda_j)(2 - c(c + \lambda_j))} - c \Rightarrow \lambda_j < \frac{2 - c^2}{c - 1}$ , yields the lemma.  $\square$

**Lemma 6.**  $0 < \frac{\lambda_j}{\Theta_i} < 1$  (which implies that player  $i$  sends a message that is some fraction (less than 1) of player  $j$ 's message) if and only if

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

*Proof.* The inequality  $0 < \frac{\lambda_j}{\Theta_i} < 1$  implies that  $\Theta_i > 0$  and so Eq. (18) must hold. We also must have

that  $\frac{\lambda_j}{\Theta_i} < 1$ . In the proof of Lemma 5 it was demonstrated that  $\frac{\lambda_j}{\Theta_i} > 1 \iff \text{Eq. (22)}$  holds. By taking the converse of Eq. (22) we have that  $\frac{\lambda_j}{\Theta_i} < 1$  if and only if

$$\lambda_i < \frac{1}{\lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)} - c \text{ or } \lambda_j \geq \frac{2-c^2}{c-1}. \quad (23)$$

From the proof of Lemma 5, we have that  $\Theta_i > 0$  if and only if

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

To see that Eq. (23) 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. (23) implies the second inequality in Eq. (24). Next, we see that if  $\lambda_j \geq \frac{2}{c} - c$  then we clearly have the second inequality in Eq. (24). 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. (23) implies that the first inequality in Eq. (23) holds, which in turn implies the first inequality in Eq. (24):

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

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

#### D.4.3 Proof of Proposition 2

*Proof.* We prove each point in turn:

1. Observe that  $c \geq \sqrt{2} \implies \frac{2-c^2}{c-1} \leq 0$ . Given that  $\lambda_j \geq 0$ , we therefore have  $\lambda_j \geq \frac{2-c^2}{c-1}$ . Lemma 6 shows that if  $\lambda_j \geq \frac{2-c^2}{c-1}$  then  $0 < \frac{\lambda_j}{\Theta_i} < 1$ . Lemma 3 implies that if  $\lambda_i > 0$  and  $0 < \frac{\lambda_j}{\Theta_i} < 1$  and  $s_j > 0$ , then  $s_i^*(s_j|\lambda_i, \lambda_j, c) < s_j$ . We therefore have that  $c \geq \sqrt{2} \implies s_i^*(s_j|\lambda_i, \lambda_j, c) < s_j$ , for each  $s_j > 0$  and  $\lambda_i > 0$ .
2. Define  $\bar{\lambda}_c \equiv \frac{2-c^2}{c-1}$ . Then by the definition of  $\bar{\lambda}_c$  and the assumption that  $\lambda_j \geq \bar{\lambda}_c$  we have that  $\lambda_j \geq \frac{2-c^2}{c-1}$  and the same steps as in part 1 yield result (a). Now, let

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

Then, under the assumption that  $0 < \lambda_i < x(\lambda_j, c)$ , Lemma 6 implies that  $0 < \frac{\lambda_j}{\Theta_i} < 1$  and the same steps as in part 1 yield result (b) i. If we instead assume that  $\lambda_i > x(\lambda_j, c)$ , Lemma 4 and Lemma 5 imply that either  $\Theta_i \leq 0$  or  $\frac{\lambda_j}{\Theta_i} > 1$ . In both of these cases, Lemma 3 implies that  $s_i^*(s_j|\lambda_i, \lambda_j, c) > s_j$ , given that  $0 < s_j < 1$ , and this yields result (b) ii.

□

## D.5 Proof of Theorem 1

*Proof.* Observe that each partnership game is identified by a pair  $(\lambda_i, \lambda_j)$  and, by the definition of  $R_i$  (in Appendix D.3), each partnership game (and each pair  $(\lambda_i, \lambda_j)$ ) corresponds to a unique pair  $(R_i, R_j)$ . Let

$$\Lambda_{0\text{-}eff}^c \equiv \left\{ (\lambda_i, \lambda_j) \subseteq [0, \infty)^2 : R_i \cdot R_j < 1 \right\},$$

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

$$\Lambda_{2\text{-}msg}^c \equiv \left\{ (\lambda_i, \lambda_j) \subseteq [0, \infty)^2 : R_i \cdot R_j > 1 > \min(R_i, R_j) \right\} \setminus \left\{ \left( 0, \frac{1+c^4-2c^2}{c(2-c^2)} \right), \left( \frac{1+c^4-2c^2}{c(2-c^2)}, 0 \right) \right\}.$$

The two points that we removed from the set  $\Lambda_{2\text{-}msg}^c$  correspond to the cases in which (I)  $\Theta_i = \lambda_j = 0$  and (II)  $\Theta_j = \lambda_i = 0$ , in which any pair of messages induce a perfect equilibrium of the partnership game (see Lemma 3). Recall that when  $R_i = \infty$  and  $R_j = 0$ , we define  $R_i \cdot R_j$  to be equal to  $\infty$ . Note that these sets are pairwise disjoint and symmetric. We now prove each point of the theorem in turn.

1. By the definition of  $\Lambda_{0\text{-}eff}^c$ ,  $\Lambda_{max}^c$ , and  $\Lambda_{2\text{-}msg}^c$ , we have

$$Cl(\Lambda_{0\text{-}eff}^c) = \left\{ (\lambda_i, \lambda_j) \subseteq [0, \infty)^2 : R_i \cdot R_j \leq 1 \right\},$$

$$Cl(\Lambda_{max}^c) = \left\{ (\lambda_i, \lambda_j) \subseteq [0, \infty)^2 : \min(R_i, R_j) \geq 1 \right\}, \text{ and}$$

$$Cl(\Lambda_{2\text{-}msg}^c) = \left\{ (\lambda_i, \lambda_j) \subseteq [0, \infty)^2 : R_i \cdot R_j \geq 1 \geq \min(R_i, R_j) \right\}.$$

Either  $R_i \cdot R_j \leq 1$  or  $R_i \cdot R_j \geq 1$  (or both). If  $R_i \cdot R_j \geq 1$  then either  $\min(R_i, R_j) \geq 1$  or  $R_i \cdot R_j \geq 1 \geq \min(R_i, R_j)$  (or both). Therefore,

$$Cl(\Lambda_{0\text{-}eff}^c) \cup Cl(\Lambda_{max}^c) \cup Cl(\Lambda_{2\text{-}msg}^c) = [0, \infty)^2.$$

2. We demonstrate each result in turn.

- (a)  $(\lambda_i, \lambda_j) \in \Lambda_{0\text{-}eff}^c \implies R_i \cdot R_j < 1$ . If  $R_i \cdot R_j < 1$ , then, by the definition of  $R_i$  and  $R_j$ ,  $\Theta_i > 0$  and  $\Theta_j > 0$  and  $\frac{\lambda_j}{\Theta_i} \cdot \frac{\lambda_i}{\Theta_j} < 1$ . First consider the case where  $\lambda_i = \lambda_j = 0$ . Fact 1 demonstrates that  $x_i^*(\vec{s}^*) = x_j^*(\vec{s}^*) = 0$ . Next, consider  $\lambda_i > \lambda_j = 0$ . The best-reply correspondence derived in Lemma 3 implies that in this case  $s_i^* = \frac{\lambda_j}{\Theta_i} \cdot \mu_{\sigma_j} = 0$ . Substituting  $s_i^* = 0$  and  $\lambda_j = 0$  into the equilibrium effort functions given by Eq. (5) yields  $x_i^*(\vec{s}^*) = x_j^*(\vec{s}^*) = 0$ . Finally, consider  $\lambda_i, \lambda_j > 0$ . By the best-reply correspondence derived in Lemma 3, 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. Substituting these messages into Eq. (5) yields that  $x_i^*(\vec{s}^*) = x_j^*(\vec{s}^*) = 0$  in the unique subgame-perfect equilibrium. We have demonstrated that  $x_i^*(\vec{s}^*) = x_j^*(\vec{s}^*) = 0$  in every subgame-perfect equilibrium with  $(\lambda_i, \lambda_j) \in \Lambda_{0\text{-}eff}^c$ , and

that  $\lambda_k > 0$  implies that  $s_k^* = 0$ . As observed at the end of Appendix A, every trembling-hand perfect equilibrium is a subgame-perfect equilibrium, and so any trembling hand perfect equilibrium has these properties.

- (b)  $(\lambda_i, \lambda_j) \in \Lambda_{2-msg}^c \implies R_i \cdot R_j > 1 > \min(R_i, R_j)$ . Assume without loss of generality that  $R_j < R_i$ . If  $R_i \cdot R_j > 1 > R_j$  then, by the definition of  $R_i$  and  $R_j$ , either (i)  $\Theta_i < 0, \Theta_j > 0$ , and  $\frac{\lambda_i}{\Theta_j} < 1$ , 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 3 implies that equilibrium messages satisfy  $s_i^* = 1$ . If  $\lambda_j > 0$ , then by Lemma 3  $s_j^* = \frac{\lambda_j}{\Theta_j} s_i$ , and these equations are simultaneously satisfied if and only if  $1 = s_i^* > s_j^* > 0$ . In case (ii), Lemma 3 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  $s_j^* = \frac{\lambda_j}{\Theta_j} s_i$ . 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 3 implies that equilibrium messages satisfy  $s_i^* = \min\{\frac{\lambda_j}{\Theta_i} s_j, 1\}$  and  $s_j^* = \frac{\lambda_j}{\Theta_j} s_i$ . 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  $s_i^* = s_j^* = 0$ . Next we show that this latter subgame-perfect equilibrium  $((0, x_1^e(s_1, s_2)), (0, x_2^e(s_1, s_2)))$  fails to satisfy trembling-hand perfection in cases (ii) and (iii). Assume to the contrary that  $((0, x_1^e(s_1, s_2)), (0, x_2^e(s_1, s_2)))$  satisfies trembling-hand perfection. This implies that  $(0, 0)$  is the weak limit as  $\epsilon_n \rightarrow 0$  of a sequence of  $\epsilon_n$ -perfect equilibria  $(\sigma_i^n, \sigma_j^n)$  of the promise game. This implies, in particular, that for each  $\epsilon > 0$ , there exists an  $\epsilon$ -perfect equilibrium  $(\sigma_i, \sigma_j) \in \Delta^{fs}([0, 1])^2$  such that  $\sigma_i(1), \sigma_j(1) < \epsilon$ . We begin by considering case (ii). The fact that  $\sigma_j$  has full support implies that  $\mu_{\sigma_j} > 0$  and that  $BR_i^c(\sigma_j) = \{1\}$ . The definition of an  $\epsilon$ -perfect equilibrium implies that  $\sigma_i(1), \sigma_j(1) > 1 - \epsilon$ , and we get a contradiction for each  $\epsilon < 0.5$ . We are left with case (iii), in which  $R_j < R_i < \infty$  and  $R_i \cdot R_j > 1$ . The fact that  $(0, 0)$  is the weak limit of a sequence of  $\epsilon_n$ -perfect equilibria  $(\sigma_i^n, \sigma_j^n)$  when  $\epsilon_n \rightarrow 0$  implies that for each  $\epsilon > 0$ , there exists an  $\epsilon$ -perfect equilibrium  $(\sigma_i, \sigma_j) \in (\Delta^{fs}([0, 1]))^2$  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  $\sigma_j$  has full support implies that  $\mu_{\sigma_j} > 0$ . Observe that  $BR_i^c(\sigma_j) = \{R_i \cdot \mu_{\sigma_j}\}$ . The fact that  $(\sigma_i, \sigma_j)$  is an  $\epsilon$ -perfect equilibrium implies that  $\sigma_i(R_i \cdot \mu_{\sigma_j}) \geq 1 - \epsilon$ , which implies that  $\mu_{\sigma_i} \geq (1 - \epsilon) \cdot R_i \cdot \mu_{\sigma_j}$ . The fact that  $(\sigma_i, \sigma_j)$  is an  $\epsilon$ -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  $\epsilon$  that satisfies  $(1 - \epsilon)^2 \cdot R_i \cdot R_j > 1$ .

- (c)  $(\lambda_i, \lambda_j) \in \Lambda_{max}^c \implies \min(R_i, R_j) > 1$ . If  $\min(R_i, R_j) > 1$ , 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 3, 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 3 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 3 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 3 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 3 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 3 implies that equilibrium messages satisfy  $s_i^* = 1$  and  $s_j^* = 1$ , which implies that  $s_i^* = s_j^* = 1$ .

This implies that in all six cases (i, ii, iii, iv, v, and vi) the strategy profile

$((1, x_1^e(s_1, s_2)), (1, x_2^e(s_1, s_2)))$  is a subgame-perfect equilibrium. It is unique (and thus satisfies trembling-hand perfection) in cases (ii), (iii), (v), and (vi). In cases (i) and (iv), the strategy profile  $((0, x_1^e(s_1, s_2)), (0, x_2^e(s_1, s_2)))$  is the only additional subgame-perfect equilibrium.

Finally, we have to show that the additional equilibrium  $((0, x_1^e(s_1, s_2)), (0, x_2^e(s_1, s_2)))$  fails to satisfy trembling-hand perfection in cases (i) and (iv). The proof of this claim in case (i) is completely analogous to the proof that  $((0, x_1^e(s_1, s_2)), (0, x_2^e(s_1, s_2)))$  fails to satisfy trembling-hand perfection in case (iii) of part (b) above, and the proof of this claim in case (iv) is completely analogous to the proof that  $((0, x_1^e(s_1, s_2)), (0, x_2^e(s_1, s_2)))$  fails to satisfy trembling-hand perfection in case (ii) of part (b) above.

3. Lemma 6 says that if  $\lambda_j \geq \frac{2-c^2}{c-1}$  then  $0 < \frac{\lambda_j}{\Theta_i} < 1$ , which implies that  $R_i < 1$ . Therefore, if  $c \geq \sqrt{2} \implies \frac{2-c^2}{c-1} \leq 0$ , then  $R_i \cdot R_j < 1$  for all  $(\lambda_i, \lambda_j)$  and so  $\Lambda_{0-eff}^c = [0, \infty)^2$ .
4. Let  $x_{\lambda_i}^c = \max \left\{ \frac{1}{(c + \lambda_j)(2 - c(c + \lambda_j))} - c, \frac{2-c^2}{c-1} \right\}$ . Given that  $c < \sqrt{2}$ , there exists  $0 \leq \lambda_j < \frac{2}{c} - c$ . Lemma 4 says that if  $\lambda_j < \frac{2}{c} - c$  and  $\lambda_i > x_{\lambda_i}^c$  then  $\Theta_i \leq 0$ , and hence  $R_i = \infty$  (using the first part of the maximum function defining  $x_{\lambda_i}^c$ ). Given that  $\lambda_i > \frac{2-c^2}{c-1} > 0$ , we have by Lemma 6 that  $1 > R_j > 0$  and so  $R_i \cdot R_j > 1 > R_j$  and so  $(\lambda_i, \lambda_j) \in \Lambda_{2-msg}^c$ .
5. We prove each point in turn. (I) The proof proceeds by showing that the set  $\Lambda_{max}^c$  is a convex set. Convexity and symmetry then imply that if  $\Lambda_{max}^c$  is nonempty then there is a  $\lambda$  such that  $(\lambda, \lambda) \in \Lambda_{max}^c$ . We then show that such a  $\lambda$  exists only if  $c < 1.25$ . By the definition of  $R_i$ , we recall that  $R_i \geq 1$  if and only if (1)  $\Theta_i \leq 0$  or (2)  $\Theta_i > 0$  and  $\frac{\lambda_i}{\Theta_i} \geq 1$ . We can recall from Lemma 4 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 5 that  $\Theta_i > 0$  and  $\frac{\lambda_j}{\Theta_i} \geq 1$  if and only if

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

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

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

$$\lambda_i \geq \frac{1}{\lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)} - c \quad and \quad \lambda_j < \frac{2-c^2}{c-1}. \quad (25)$$

We will now show that the set of points that satisfy Eq. (25) is convex. First, observe that the second derivative of the right-hand side of the first inequality of Eq. (25) (the lower bound on  $\lambda_i$ ) with respect to  $\lambda_j$  is

$$\frac{2[3c^4 + (6\lambda_j - 3)c^3 + (3\lambda_j^2 - 9\lambda_j - 5)c^2 + 3\lambda_j^2 + 6\lambda_j + 4]}{(\lambda_j + c)[2 - c^2 - \lambda_j(c - 1)]}. \quad (26)$$

The numerator of this expression is positive for all  $\lambda_j > 0$  and<sup>16</sup>  $c > 1$ . 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. (25). Therefore, the set of points that satisfy Eq. (25) 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  $\min(R_i, R_j) > 1$  ( $\Lambda_{max}^c$ ) is convex.

We now establish the interval of  $c$  in which  $\Lambda_{max}^c$  is nonempty. By the convexity and symmetry of  $\Lambda_{max}^c$ , if this set is nonempty there must be a maximum and a minimum  $\lambda$  such that  $(\lambda, \lambda) \in Cl(\Lambda_{max}^c)$ . We now show that such maximum and minimum elements exist if and only if  $c < 1.25$ . Clearly, the maximum and minimum  $\lambda$  such that  $(\lambda, \lambda) \in Cl(\Lambda_{max}^c)$  are the largest and smallest values of  $\lambda$  such that the weak counterpart of Eq. (25) holds when  $\lambda_i = \lambda_j = \lambda$ . Given that  $Cl(\Lambda_{max}^c)$  is convex and closed, these maximum and minimum values must obtain when at least one of the inequalities in Eq. (25) holds with equality. To find the maximum and minimum values of  $\lambda$  that satisfy the first inequality in Eq. (25), we solve the corresponding equation when  $\lambda_i = \lambda_j = \lambda$ . We then show that these are the largest and smallest values satisfying both

<sup>16</sup>Eq. (26) and the conditions for the positive numerator are derived using Mathematica. The code is available in the supplementary appendix of this paper.

inequalities simultaneously. Imposing  $\lambda_i = \lambda_j = \lambda$  on the first inequality in Eq. (25), we obtain

$$\lambda = \frac{1}{\lambda^2(1-c) + \lambda(2-2c^2+c) + c(2-c^2)} - c. \quad (27)$$

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

$$\lambda^3[1-c] + \lambda^2[2+2c-3c^2] + \lambda[4c-3c^3+c^2] - [c^2-1]^2 = 0. \quad (28)$$

Eq. (28) has two solutions when  $\lambda$  is positive:

$$\lambda = \frac{1+2c-2c^2}{2(c-1)} - \frac{\sqrt{5-4c}}{2(c-1)} \equiv \lambda_c^- \quad (29)$$

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

Clearly, these two solutions are defined if and only if  $c < 1.25$ . By inspection of Eq. (29) and Eq. (30), it is straightforward to see that for all  $1 < c < 1.25$ ,  $0 < \lambda_c^- < \lambda_c^+ < \infty$ . To see point (II) we then simply note that by the definition of  $R_k$ ,  $\lambda_i = \lambda_j$  implies that  $R_i > 1 \iff R_j > 1$  and either  $\min(R_i, R_j) \geq 1$  or  $R_i, R_j < 1 \implies R_i \cdot R_j < 1 \implies (\lambda, \lambda) \in \Lambda_{0-eff}^c$ . Therefore, given that  $\Lambda_{max}^c = \emptyset$  when  $c > 1.25$ , we have that  $(\lambda, \lambda) \in \Lambda_{0-eff}^c$  for  $\lambda \geq 0$ . To see (III) first note that Lemma 6 implies that if  $\lambda_i, \lambda_j \geq \frac{2-c^2}{c-1}$  then  $R_i \cdot R_j < 1$  and  $(\lambda_i, \lambda_j) \in \Lambda_{0-eff}^c$ . It is straightforward to see that  $\frac{2-c^2}{c-1}$  is decreasing for  $c > 1$  and obtains the value 1.75 when  $c = 1.25$ . Therefore  $\lambda_i, \lambda_j > 1.75 \implies (\lambda_i, \lambda_j) \in \Lambda_{0-eff}^c$  for  $c \in (1.25, \sqrt{2})$ . Next, note that Lemma 6 implies that if  $\lambda_i < \frac{1}{\lambda_j^2(1-c) + \lambda_j(2-2c^2+c) + c(2-c^2)} - c$  and the expression holds also with  $i$  and  $j$  interchanged then  $R_i \cdot R_j < 1$  and  $(\lambda_i, \lambda_j) \in \Lambda_{0-eff}^c$ . We now derive a lower bound on the value taken by the right-hand side of this inequality. The second derivative of this expression with respect to  $\lambda_j$  is strictly positive for all  $\lambda_j < \frac{2-c^2}{c-1}$  for all  $c \in (1.25, \sqrt{2})$  and hence this function is strictly convex in  $\lambda_j$  and achieves at most its global minimum over this interval<sup>17</sup>  $0 \leq \lambda_j < \frac{2-c^2}{c-1}$ . We next derive this minimum value by setting the first derivative of this function with respect to  $\lambda_j$  equal to zero, solving for  $\lambda_j$ . and substituting this into the function. This yields a lower bound as a function of  $c$ :

$$\frac{4c^2 - c^3 - 4}{(c-2)^2}. \quad (31)$$

The first derivative of this expression with respect to  $c$  is

$$\frac{8 - 16c + 6c^2 - c^3}{(c-2)^3}.$$

---

<sup>17</sup>This fact is proven using Mathematica. The code used to prove this fact is available in the online appendix.

This expression is positive for all<sup>18</sup>  $c \in (1.25, \sqrt{2})$ . Therefore, the lower bound takes its lowest value (with respect to  $c$ ) when  $c$  takes its lowest value, i.e.,  $c = 1.25$ . Evaluating Eq. (31) at  $c = 1.25$  gives a value of  $\frac{19}{36} \approx 0.53$ . Therefore,  $\lambda_i, \lambda_j < \frac{19}{36} \approx 0.53 \implies (\lambda_i, \lambda_j) \in \Lambda_{0-eff}^c$  for  $c \in (1.25, \sqrt{2})$ .

6. Let  $\lambda_c^- = \frac{1+2c-2c^2}{2(c-1)} - \frac{\sqrt{5-4c}}{2(c-1)}$  and  $\lambda_c^+ = \frac{1+2c-2c^2}{2(c-1)} + \frac{\sqrt{5-4c}}{2(c-1)}$ . Then we can see that points (I) and (II) are proven in the proof of the previous part (part 5). To see (III), note first that by the definition of  $\lambda_c^-$  and  $\lambda_c^+$  as the minimum and maximum  $\lambda$  (respectively) such that  $(\lambda, \lambda) \in Cl(\Lambda_{max}^c)$  and by the convexity of  $\Lambda_{max}^c$ , we have that  $(\lambda, \lambda) \notin Cl(\Lambda_{max}^c)$  for each  $\lambda \in [0, \lambda_c^-) \cup (\lambda_c^+, \infty)$ . Assume that there exist  $\lambda_i, \lambda_j < \lambda_c^-$  such that  $(\lambda_i, \lambda_j) \in \Lambda_{max}^c$ . By the symmetry of  $\Lambda_{max}^c$ , we have that  $(\lambda_j, \lambda_i) \in \Lambda_{max}^c$ . Let  $\lambda_k = \frac{\lambda_i + \lambda_j}{2} < \lambda_c^-$ . By the convexity of  $\Lambda_{max}^c$ , we have that  $(\lambda_k, \lambda_k) \in \Lambda_{max}^c$ , which is a contradiction. This establishes that  $(\lambda_i, \lambda_j) \in \Lambda_{max}^c$  implies that  $\lambda_c^- \leq \max(\lambda_i, \lambda_j)$ . By assuming that there exist  $\lambda_i, \lambda_j > \lambda_c^+$  such that  $(\lambda_i, \lambda_j) \in \Lambda_{max}^c$ , we can derive a contradiction in an analogous way. Finally, we consider the case where  $\lambda_c^- < \lambda_i \leq \lambda_c^+ \leq \lambda_j$ . We have established in the proof of the previous part (part 5) that the right-hand side of the first inequality in Eq. (25) (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  $\lambda_j$  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  $\max(\lambda_i, \lambda_j) < \lambda_c^+$ . To see (IV), let  $\bar{\lambda}_c = \frac{2-c^2}{c-1}$  and observe that Lemma 6 implies that if  $\lambda_i, \lambda_j > \bar{\lambda}_c$  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  $(\lambda_i, \lambda_j) \in \Lambda_{0-eff}^c$ . Let

$$Z_c \equiv \min_{\lambda < \frac{2-c^2}{c-1}} \left\{ \frac{1}{\lambda^2(1-c) + \lambda(2-2c^2+c) + c(2-c^2)} - c \right\} = \frac{4c^2 - c^3 - 4}{(c-2)^2},$$

where the equality is derived in the proof of part 5 (III) of the theorem (Eq. (31)). Assume that  $Z_c > 0$  and let  $\underline{\lambda}_c = Z_c$ . By equivalent arguments to those in the proof of part 5 (III), we then have that  $\lambda_i, \lambda_j < \underline{\lambda}_c \implies (\lambda_i, \lambda_j) \in \Lambda_{0-eff}^c$ . Assume conversely that  $Z_c \leq 0$  and let  $\underline{\lambda}_c = \lambda_c^-$ . Observe that given the strict convexity of the right-hand side of the first inequality in Eq. (25) and the definition of  $\lambda_c^-$  as the first of two points at which this function crosses the 45 degree line, we must have that this function is strictly decreasing for  $\lambda_j < \lambda_c^-$  and hence  $\lambda_i, \lambda_j < \lambda_c^-$  implies that  $R_i < 1$ . By the symmetry of the condition defining  $R_j < 1$  we have that  $\lambda_i, \lambda_j < \lambda_c^-$  implies that  $R_j < 1$  and hence  $\lambda_i, \lambda_j < \underline{\lambda}_c$  implies that  $R_i \cdot R_j < 1$  and hence  $(\lambda_i, \lambda_j) \in \Lambda_{0-eff}^c$ . Therefore, given that  $0 < \lambda_c^-$  for  $c \in (1, 1.25)$ , letting  $\underline{\lambda}_c = Z_c$  if  $Z_c > 0$  and  $\underline{\lambda}_c = \lambda_c^-$  if  $Z_c \leq 0$ , we have that  $\lambda_i, \lambda_j < \underline{\lambda}_c$  implies that  $(\lambda_i, \lambda_j) \in \Lambda_{0-eff}^c$ . □

---

<sup>18</sup>This fact is proven using Mathematica. The code used to prove this fact is available in the online appendix.

## D.6 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  $\Lambda_{2-msg}^c$ ) 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  $\Theta_i$ , 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  $s_i = s_j$ .  $\square$

## D.7 Corollary of Theorem 1 and Lemma 3

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  $\lambda_c^- = \min\{\lambda : (\lambda, \lambda) \in Cl(\Lambda_{max}^c)\}$  and  $\lambda_c^+ = \max\{\lambda : (\lambda, \lambda) \in Cl(\Lambda_{max}^c)\}$ ), 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 5. 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  $\left((s, s'), (x_i^*(\vec{s}), x_j^*(\vec{s}))\right)$  is a perfect equilibrium of the partnership game if and only if  $s = s'$ . (2) If  $\left((1, 1), (x_i^*(\vec{s}), x_j^*(\vec{s}))\right)$  is a perfect equilibrium of the partnership game then  $(\lambda, \lambda) \in Cl(\Lambda_{max}^c)$ .*

*Proof.* Part 1: The proof of part 5 of Theorem 1 demonstrates that by the definitions  $\lambda_c^- = \min\{\lambda : (\lambda, \lambda) \in Cl(\Lambda_{max}^c)\}$  and  $\lambda_c^+ = \max\{\lambda : (\lambda, \lambda) \in Cl(\Lambda_{max}^c)\}$ , 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 3 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  $\left((s^*, s^*), (x_i^*(\vec{s}), x_j^*(\vec{s}))\right)$  is a trembling-hand perfect equilibrium, observe that for each  $\epsilon > 0$  there exists an  $\epsilon$ -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  $\lambda_c^-$  and  $\lambda_c^+$  ( $\lambda_c^- = \min\{\lambda : (\lambda, \lambda) \in Cl(\Lambda_{max}^c)\}$  and  $\lambda_c^+ = \max\{\lambda : (\lambda, \lambda) \in Cl(\Lambda_{max}^c)\}$ ) and from the fact that  $\Lambda_{max}^c$  is a convex set (part 6 (I) of Theorem 1) and so its closure is too.  $\square$

## D.8 Proof of Theorem 2

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

### D.8.1 Lemma: Positive Payoff Always Possible in the Population Game

**Lemma 7.** 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  $(\lambda_i, \lambda_j)$ , player  $i$  achieves a strictly positive material payoff, i.e.,  $\pi(\lambda_i, \lambda_j) > 0$ .

*Proof.* Theorem 1 and the definition of  $\Lambda_{max}^c$  and  $\Lambda_{2-msg}^c$  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  $\Theta_j$  to write the condition  $\Theta_j < 0$  when  $\lambda_i = 0$ , and rearrange it to yield a lower bound on  $\lambda_j$ :

$$c^2 + \frac{1}{(c + \lambda_j)c} - 2 < 0 \iff \frac{1}{c + \lambda_j} < c(2 - c^2) \iff \frac{1}{c(2 - c^2)} - c < \lambda_j. \quad (32)$$

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

$$\frac{3c^2 - 2}{(2c - c^3)^2} - 1. \quad (33)$$

Eq. (33) is positive for  $c < 1.25$ . The lower bound on  $\lambda_j$  given by Eq. (32) 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  $\lambda_i$  sufficiently large, either  $\Theta_i \leq 0$  or  $R_i \cdot R_j > 1$ . We take the limit of  $\Theta_i$  as  $\lambda_i \rightarrow \infty$  and find the conditions under which this is negative:

$$\lim_{\lambda_i \rightarrow \infty} \Theta_i \leq 0 \iff c(c + \lambda_j) - 2 \leq 0 \iff \lambda_j \leq \frac{2}{c} - c. \quad (34)$$

Next, we check the condition for satisfying  $\frac{\lambda_i \cdot \lambda_j}{\Theta_i \cdot \Theta_j} > 1$ , 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  $\lambda_i \rightarrow \infty$ :

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

where the second equality is derived from neglecting the term  $\frac{1}{(c + \lambda_i)(c + \lambda_j)}$ , which converges to zero as  $\lambda_i \rightarrow \infty$ , 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  $\lambda_i \rightarrow \infty$ , in the second denominator.

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

$$\begin{aligned} \frac{\lambda_j}{c[c(c + \lambda_j) - 2]} > 1 &\iff c(c + \lambda_j) - 2 > 0 \quad \text{and} \quad \lambda_j < c[c(c + \lambda_j) - 2] \\ &\iff \frac{2}{c} - c < \lambda_j < \frac{c}{c^2 - 1} - c. \end{aligned} \quad (35)$$

Observe that the first inequality in Eq. (35) holds precisely when Eq. (34) does not hold. The first derivative of the right-hand side of the second inequality in Eq. (35) 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. (35) is  $\frac{35}{36} > \frac{81}{140}$ . Therefore, for all  $c < 1.25$  and  $\lambda_j \leq \frac{81}{140}$ , this second inequality holds. We therefore have that for all  $c < 1.25$  and  $\lambda_j \leq \frac{81}{140}$ , either  $\Theta_i \leq 0$  or  $\frac{\lambda_i \lambda_j}{\Theta_j \cdot \Theta_i} > 1$ , when  $\lambda_i$  is sufficiently high. Therefore, for all  $c < 1.25$  and for all  $\lambda_j \geq 0$ , there exists a  $\lambda_i \geq 0$  such that either  $s_i = 1$  or  $s_j = 1$  in the unique equilibrium of the game  $(\lambda_i, \lambda_j)$  (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 3, 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.  $\square$

### D.8.2 Lemma: Additional Properties of $\Lambda_{max}^c$

**Lemma 8.** Fix  $c \in (1, 1.24)$ . (1) For all  $\lambda \in [\lambda_c^-, \lambda_c^+)$ , there exists  $\delta_\lambda > 0$  such that for all  $\lambda' \in (\lambda, \lambda + \delta_\lambda)$ ,  $(\lambda', \lambda) \in \Lambda_{max}^c$  (2) For all  $\lambda' \neq \lambda_c^+$ ,  $(\lambda', \lambda_c^+) \notin Cl(\Lambda_{max}^c)$ .

*Proof.* The proof of Theorem 1 yields Eq. (25) and the corresponding condition for player  $j$ , which together define  $\Lambda_{max}^c$ . The strict convexity of the first inequality of Eq. (25) defining the boundary of  $\Lambda_{max}^c$ , implies that for all  $\lambda \in (\lambda_c^-, \lambda_c^+)$ ,  $(\lambda, \lambda)$  is not on the boundary of  $\Lambda_{max}^c$  and is therefore in the interior of  $\Lambda_{max}^c$  (i.e., it is in  $\Lambda_{max}^c$  but not in  $Cl(\Lambda_{max}^c)$ ). 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  $\lambda' \in [\lambda, \lambda + \delta_\lambda)$ ,  $(\lambda', \lambda) \in \Lambda_{max}^c$ . We next show that there exists  $\delta_\lambda > 0$  such that for all  $\lambda' \in (\lambda_c^-, \lambda_c^- + \delta_\lambda)$ ,  $(\lambda', \lambda_c^-) \in \Lambda_{max}^c$ . This will be the case if and only if there is  $\delta_\lambda > 0$  such that Eq. (25) 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. (25) does not become “tighter” as  $\lambda_j$  increases, i.e., if and only if the derivative of the right-hand side of the first inequality of Eq. (25) 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. (25) with

respect to  $\lambda_j$  is

$$\frac{c(2c + 2\lambda_j - 1) - 2(1 + \lambda_j)}{(\lambda_j + c)^2[\lambda_j(c - 1) + c^2 - 2]^2}. \quad (36)$$

When evaluated at  $\lambda_j = \lambda_c^-$ , Eq. (36) is nonpositive if<sup>19</sup>  $c > \sqrt{5} - 1 \approx 1.24$ . Therefore, we have that for all  $\lambda \in [\lambda_c^-, \lambda_c^+)$ , there exists  $\delta_\lambda > 0$  such that for all  $\lambda' \in (\lambda, \lambda + \delta_\lambda)$ ,  $(\lambda', \lambda) \in \Lambda_{max}^c$ .

Point (2) is established by noting first that the right-hand side of the first inequality of Eq. (25) must be increasing in  $\lambda_j$  when evaluated at  $\lambda_c^+$  (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  $\lambda_c^-$ ). Therefore, given that  $\lambda_c^+$  satisfies Eq. (27), an increase in  $\lambda_j$  with  $\lambda_i$  fixed at  $\lambda_c^+$  means that the weak counterpart of Eq. (25) does not hold. By symmetry, an increase in  $\lambda_i$  with  $\lambda_j$  fixed at  $\lambda_c^+$  means that the equivalent condition on  $\lambda_j$  is violated. Secondly, it is straightforward to see that given that  $\lambda_c^+$  satisfies Eq. (27), when  $\lambda_j$  is fixed at  $\lambda_c^+$ , any  $\lambda_i < \lambda_c^+$  must violate the weak counterpart of the first inequality in Eq. (25). Therefore, for any  $\lambda' \neq \lambda_c^+$ ,  $\min(R_i, R_j) < 1$  and so  $(\lambda', \lambda_c^+) \notin Cl(\Lambda_{max}^c)$ .  $\square$

### D.8.3 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  $(\lambda_i, \lambda_j) \in \Lambda_{0-eff}^c$  or  $\min\{R_i, R_j\} \geq 1$  and hence  $(\lambda_i, \lambda_j) \in Cl(\Lambda_{max}^c)$ . Therefore,  $\lambda_i = \lambda_j \implies (\lambda_i, \lambda_j) \in Cl(\Lambda_{max}^c) \cup \Lambda_{0-eff}^c$ . For any  $\lambda$  such that  $(\lambda, \lambda) \in \Lambda_{0-eff}^c$ ,  $x_i = x_j = 0$  and so  $\pi(\lambda, \lambda) = 0$ . To find the material payoff when  $(\lambda, \lambda) \in Cl(\Lambda_{max}^c)$ , we recall Eq. (1) for material payoff, and impose  $x_i = x_j = x$  on the equation, which yields

$$\pi(\lambda, \lambda) = x^2 - \frac{cx^2}{2}, \quad (37)$$

which is clearly positive and increasing in  $x$  for all  $c < 2$ . The level of renegeing aversion that maximises the material payoff in a symmetric game is therefore that which maximises equilibrium effort. Theorem 1 tells us that for  $(\lambda, \lambda) \in \Lambda_{max}$ ,  $s_i = s_j = 1$  in the unique equilibrium. For  $(\lambda, \lambda) \in Cl(\Lambda_{max}) \setminus \Lambda_{max}$ , Corollary 2 tells us that any  $s_i = s_j \in [0, 1]$  is supported as an equilibrium and by our equilibrium selection assumption we have that  $s_i = s_j = 1$ . We therefore impose  $s_i = s_j = 1$  and  $\lambda_i = \lambda_j = \lambda$  on the equation for equilibrium effort (Eq. 5):

$$\frac{(c + \lambda)\lambda + \lambda}{(c + \lambda)(c + \lambda) - 1} = \frac{\lambda}{c + \lambda - 1}. \quad (38)$$

The derivative of this expression with respect to  $\lambda$  is

$$\frac{(c + \lambda - 1) - \lambda}{[c + \lambda - 1]^2} = \frac{(c - 1)}{[c + \lambda - 1]^2}, \quad (39)$$

which is clearly positive for all  $c > 1$ . Therefore, the renegeing cost that maximises effort, and

---

<sup>19</sup>This final result is obtained using Mathematica. The code is available in the supplementary appendix of this paper.

therefore maximises the material payoff, in a symmetric game is the highest  $\lambda$  such that  $(\lambda, \lambda) \in Cl(\Lambda_{max}^c)$ . By definition, this is  $\lambda_c^+$ .

2. Recall that

$$\lambda_c^+ = \frac{1 + 2c - 2c^2}{2(c-1)} + \frac{\sqrt{5-4c}}{2(c-1)} = \frac{1 + 2c(1-c) + \sqrt{5-4c}}{2(c-1)}. \quad (40)$$

Note that as  $c \rightarrow 1$ , the numerator of Eq. (40) is increasing and the denominator of Eq. (40) converges to zero. Hence  $\lim_{c \rightarrow 1} \lambda_c^+ = \infty$ . To find the limit of the players' material payoff in the game  $(\lambda_c^+, \lambda_c^+)$  as  $c \rightarrow 1$ , we substitute the expression for effort in a maximum message equilibrium (Eq. 38) into that for material payoff in a symmetric equilibrium (Eq. (37)) when  $\lambda = \lambda_c^+$ :

$$\pi(\lambda_c^+, \lambda_c^+) = \left[ \frac{\lambda_c^+}{c + \lambda_c^+ - 1} \right]^2 \left[ 1 - \frac{c}{2} \right]. \quad (41)$$

As  $c \rightarrow 1$ ,  $\lambda_c^+ \rightarrow \infty$  and therefore the limit of Eq. (41) is given by

$$\lim_{c \rightarrow 1} \pi(\lambda_c^+, \lambda_c^+) = \lim_{c \rightarrow 1} \left[ \frac{\lambda_c^+}{c + \lambda_c^+ - 1} \right]^2 \left[ 1 - \frac{c}{2} \right] = \left( 1 - \frac{1}{2} \right) = \frac{1}{2}. \quad (42)$$

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.,  $\pi(\lambda', \lambda_c^+) < \pi(\lambda_c^+, \lambda_c^+)$  for  $\lambda' \in [0, \lambda_c^+)$ . Point (2) of Lemma 8 implies that for all  $\lambda' \in [0, \lambda_c^+)$ ,  $(\lambda', \lambda_c^+) \notin Cl(\Lambda_{max}^c)$ . Therefore for all such deviations,  $(\lambda', \lambda_c^+) \in Cl(\Lambda_{2-msg}^c)$  or  $(\lambda', \lambda_c^+) \in Cl(\Lambda_{0-eff}^c)$ . Suppose first that  $(\lambda', \lambda_c^+) \in \Lambda_{0-eff}^c$ . Then the effort levels of both players are zero and so we have  $\pi(\lambda_c^+, \lambda_c^+) > \pi(\lambda', \lambda_c^+) = 0$ . Suppose instead that  $(\lambda', \lambda_c^+) \in \Lambda_{2-msg}^c$ ; 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  $s_i = \frac{\lambda_j}{\Theta_i}$  and  $s_j = 1$  and  $\lambda_j = \lambda_c^+$  (player  $i$  is therefore the deviating player):

$$\pi_i(\lambda_i, \lambda_c^+) = \frac{[(c + \lambda_c^+) \lambda_i \frac{\lambda_j}{\Theta_i} + \lambda_c^+][(c + \lambda_i) \lambda_c^+ + \lambda_i \frac{\lambda_j}{\Theta_i}]}{[(c + \lambda_i)(c + \lambda_c^+) - 1]^2} - \frac{c[(c + \lambda_c^+) \lambda_i \frac{\lambda_j}{\Theta_i} + \lambda_c^+]^2}{2[(c + \lambda_i)(c + \lambda_c^+) - 1]^2}. \quad (43)$$

The derivative of this expression with respect to  $\lambda_i$  is<sup>20</sup>

$$\frac{[\lambda_c^+]^2 (c(\lambda_c^+ + c) - 1)^2}{[1 + (c + \lambda_c^+)(c + \lambda_i)(c(\lambda_c^+ + c) - 2)]^3}. \quad (44)$$

Clearly, the numerator of Eq. (44) is always positive. A *sufficient* condition for the denominator, and hence for the whole expression, to be strictly positive is that

$$c(\lambda_c^+ + c) - 2 > 0 \iff \lambda_c^+ > \frac{2}{c} - c. \quad (45)$$

---

<sup>20</sup>This derivative was calculated using Mathematica. The code is available in the supplementary appendix of this paper.

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,  $\pi(\lambda_c^+, \lambda_c^+) > \pi(\lambda', \lambda_c^+)$  for any  $\lambda' \in [0, \lambda_c^+)$  such that  $(\lambda', \lambda_c^+) \in \Lambda_{2-msg}^c$ . Finally, suppose that  $(\lambda', \lambda_c^+) \in \{Cl(\Lambda_{2-msg}) \setminus \Lambda_{2-msg}\} = \{(\lambda_i, \lambda_c^+) \subseteq [0, \infty)^2: R_i \cdot R_j = 1 > R_i\}$ . By Lemma 3 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):

$$\pi_i(\lambda_i, \lambda_c^+) = \left( \frac{[(c + \lambda_c^+) \lambda_i \frac{\lambda_j}{\Theta_i} + \lambda_c^+][(c + \lambda_i) \lambda_c^+ + \lambda_i \frac{\lambda_j}{\Theta_i}]}{[(c + \lambda_i)(c + \lambda_c^+) - 1]^2} - \frac{c[(c + \lambda_c^+) \lambda_i \frac{\lambda_j}{\Theta_i} + \lambda_c^+]^2}{2[(c + \lambda_i)(c + \lambda_c^+) - 1]^2} \right) s_j^2. \quad (46)$$

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. (43) and, by the above arguments, it cannot represent a profitable deviation. We therefore have that  $\pi(\lambda', \lambda_c^+) < \pi(\lambda_c^+, \lambda_c^+)$  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.,  $\pi(\lambda', \lambda_c^+) < \pi(\lambda_c^+, \lambda_c^+)$  for  $\lambda' > \lambda_c^+$ . By Lemma 8,  $\lambda' > \lambda_c^+$  implies that  $(\lambda', \lambda_c^+) \notin Cl(\Lambda_{max}^c)$ . Suppose first that  $(\lambda', \lambda_c^+) \in \Lambda_{0-eff}^c$ . In this case, the effort levels of both players are zero and so we have  $\pi(\lambda_c^+, \lambda_c^+) > \pi(\lambda', \lambda_c^+) = 0$ . Suppose instead that  $(\lambda', \lambda_c^+) \in \Lambda_{2-msg}^c$ . 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):

$$\pi_i(\lambda_i, \lambda_c^+) = \frac{[(c + \lambda_c^+) \lambda_i + \lambda_c^+ \frac{\lambda_i}{\Theta_j}][(c + \lambda_i) \lambda_c^+ \frac{\lambda_i}{\Theta_j} + \lambda_i]}{[(c + \lambda_i)(c + \lambda_c^+) - 1]^2} - \frac{c[(c + \lambda_c^+) \lambda_i + \lambda_c^+ \frac{\lambda_i}{\Theta_j}]^2}{2[(c + \lambda_i)(c + \lambda_c^+) - 1]^2}. \quad (47)$$

In the supplementary appendix of this paper, we present the explicit formula for the derivative of Eq. (47) with respect to  $\lambda_i$  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  $(\lambda_i, \lambda_c^+) \in \Lambda_{2-msg}^c$ , we have that  $\pi(\lambda_i, \lambda_c^+) < \pi(\lambda_c^+, \lambda_c^+)$ . Finally, suppose that  $(\lambda', \lambda_c^+) \in \{Cl(\Lambda_{2-msg}) \setminus \Lambda_{2-msg}\} = \{(\lambda_i, \lambda_c^+) \subseteq [0, \infty)^2: R_i \cdot R_j = 1 > R_j\}$ . 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. (47) and therefore not profitable. Hence for any  $\lambda_i > \lambda_c^+$ ,  $\pi(\lambda_i, \lambda_c^+) < \pi(\lambda_c^+, \lambda_c^+)$ .

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 renegeing costs) let player  $i$  make his effort choice first with player  $j$  best-replying to this. Then, in equilibrium,  $x_j = \underset{x_j}{\operatorname{argmax}} \left\{ x_i x_j - \frac{c \cdot x_j^2}{2} \right\} = \frac{x_j}{c}$  and hence  $x_i = \underset{x_i}{\operatorname{argmax}} \left\{ \frac{x_i^2}{c} - \frac{c \cdot x_i^2}{2} \right\} = \underset{x_i}{\operatorname{argmax}} \left\{ \frac{(2-c^2)x_i}{2c} \right\} = 1$ , 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  $\frac{1}{c} - \frac{c}{4} \left(1 + \frac{1}{c^2}\right) = \frac{3-c^2}{4c}$ . The payoff to either player in the equilibrium induced by  $(\lambda_c^+, \lambda_c^+)$  is given by Eq. (41). We then have that

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

The final step is proven using Mathematica.<sup>21</sup>

5. Recall from part 1 that  $\lambda_i = \lambda_j \implies (\lambda_i, \lambda_j) \in Cl(\Lambda_{max}^c) \cup \Lambda_{0-eff}^c$ . 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  $\lambda$  such that the unique equilibrium in the corresponding partnership game  $(\lambda, \lambda) \in \Lambda_{0-eff}$ , we have that  $\pi(\lambda, \lambda) = 0$ . Lemma 7 shows that for  $c < 1.25$  and for  $\lambda \geq 0$ , there exists  $\lambda' \geq 0$  such that  $\pi(\lambda', \lambda) > 0$ . Therefore, for all  $\lambda$  such that  $\pi(\lambda, \lambda) = 0$ ,  $(\lambda, \lambda)$  cannot be a Nash equilibrium of the population game. For any  $\lambda$  such that  $(\lambda, \lambda) \in Cl(\Lambda_{max}^c)$ , we say that such an equilibrium “admits an upward deviation within  $\Lambda_{max}^c$ ” if there exists  $\delta_\lambda > 0$  such that for all  $\lambda' \in (\lambda, \lambda + \delta_\lambda)$ ,  $(\lambda', \lambda) \in \Lambda_{max}^c$ . For all  $\lambda$  such that  $(\lambda, \lambda) \in Cl(\Lambda_{max}^c)$ , the equilibrium payoff to both players is obtained by substituting  $s_i = s_j = 1$  into Eq. (11):

$$\pi_i(\lambda_i, \lambda_j) = \frac{[(c + \lambda_j)\lambda_i + \lambda_j][(c + \lambda_i)\lambda_j + \lambda_i]}{[(c + \lambda_i)(c + \lambda_j) - 1]^2} - \frac{c[(c + \lambda_j)\lambda_i + \lambda_j]^2}{2[(c + \lambda_i)(c + \lambda_j) - 1]^2}. \quad (48)$$

The first derivative of this function with respect to  $\lambda_i$  is

$$\frac{(c-1)(1+c+\lambda_j)[\lambda_i c^3 + 2c^2 \lambda_i \lambda_j + c \lambda_i (\lambda_j^2 - \lambda_j - 2) - \lambda_j (1 + \lambda_i (2 + \lambda_j))]}{[c^2 - 1 + \lambda_i \lambda_j + c(\lambda_i + \lambda_j)]^3}. \quad (49)$$

Imposing the condition  $\lambda_i = \lambda_j = \lambda$ , we can simplify this expression to<sup>22</sup>

$$\frac{(c-1)[c(c+\lambda-1) - 1 - \lambda]\lambda}{[c+1+\lambda][c-1+\lambda]^3}. \quad (50)$$

<sup>21</sup>The code available in the supplementary appendix of this paper.

<sup>22</sup>The derivative given by Eq. (49) and its simplification when  $\lambda_i = \lambda_j$  is obtained using Mathematica. The code available in the supplementary appendix of this paper.

This expression is strictly positive if and only if

$$c(c + \lambda - 1) - 1 - \lambda > 0 \iff \lambda < \frac{1 + c - c^2}{c - 1}. \quad (51)$$

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 4 and Lemma 5 each of these conditions implies that

$$\lambda_j < \frac{2 - c^2}{c - 1}. \quad (52)$$

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

$$\lambda < \frac{2 - c^2}{c - 1} < \frac{1 + c - c^2}{c - 1}, \quad (53)$$

where the second inequality clearly follows when  $c > 1$ . We can see that this yields the second inequality in Eq. (51) and hence Eq. (50) is always positive in a maximum-message equilibrium. Therefore, for any  $\lambda$  such that  $(\lambda, \lambda)$  “admits an upward deviation within  $\Lambda_{max}^c$ ,” there exists some  $\lambda' > \lambda$  such that  $\pi(\lambda', \lambda) > \pi(\lambda, \lambda)$  and hence no such strategy profile is a Nash equilibrium of the population game. We have shown that the only potential symmetric pure Nash equilibria of the population game are those that admit a maximum-message equilibrium and do not “admit an upward deviation within  $\Lambda_{max}^c$ .” Lemma 8 implies that there is a unique pair  $(\lambda_c^+, \lambda_c^+)$  that fulfills these conditions when  $c \in (1, 1.24)$ .

□

## D.9 Proof of Proposition 3

*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 4.3 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.<sup>23</sup>

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 4.3 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

$$\begin{aligned} U_i(s_i, c) &= \frac{[(c + \lambda_j)\lambda_i s_i + \lambda_j R_j s_i][(c + \lambda_i)\lambda_j R_j s_i + \lambda_i s_i]}{[(c + \lambda_i)(c + \lambda_j) - 1]^2} \\ &\quad - \frac{c[(c + \lambda_j)\lambda_i s_i + \lambda_j R_j s_i]^2}{2[(c + \lambda_i)(c + \lambda_j) - 1]^2} - \frac{\lambda_i}{2} \left[ s_i - \frac{(c + \lambda_j)\lambda_i s_i + \lambda_j R_j s_i}{(c + \lambda_i)(c + \lambda_j) - 1} \right]^2 \\ &= \Psi(\lambda_i, \lambda_j, c) s_i^2. \end{aligned} \quad (54)$$

Here,  $\Psi(\lambda_i, \lambda_j, c)$  is a function of the parameters  $\lambda_i, \lambda_j, c$  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 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

<sup>23</sup>In the case where  $\Theta_k = 0$  and  $[\lambda_l = 0 \text{ or } R_l = 0]$  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  $s_j = 0$ .

same optimisation problem as under simultaneous communication. From Section 4.3 we know that  $i$ 's best reply from this interval is therefore  $s_i = \min \left\{ \max \left\{ R_i, \frac{1}{R_j} \right\}, 1 \right\}$ . From Section 4.3 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 4.3 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  $\lambda_c^+$  we have  $\min(R_i, R_j) \geq 1 \implies R_i R_j \geq 1$ . 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 8 that  $\lambda' \neq \lambda_c^+ \implies (\lambda', \lambda_c^+) \notin Cl(\Lambda_{max})$  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. (26)), the set of points satisfying Eq. (25) 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  $(\lambda', \lambda_c^+)$ ,  $R_k > 1$ . Reproducing Eq. (25) but imposing  $\lambda_i = \lambda_k = \lambda_c^+$  and

$\lambda_j = \lambda' = 0$ , we have that  $R_k > 1$  if

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

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  $(\lambda', \lambda_c^+)$  such that  $\lambda' < \lambda_c^+$ ,  $\pi(\lambda', \lambda_c^+) < \pi(\lambda_c^+, \lambda_c^+)$  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  $(\lambda_i, \lambda_c^+)$  such that  $\lambda_i \leq \lambda_c^+$ , if  $s_j = 1$  and  $s_i = R_i$ , then  $\pi(\lambda_i, \lambda_c^+) < \pi(\lambda_c^+, \lambda_c^+)$ . 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 <sup>24</sup>

$$\frac{[(c(c + \lambda') - 1)][c - 3c\sqrt{5-4c} + c^3(1 + \sqrt{5-4c}) + c^2\lambda'(1 + \sqrt{5-4c}) - 2(2 + \lambda' + \sqrt{5-4c}\lambda')]}{2(1 + \sqrt{5-4c})(c + \lambda')^2}. \quad (55)$$

The payoff  $\pi(\lambda_c^+, \lambda_c^+)$  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 <sup>25</sup>

$$\frac{(c-2)(1 + \sqrt{5-4c} - 2(c-1)c)^2}{2(3 + \sqrt{5-4c} - 2c)^2}. \quad (56)$$

The value of Eq. (56) is strictly greater than that of Eq. (55) for all  $\lambda' < \lambda_c^+$  for all <sup>26</sup>  $c < 1.25$ . Therefore, in all possible cases (1), (2), (3) and (4), if  $\lambda_k = \lambda_c^+$  then a deviation by player  $l$  from  $\lambda_c^+$  to  $\lambda' < \lambda_c^+$  yields a strictly lower payoff and hence for  $\lambda' < \lambda_c^+$ , we have that  $\pi(\lambda', \lambda_c^+) < \pi(\lambda_c^+, \lambda_c^+)$ .

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. (25), if  $\lambda_l > \lambda_c^+$  then the pair  $(\lambda_c^+, \lambda_c^+)$  must also satisfy Eq. (25) with  $l = i$  and  $k = j$ , as  $\lambda_i$  is increased while  $\lambda_j$  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

<sup>24</sup>This simplification was obtained using Mathematica. The code is available in the supplementary appendix of this paper.

<sup>25</sup>This simplification was obtained using Mathematica. The code is available in the supplementary appendix of this paper.

<sup>26</sup>This result is obtained using Mathematica. The code is available in the supplementary appendix of this paper.

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  $(\lambda', \lambda_c^+)$  such that  $\lambda' > \lambda_c^+$ , it is the case that  $\pi(\lambda', \lambda_c^+) < \pi(\lambda_c^+, \lambda_c^+)$  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  $\pi(\lambda_c^+, \lambda_c^+)$ . 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<sup>27</sup>

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

The value of Eq. (56) is strictly greater than that of Eq. (57) for all  $\lambda' < \lambda_c^+$  for all<sup>28</sup>  $c < 1.25$ . Therefore in all possible cases (1), (2), (3) and (4), if  $\lambda_k = \lambda_c^+$  then deviation by player  $l$  from  $\lambda_c^+$  to  $\lambda' < \lambda_c^+$  yields a strictly lower payoff and hence for  $\lambda' \neq \lambda_c^+$  we have that  $\pi(\lambda', \lambda_c^+) < \pi(\lambda_c^+, \lambda_c^+)$  and  $(\lambda_c^+, \lambda_c^+)$  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  $\pi(\lambda_c^+, \lambda_c^+)$  is the same under both forms of communication (these results are points 3 and 5 of the proposition).

Finally, we prove that if  $c < 1.2$ , then  $\pi(\lambda_c^+, \lambda_c^+) > \pi(\lambda', \lambda')$  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, payoffs are the same as in the game with simultaneous communication, which by point 1 of Theorem 2 are less than  $\pi(\lambda_c^+, \lambda_c^+)$ . In the latter case, either the payoffs are the same as in the game with simultaneous communication (and both players' payoffs are zero and therefore less than  $\pi(\lambda_c^+, \lambda_c^+)$ ) or there is an equilibrium in which  $s_i = 1$  and  $s_j = R_j$ . In this case 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  $\pi_i(\lambda', \lambda')$  denote the payoff to the player selected to send his message first, we have that

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

Player  $j$ 's payoff is

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

By using Mathematica<sup>29</sup> we have verified that for all  $c < 1.2$ ,  $\frac{\pi_i(\lambda', \lambda') + \pi_j(\lambda', \lambda')}{2} = \pi(\lambda', \lambda') < \pi(\lambda_c^+, \lambda_c^+)$ ; 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  $\pi(\lambda_c^+, \lambda_c^+)$ , which completes the proof.  $\square$

<sup>27</sup>This simplification was obtained using Mathematica. The code is available in the supplementary appendix of this paper.

<sup>28</sup>This result is obtained using Mathematica. The code is available in the supplementary appendix of this paper.

<sup>29</sup>The code is available in the supplementary appendix of this paper.

## D.10 Proof of Proposition 4

*Proof.* In the exposition of this proof, let  $\pi^1(\lambda_i, \lambda_j)$  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,  $\pi^1(\lambda_i, \lambda_j)$  denotes the expected payoff) and use  $\pi(\lambda_i, \lambda_j)$  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  $\mu_{\chi_j}$ . Their expected utility function is

$$U_i(x_i, \mu_{\chi_j}, s_i, c) = x_i \mu_{\chi_j} - \frac{cx_i^2}{2} - \mathbf{1}_{s_i > x_i} \frac{\lambda_i}{2} (s_i - x_i)^2. \quad (58)$$

Suppose first that  $s_i \leq \frac{\mu_{\chi_j}}{c}$ . 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:

$$x_i^*(x_j, s_i, s_j, \lambda_i, \lambda_j, c) = \begin{cases} \frac{x_j}{c} & \text{if } s_i \leq \frac{x_j}{c} \\ \frac{x_j + \lambda_i s_i}{c + \lambda_i} & \text{if } s_i > \frac{x_j}{c} \end{cases}. \quad (59)$$

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  $x_i^* = \frac{x_j + \lambda_i s_i}{c + \lambda_i} < \frac{c \cdot s_i + \lambda_i s_i}{c + \lambda_i} = s_i$  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  $s_i < x_i = \frac{x_j}{c} < x_j = \frac{x_i + \lambda_j s_j}{c + \lambda_j} < s_j$ , which implies the following effort choices in equilibrium:

$$x_i^e(x_j, s_i, s_j, \lambda_i, \lambda_j, c) = \frac{\lambda_j}{c^2 + c\lambda_j - 1} s_j \equiv \alpha_i s_j \quad (60)$$

$$x_j^e(x_i, s_i, s_j, \lambda_i, \lambda_j, c) = \frac{\lambda_j}{c + \lambda_j - \frac{1}{c}} s_j \equiv c\alpha_i s_j. \quad (61)$$

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  $(\lambda_i, \lambda_j)$ ,  $\alpha_i < R_i$ . Either  $\Theta_i \leq 0$ , and so by definition  $R_i = \infty > 1 > \alpha_i$ , or  $\Theta_i > 0$ , in which case

$$\begin{aligned}
& \alpha_i < R_i \\
& \iff \alpha_i < \frac{\lambda_j}{\Theta_i} \\
& \iff \frac{\lambda_j}{c^2 + c\lambda_j - 1} < \frac{\lambda_j}{c(c + \lambda_j) + \frac{1}{(c + \lambda_i)(c + \lambda_j)} - 2} \\
& \iff c(c + \lambda_j) + \frac{1}{(c + \lambda_i)(c + \lambda_j)} - 2 < c^2 + c\lambda_j - 1 \\
& \iff c^2 + c\lambda_j + \frac{1}{(c + \lambda_i)(c + \lambda_j)} - 2 < c^2 + c\lambda_j - 1 \\
& \iff \frac{1}{(c + \lambda_i)(c + \lambda_j)} < 1, \tag{62}
\end{aligned}$$

which always holds as  $c > 1$ .

We can now show that for any  $(\lambda_i, \lambda_j)$  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  $s_i \in (\alpha_i s_j, \min\{\frac{s_j}{\alpha_i}, 1\}]$  and  $s_j \in (\alpha_j s_i, \min\{\frac{s_i}{\alpha_j}, 1\}]$ . Lemma 3 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  $s_i^* = \min\{R_i s_j, \frac{s_j}{\alpha_i}, 1\}$ . 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

$$s_i = \frac{s_j}{R_j} < \frac{s_j}{\alpha_j} \tag{63}$$

and

$$s_i = \frac{s_j}{R_j} < \frac{s_j}{R_j} R_i \cdot R_j = R_i s_j, \tag{64}$$

where the inequality in Eq. (64) follows given that  $R_i \cdot R_j > 1$ . Eq. (63) and Eq. (64) 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  $(\lambda_i, \lambda_j)$  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. (60) and Eq. (61). 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  $\alpha_i$ ) 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<sup>30</sup>

$$U_i(c) = \frac{c(\sqrt{5-4c}-1)+2}{4}. \quad (65)$$

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. (60) and Eq. (61)) into the utility function. Simplifying the resulting expression yields<sup>31</sup>

$$U_i(c) = \frac{c(\sqrt{5-4c}-2c+3)}{4}. \quad (66)$$

---

<sup>30</sup>This simplification was obtained using Mathematica. Code available in the supplementary appendix accompanying this paper.

<sup>31</sup>This simplification was obtained using Mathematica. The code is available in the supplementary appendix of this paper.

Therefore, the deviation is not profitable if

$$\begin{aligned}
& \frac{c(\sqrt{5-4c}-1)+2}{4} > \frac{c(\sqrt{5-4c}-2c+3)}{4} \\
& \iff 2-c > 3c-2c^2 \\
& \iff 2+2c^2 > 4c & (67) \\
& \iff 1+c^2 > 2c. & (68)
\end{aligned}$$

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. (66) multiplied by  $s_j^2$ . Note also that utility from deviating to  $s_i = s_j$  is equal to the expression in Eq. (65) 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  $\pi^1(\lambda_c^+, \lambda_c^+) = \pi(\lambda_c^+, \lambda_c^+)$ . 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  $(\lambda', \lambda_c^+)$  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  $\pi^1(\lambda_c^+, \lambda_c^+) > \pi^1(\lambda', \lambda_c^+)$ . In the case where one player,  $i$ , reneges upward, the equilibrium effort levels are of the form given by Eq. (60) and Eq. (61). 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. (60) and Eq. (61) we see that the equilibrium efforts when  $s_j = 1$  are the same as those in the equilibrium of the two-sided game where  $\lambda_i = 0$ ; i.e., if  $(\lambda_i, \lambda_j)$  induces an equilibrium in which player  $i$  reneges upward, then  $\pi_1(\lambda_i, \lambda_j) = \pi(0, \lambda_j)$ . We know from Theorem 2 that  $\pi(\lambda_c^+, \lambda_c^+) > \pi(0, \lambda_c^+)$ . Therefore, for any  $(\lambda_i, \lambda_c^+)$  that induces an equilibrium in which player  $i$  reneges upward in the one-sided game, we have that  $\pi_1(\lambda_i, \lambda_c^+) = \pi(0, \lambda_c^+) < \pi(\lambda_c^+, \lambda_c^+) = \pi^1(\lambda_c^+, \lambda_c^+)$ . Finally, consider any  $(\lambda_i, \lambda_c^+)$  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. (60) and Eq. (61) and substituting the resulting expressions into the expression for the material payoff (Eq. (1) yields

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

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

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

which never holds for  $c \in (1, 1.25)$  and<sup>32</sup>  $\lambda_i \geq 0$ . 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  $\pi^1(\lambda', \lambda_c^+)$ , we have that  $\pi^1(\lambda_c^+, \lambda_c^+) > \pi^1(\lambda', \lambda_c^+)$  for all  $\lambda' \neq \lambda_c^+$ .

Finally, we show that if  $c < 1.22$ , then  $\pi^1(\lambda_c^+, \lambda_c^+) > \pi^1(\lambda', \lambda')$  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  $\pi(\lambda_c^+, \lambda_c^+)$ . 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  $\pi(\lambda_c^+, \lambda_c^+)$ ); (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,  $\pi^1(\lambda_c^+, \lambda_c^+) > \pi^1(\lambda', \lambda')$ . 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,  $\pi^1(\lambda', \lambda')$  is the average of the equilibrium payoffs of players in the two roles (denoted by  $i$  and  $j$ ). To obtain the payoff function  $\pi^1(\lambda', \lambda')$  we therefore impose  $s_j = 1$  on Eq. (60) and Eq. (61) 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 average to yield the payoff:

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

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

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

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  $\pi(\lambda_c^+, \lambda_c^+)$  if  $c < 1.22$ , yielding the result.  $\square$

<sup>32</sup>This result was obtained using Mathematica. The code is available in the supplementary appendix of this paper.

## D.11 Proof of Proposition 5

Fix  $c \in (1, 2)$ . Let  $0 < \beta_c^+ \equiv \frac{1}{2c} + \frac{c}{2} - 1 = \frac{1}{2} \cdot \left(\frac{1}{c} + c\right) - 1$ . Consider the partnership game with fixed reneging costs  $\beta_i = \beta_j = \beta_c^+$ . For each player  $i$ , let  $x_i^* : [0, 1]^2 \rightarrow [0, 1]$  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 4.2), 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,  $U_i\left(\frac{1}{c} \cdot x_j, x_j, 1, c\right) - U_i(1, x_j, 1, c)$ , is equal to

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

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

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

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  $\left( (1, 1), (x_i^*, x_j^*) \right)$  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  $\left( (1, 1), (x_i^*, x_j^*) \right)$  is a subgame-perfect equilibrium of the partnership game with fixed reneging costs  $\beta_i = \beta_j = \beta_c^+$ .

## D.12 Proof of Proposition 6

*Proof.* Parts (1), (2), and (4) of Proposition 6 are immediate from the fact that equilibrium payoffs remain the same as in the baseline model of full observability. We have to prove that for each  $c \in (1, 1.25)$ , there exists  $\bar{q} < 1$  such that  $(\lambda_c^+, \lambda_c^+)$  is a strict Nash equilibrium of the population game with observability  $q$  for each  $q \in [\bar{q}, 1)$ , i.e.,  $\pi_q(\lambda', \lambda_c^+ | \lambda_c^+) < \pi(\lambda_c^+, \lambda_c^+)$  for all  $\lambda' \neq \lambda_c^+$ . We first note that by Lemma 8, for any  $\lambda' \neq \lambda_c^+$ , we have that  $(\lambda', \lambda_c^+) \notin Cl(\Lambda_{max})$ . If  $(\lambda', \lambda_c^+) \in \Lambda_{0-eff}$ , the material payoff to both players is zero and hence any mutant  $\lambda'$  such that  $G(\lambda', \lambda_c^+)$  induces a no-effort equilibrium achieves a strictly lower material payoff than the incumbent type  $\lambda_c^+$  in encounters where levels of reneging aversion are observed. In the proof of Theorem 2 it was shown that when  $\lambda' \leq \lambda_c^+$ , the derivative is equal to (the left derivative when  $\lambda' = \lambda_c^+$ ):

$$\frac{\partial \pi(\lambda', \lambda_c^+)}{\partial \lambda'} = \frac{[\lambda_c^+]^2 (c(\lambda_c^+ + c) - 1)^2}{[1 + (c + \lambda_c^+)(c + \lambda')(c(\lambda_c^+ + c) - 2)]^3} \quad (71)$$

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

$$\lim_{\lambda' \nearrow \lambda_c^+} \frac{\partial \pi(\lambda', \lambda_c^+)}{\partial \lambda'} > 0. \quad (72)$$

The fact that the derivative of the material payoff function with respect to  $\lambda'$  is strictly positive for all  $\lambda' < \lambda_c^+$  and that the left derivative at  $\lambda_c^+$  is bounded away from zero implies that when levels of reneging aversion are observed and  $(\lambda', \lambda_c^+) \in \Lambda_{2-msg}$ , there is a first-order material payoff loss for a mutant with  $\lambda' < \lambda_c^+$ , compared to the incumbent type  $\lambda_c^+$ . Now, considering the case where  $\lambda' > \lambda_c^+$  and  $(\lambda', \lambda_c^+) \in \Lambda_{2-msg}$ , we note that, analogously, in the proof of Theorem 2 it was shown that when  $\lambda' \geq \lambda_c^+$ , the derivative of the payoff function with respect to  $\lambda'$  is strictly negative and the right derivative of the payoff function, evaluated at  $\lambda_c^+$ , is strictly negative; i.e., the payoff increases as  $\lambda'$  decreases toward  $\lambda_c^+$

(Mathematica code demonstrating this is available in the online appendix). Therefore, there is also a first-order loss for a mutant with  $\lambda' > \lambda_c^+$  when renegeing costs are observed.

We have therefore demonstrated that any mutant achieves a strictly lower payoff in the partnership games played after renegeing costs are observed as compared to an incumbent, i.e.,  $\pi(\lambda', \lambda_c^+) < \pi(\lambda_c^+, \lambda_c^+)$  for all  $\lambda' \neq \lambda_c^+$ , and, further, that the first-order loss of a mutant is bounded away from zero when  $\lambda' \rightarrow \lambda_c^+$ .

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

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

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

$$q \cdot [\pi(\lambda', \lambda_c^+) - \pi(\lambda_c^+, \lambda_c^+)] + (1-q) \frac{1}{2c}. \quad (73)$$

We therefore have that a mutant of type  $\lambda'$  is strictly outperformed by an incumbent of type  $\lambda_c^+$  when Eq. (73) is strictly negative. Imposing this strict negativity and rearranging for  $q$  yields

$$q > \frac{1}{1 + 2c[\pi(\lambda_c^+, \lambda_c^+) - \pi(\lambda', \lambda_c^+)]} \equiv \widetilde{q}_{\lambda'} \quad (74)$$

From the fact that the term in square brackets in the denominator of Eq. (74) is strictly positive, it is immediate that  $\widetilde{q}_{\lambda'} \in (0, 1)$ . We then define  $\bar{q} \equiv \sup\{\widetilde{q}_{\lambda'} : \lambda' \in \mathbb{R}^+\}$ . It follows that for all  $c \in (1, 1.25)$ , there exists  $\bar{q}$  such that for all  $q \in [\bar{q}, 1]$ ,  $(\lambda_c^+, \lambda_c^+)$  is a strict Nash equilibrium of the population game with partial observability.

The stable population in the setting with partial observability has been proven to be identical to the population in the setting with perfect observability. Therefore, results 1, 2, and 4 of Theorem 2 that pertain to the payoffs of this stable population hold also in the partial observability setting.  $\square$

### D.13 Proof of Proposition 7

*Proof.* If the incumbents have  $\lambda = 0$ , then they exert no effort by Fact 1. If the incumbents have  $\lambda > 0$ , then assume to the contrary that agents exert a positive level of effort on the equilibrium path. By Theorem 1, this implies that all agents make the maximal promise 1 and, due to the payoff function being strictly convex, that they exert the same positive level of effort  $x_i^e(1, 1, \lambda, \lambda, c) > 0$  on the equilibrium path in the second stage (see Eq. (5)). Consider a mutant with zero renegeing cost who sends message 1 and then exerts effort  $\frac{1}{c} \cdot x_i^e(1, 1, \lambda, \lambda, c)$ . It is immediate that such a mutant achieves a strictly higher payoff than the incumbents because the mutant exerts the unique amount of effort that maximises the payoff, given that the partner exerts effort  $x_i^e(1, 1, \lambda, \lambda, c)$ .  $\square$

## D.14 Proof of Proposition 8

*Proof.* We show that there can be no symmetric pure Nash equilibrium of the population game in which players exert no effort on the equilibrium path. Consider any symmetric population in which players have a level of renegeing aversion  $\lambda$  and in which, in game  $G(\lambda, \lambda)$ , the unique equilibrium is a no-effort equilibrium and hence all players achieve a material payoff of zero, i.e.,  $\pi(\lambda, \lambda) = 0$ . Lemma 7 implies that for any such  $\lambda$ , there exists  $\lambda'$  such that in  $G(\lambda, \lambda')$  - the partnership game played where players of type  $\lambda$  and type  $\lambda'$  meet and observe each other's level of renegeing aversion - both players exert positive effort in equilibrium and achieve strictly positive material payoffs. As any player can always guarantee a level of utility of at least zero in any interaction, a player of type  $\lambda'$  achieves a weakly positive payoff from the partnership game played after players of types  $\lambda$  and  $\lambda'$  meet but do not observe each other's level of renegeing aversion. Therefore, when  $q > 0$  (i.e., players in a population observe each other's level of renegeing aversion at least some of the time), any mutant of type  $\lambda'$  achieves a strictly positive payoff in the population game with partial observability, i.e.,  $\pi_q(\lambda', \lambda|\lambda) > 0 = \pi(\lambda, \lambda)$ .  $\square$