School of Economics Working Paper
2018-06

University of East Anglia

SCHOOL OF ECONOMICS

Communication as Gift-Exchange

by Mark T. Le Quement*
Amrish Patel*

*University of East Anglia

School of Economics
University of East Anglia
Norwich Research Park
Norwich NR4 7TJ
United Kingdom
www.uea.ac.uk/economics

Communication as Gift-Exchange*

Mark T. Le Quement† and Amrish Patel‡

October 2018

Abstract

We study psychological games of cheap talk communication involving players who have misaligned material interests and reciprocity preferences. We find that full and efficient information transmission is often impossible if reciprocity concerns are too high. Furthermore, higher material preference misalignment may facilitate the achievement of full information transmission. A key driver of our results is that truth-telling is not *per se* a kind action by the sender. We contrast discrete and continuous environments, alternative conceptions of reciprocity preferences and consider one-sided reciprocity models.

Keywords: Cheap talk, Gift-Exchange, Incomplete Information, Psychological Game, Reciprocity.
JEL classification: D81, D83, D91

*We thank Martin Dufwenberg and participants of the 2^{nd} Workshop on Psychological Game Theory for helpful comments. All errors remain our own.

†School of Economics, University of East Anglia, M.Le-Quement@uea.ac.uk
‡School of Economics, University of East Anglia, Amrish.Patel@uea.ac.uk

"*The gift of truth excels all other gifts.*" Gautama Buddha, *Dhammapada*, verse 354.
"*A truth that's told with bad intent, beats all the lies you can invent.*" William Blake, *Auguries Of Innocence*.

Good information transmission between asymmetrically informed agents is often critical in economic interaction. The possibility of such transmission has primarily been examined in relation to the exogenous alignment in material payoffs. In Crawford and Sobel (1982)'s seminal paper on cheap talk communication, the authors find that information transmission is negatively affected by the degree of misalignment of material preferences.

Communication problems are inherently situations where there are gains from cooperation. In unrelated contexts with gains from cooperation (e.g. prisoners' dilemmas; employer-employee relations (Akerlof 1982, Fehr et al. 1993, Charness et al. 2004); charities and donors (Falk 2007)), gift-exchange behaviour is often observed. A leading explanation for such behaviour is that individuals have reciprocity preferences (Rabin 1993; Dufwenberg and Kirchsteiger 2004; Falk and Fischbacher 2006). That is, they want to be kind to those who are kind to them, and unkind to those who are unkind to them.

Could the behaviour of parties involved in communication constitute a form of gift-exchange? And if so, could this potentially provide a rationale for why more truth-telling is observed in experiments than predicted by the standard model (Cai and Wang 2006; Sanchez-Pages and Vorsatz 2007) and why subjects' lying in experimental cheap talk games is sensitive to the payoff consequences for receivers (Gneezy 2005).

That communicative interaction may constitute an instance of gift-exchange is not obvious. Most gifts have a direct impact on material interests either because they are physical objects or actions that directly impact material payoffs. By contrast, communication involves a gift of information, which is often unverifiable, subject to interpretation, etc.

We re-examine strategic information transmission in light of reciprocity preferences. How do reciprocity preferences affect the possibility of informative communication? We tackle this question by studying simple cheap talk games involving agents (a sender, S, and a receiver, R) who have heterogeneous material preferences (i.e. S is biased) and have a preference for reciprocating the kindness of their co-agent. We address the following specific questions. Does there exist a fully revealing equilibrium? For a fixed material in-

centive misalignment, does an increase in the reciprocity concern always foster existence of such an equilibrium? For a fixed reciprocity concern, is an increase in S's bias always hurtful? Does truth-telling always imply that S is being kind? What are the implications for welfare?

Our main analysis examines a simple binary state game in which the low state implies identical materially optimal actions for both parties whereas the high state implies different materially optimal actions. The most interesting case is if the sender's materially optimal action in the high state is closer to the receiver's materially optimal action in the low state than to the latter's materially optimal action in the high state. In this case, truth-telling is impossible if agents are motivated purely by material payoffs. We append a model of reciprocity preferences inspired by Rabin (1993) and Dufwenberg and Kirchsteiger (2004) appropriate to our game to arrive at a psychological game of incomplete information. We then characterise equilibria with full information transmission (which are arguably most focal).

Our first main finding is that reciprocity preferences indeed can improve communication. Reciprocity concerns can generate new equilibria with more information transmission. In such equilibria, agents reciprocate each other's kindness or unkindness. In so-called 'positive truth-telling equilibria', the sender is being kind by refraining from lying despite her potential material gain, and the receiver is in response kind by taking a compromise action biased towards the sender's materially optimal action given the revealed state. In so-called 'negative truth-telling' equilibria, S truth-tells believing that R will take an action which is materially detrimental for both agents and R, in response, indeed takes such a destructive action to reciprocate S's unkindess. In such equilibria, communicative interaction constitutes gift-exchange in the germanic sense of the word gift (i.e. poison). Summarizing, reciprocity can provide a potential explanation for the overcommunication observed in cheap talk experiments (Cai and Wang 2006; Sanchez-Pages and Vorsatz 2007).

Our second main finding is that for a given degree of material preference misalignment, a higher reciprocity concern can be detrimental to the possibility of truthful communication. A key underlying mechanism originates in the fact that R's action, con-

ditional on S's message being kind, becomes kinder as R's kindness concern increases. With a binary state space, if R's action is excessively kind in response to S's truth-telling (as is the case when the reciprocity concern is high enough), deviating from the truth is Pareto-dominated at all values of the state. This in turn, using Rabin (1993)'s definition of kindness, negates the assumption that truth-telling by S is kind, and the gift-exchange unravels. In the above scenario, reciprocity in kindness has a self-infirming property. One player's kindness undoes the other's kindness in a zero-sum fashion, so that players are de facto engaged in a kindness conflict.

Our third main finding is that for a given degree of reciprocity concerns, more material preference *mis*alignment can be beneficial for truthful communication. A first intuition is that as material payoff misalignment increases, the potential for being kind increases as the size of feasible favours mechanically increases (the size of the cake to be divided grows). A second key intuition is that as the degree of material payoff misalignment increases, the kindness of S in a truth-telling equilibrium inflates mechanically. The kindness of an agent i towards agent j is the difference between the equilibrium payoff received by j given the action chosen by i and a so-called equitable payoff. The latter is the average of the highest and the lowest payoff that j could obtain conditional on i's action being Pareto-efficient. The key is that this equitable payoff decreases mechanically as material payoff misalignment increases.

Our first insight, that reciprocity can generate new and more efficient equilibria has been noted in other games (e.g. the sequential prisoner's dilemma game in Dufwenberg and Kirchsteiger 2004). Analogues of our second and third insights do not appear in Rabin (1993) and Dufwenberg and Kirchsteiger (2004).

An extensions section considers three alternative environments. First, we study a simple version of the Crawford and Sobel (1982) model with a continuous state and action space and constant bias. The three central results from our main analysis hold in this environment.

Second, we examine an alternative notion of reciprocity preferences where the kindness reference point is independent of beliefs (cf. Dufwenberg and Kirchsteiger 2004). Under this assumption, truth-telling is feasible for a larger share of the parameter space.

This is because messages and actions are seen as kinder than they are under the belief-dependent definition of the kindness reference point, thereby making gift-exchange via communication easier. Nonetheless, the result that greater reciprocity concerns can be detrimental to truth-telling still holds.

Third, we consider one-sided models of reciprocity. While both agents being reciprocal may be appropriate for contexts where they are peers (e.g. asymmetrically informed work colleagues), it seems less likely when the relationship is vertical (e.g. a firm with private information trying to sell to a consumer). We briefly consider a model where S only cares about material payoffs and R has reciprocity preferences. We then examine a model where S has a desire to be kind and R has reciprocity preferences. This latter model may be relevant for examples like a socially responsible firm interacting with reciprocal consumers. We find that there is more truth-telling here than in our main model, but our three main results still arise in this environment.

Finally, we compare our results to those obtained in Rabin (1993) and Dufwenberg and Kirchstegier 2004) to examine whether insights analogous to ours have been found in the games that they consider.

Our paper lies at the intersection of two literatures studying respectively psychological games (Geneakoplos et al. 1989, Battigalli and Dufwenberg 2009) and strategic information transmission (Crawford and Sobel 1982). Our paper adds to the psychological games literature on reciprocity (Rabin 1993; Battigalli and Dufwenberg 2009; Dufwenberg and Kirchsteiger 2004) and that on Bayesian psychological games (Battigalli and Dufwenberg 2009; Attanasi et al. 2016; Battigalli et al. 2018). There is relatively little research on the implications of reciprocity in incomplete information environments. Previous work has studied mechanism design (Bartling and Netzer 2016; Bierbrauer and Netzer 2016; Bierbrauer et al. 2017); employer-employee relations where preferences are private information (von Siemens 2013) and principal-multiple agent settings (De Marco and Immordino 2014).

Second, we contribute to a vast literature on strategic information transmission initiated by Crawford and Sobel (1982) (see Farrell and Rabin (1996), Krishna and Morgan (2008) or Sobel (2013) for general reviews of this literature). While much research has

been devoted to understanding when informative communication is incentive compatible, most existing studies rely on bounded rationality (e.g. Cai and Wang 2006) or an exogenous alignment of incentives, via material payoffs (Crawford and Sobel 1982) or via exogenous lying-costs (e.g. Kartik 2009).

Previous work at the intersection of these two literatures includes the following contributions. Battigalli et al. (2013) examine how guilt aversion affects behaviour in cheap talk games. The series of papers (Khalmetski and Silwka 2017; Dufwenberg and Dufwenberg 2018; Gneezy et al. 2018) study how an aversion to "perceived lying" explains behaviour in the dice-roll reporting experiment (Fischbacher and Föllmi-Heusi 2013, Abeler et al. 2018). To the best of our knowledge, ours is the first paper to study the implications of reciprocity in cheap talk sender-receiver games.

We proceed as follows. Section 1 introduces the general type of game that we study and defines how we model reciprocity preferences. Section 2 presents our main results and Section 3 examines alternative environments. We then conclude.

1 General environment

We first introduce the general cheap talk game environment that we work within (1.1) and then present how we model reciprocity (1.2). Our main results in Section 2 examine specific games within the general environment presented here.

1.1 The cheap talk game

There are two agents S and R. S privately observes the state ω, which is drawn from a commonly known distribution over a state space Ω. After observing ω, S can choose a costless message m taken from a message set $M = \Omega$. After observing S's message, R can choose an action a from the action space A.[1] The latter can be bounded, for example $A = [\underline{a}, \overline{a}]$, or it could be $\mathbb{R}$. The timing of the game is as follows. The state ω is drawn

[1]To avoid confusion, note that we shall never use the word "action" to refer to a choice by an arbitrary player. The word action is reserved only for a choice by R.

according to c.d.f. F (endowed with pdf f) and privately observed by S. S chooses $m \in M$. R picks a after observing m.

A profile (a, ω) determines a material payoff for each agent. An example would be that the material payoff function of S is given by $\pi_S(a, \omega) = -|a + b - \omega|$, for $b > 0$, while that of R is given by $\pi_R(a, \omega) = -|a - \omega|$. Note that the difference between the ideal actions of agents does not have to be constant across states.

A (communication) strategy σ_S of S is a conditional distribution over messages in the message set M at every information set of S (every possible ω). A strategy of R is a conditional distribution over the action set for each message in M. We allow for mixed strategies throughout Section 1, however our main results focus on pure strategy equilibria.

Let $a(m, \sigma_R)$ be the pure action picked by R in response to message m if R uses the pure strategy σ_R. Similarly, let $m(\omega, \sigma_S)$ denote the message picked by S given that the state is ω if S uses the pure strategy σ_S.

The definitions that follow correspond to a setup with continuous as well as bounded action and state spaces defined as $A = [\underline{a}, \overline{a}]$ and $\Omega = [\underline{\omega}, \overline{\omega}]$. Modifications to the notation implied by the discrete and/or unbounded cases are trivial and therefore left to the reader.

We now introduce notation for expected payoffs given a particular strategy profile. For any given σ_R, ω, let $E\pi_i(m|\sigma_R, \omega)$ denote the expected material payoff of agent $i \in \{S, R\}$ given that the state is ω, that m is sent and R uses σ_R. We thus have

$$E\pi_i(m|\sigma_R, \omega) = \int_{\underline{a}}^{\overline{a}} \pi_i(a, \omega) g(a|m, \sigma_R)\, da,$$

where $g(a|m, \sigma_R)$ is a (possibly non-degenerate) distribution over $A = [\underline{a}, \overline{a}]$.

For any given σ_S, m, let $E\pi_i(a|\sigma_S, m)$ denote the expected material payoff of agent $i \in \{S, R\}$ conditional on R observing m, S being known to use σ_S and R taking action a. We thus have

$$E\pi_i(a|m, \sigma_S) = \int_{\underline{\omega}}^{\overline{\omega}} \pi_i(a, \omega) f(\omega|m, \sigma_S)\, d\omega,$$

where $f(\omega|m, \sigma_S) = \frac{P(\omega, m|\sigma_S)}{\int_{\underline{\omega}}^{\overline{\omega}} P(t, m|\sigma_S) dt}$.

1.2 Modelling reciprocity

We now incorporate reciprocity preferences following an approach inspired by Rabin (1993) and Dufwenberg and Kirchsteiger (2004). Since their models are only defined for games of complete information, we define one appropriate for our game of incomplete information (cf. Attanasi et al. 2016). As in the previous subsection, our definitions correspond to a setup with a continuous state space; modifications for the discrete case are trivial.

The Rabin-Dufwenberg-Kirchsteiger approach to modelling reciprocity uses "kindness functions". To define how kind i is to j we need notation to explicitly represent agents' beliefs about their co-agent's strategies and beliefs. Let $\overline{\sigma}_{ij}$ denote i's belief about j's strategy and $\overline{\sigma}_{iji}$ denote i's belief about j's belief about i's strategy. We restrict attention to point beliefs.

To determine whether i is kind to j one needs a reference point, the *equitable payoff*. When defining this reference point it has been argued that only *efficient* strategies, those not involving "wasteful" play, are relevant. We adopt Rabin's approach to defining efficient strategies here where efficiency depends on an agent's beliefs[2] (see Section 3.2 for the implications of the belief-independent approach of Dufwenberg and Kirchsteiger (2004) in our context).

Efficient actions, messages and strategies: We first define the inefficient messages of S. A message m is *inefficient at ω conditional on $\overline{\sigma}_{SR}$* if and only if there exists $m' \neq m$ such that

$$E\pi_i\left(m' \left| \overline{\sigma}_{SR}, \omega\right.\right) \geq E\pi_i\left(m \left| \overline{\sigma}_{SR}, \omega\right.\right), \ \forall i \in \{S, R\},$$

with at least one of the above inequalities holding strictly. Define $\Sigma_S^-\left(\omega, \overline{\sigma}_{SR}\right)$ as the set of messages that are inefficient given $\omega, \overline{\sigma}_{SR}$.

We now define the inefficient actions of R. An action a is *inefficient given m conditional on $\overline{\sigma}_{RS}$* if and only if there exists $a' \neq a$ such that

$$E\pi_i\left(a' \left| \overline{\sigma}_{RS}, m\right.\right) \geq E\pi_i\left(a \left| \overline{\sigma}_{RS}, m\right.\right), \ \forall i \in \{S, R\},$$

[2]This approach was also used by Netzer and Schmutzler (2014) and Bierbrauer and Netzer (2016).

with at least one of the above inequalities holding strictly. Define $\Sigma_R^-(m, \overline{\sigma}_{RS})$ as the set of actions that are inefficient given $m, \overline{\sigma}_{RS}$.

Using the notions of inefficient actions and messages, we can now define inefficient strategies.

Definition 1 ***a)*** *A sender strategy σ_S is efficient given $\overline{\sigma}_{SR}$ if for any ω and m such that $m \in \Sigma_S^-(\omega, \overline{\sigma}_{SR})$ (i.e. m is inefficient given ω and $\overline{\sigma}_{SR}$), it holds true that m is sent with probability 0 given that the state is ω.* ***b)*** *A strategy σ_R is efficient given $\overline{\sigma}_{RS}$ if for any m and a such that $a \in \Sigma_R^-(m, \overline{\sigma}_{RS})$ (i.e. a is inefficient given m and $\overline{\sigma}_{RS}$), it holds true that action a is picked with probability 0 given that R observes m.* ***c)*** *A strategy of i that is not efficient given $\overline{\sigma}_{ij}$ is said to be inefficient given $\overline{\sigma}_{ij}$.*

Kindness functions and utility: We now define kindness functions and explain how they are used in a utility function to represent agents' reciprocity preferences.

First we define how kind S is to R. Given $\overline{\sigma}_{SR}$ and state ω, the kindness of S towards R if sending message m is given by:

$$\begin{aligned} & K_{SR}(m\,|\omega, \overline{\sigma}_{SR}\,) \\ = \;& E\pi_R\,(m\,|\omega, \overline{\sigma}_{SR}\,) - \frac{\min\limits_{m \notin \Sigma_S^-(\omega, \overline{\sigma}_{SR})} E\pi_R\,(m\,|\overline{\sigma}_{SR}, \omega\,) + \max\limits_{m \notin \Sigma_S^-(\omega, \overline{\sigma}_{SR})} E\pi_R\,(m\,|\overline{\sigma}_{SR}, \omega\,)}{2}. \end{aligned}$$

The first term on the RHS is the expected payoff that S believes R will receive if S chooses m, the second term, the equitable payoff, is the average of the highest and lowest payoff that S believes she could give to R by choosing an efficient message. If $K_{SR} > 0$ then S is said to be kind to R, if $K_{SR} < 0$ then S is unkind to R and if $K_{SR} = 0$ then S exhibits zero kindness towards R.

We analogously define how kind R is to S. Given $\overline{\sigma}_{RS}$ and message m, the kindness of R towards S if picking action a is given by:

$$\begin{aligned} & K_{RS}(a\,|m, \overline{\sigma}_{RS}\,) \\ = \;& E\pi_S(a\,|m, \overline{\sigma}_{RS}\,) - \frac{\min\limits_{a \notin \Sigma_R^-(m, \overline{\sigma}_{RS})} E\pi_S(a\,|m, \overline{\sigma}_{RS}\,) + \max\limits_{a \notin \Sigma_R^-(m, \overline{\sigma}_{RS})} E\pi_S(a\,|m, \overline{\sigma}_{RS}\,)}{2}. \end{aligned}$$

As above, if $K_{RS} > 0$ then R is said to be kind to S, if $K_{RS} < 0$ then R is unkind to S and if $K_{RS} = 0$ then R exhibits zero kindness towards S.

To capture an agent's reciprocity incentives, one also needs a measure of how kind she perceives her-co agent as being.

We first define how kind S perceives R as being towards S. Given $\overline{\sigma}_{SR}, \overline{\sigma}_{SRS}$, S's *perceived kindness of* R to S is given by:

$$K_{SRS}(\overline{\sigma}_{SR}, \overline{\sigma}_{SRS}) = \int_{\omega} \int_{m} f(\omega) P(m \,|\overline{\sigma}_{SRS}, \omega) K_{RS}(a(m, \overline{\sigma}_{SR}) \,|m, \overline{\sigma}_{SRS}) dm d\omega.$$

where $a(m, \overline{\sigma}_{SR})$ is the action that S anticipates that R takes after m. S's perception of R's kindness to S is a weighted average of the kindness of R conditional on each possible message, each kindness being weighted by the ex ante probability that S thinks that R assigns to it**.** As before, if $K_{SRS} > 0$ then S perceives R as kind to S, if $K_{SRS} < 0$... etc.

We now define how kind R perceives S as being towards R. Given $\overline{\sigma}_{RS}, \overline{\sigma}_{RSR}$ and m, R's *perceived kindness of* S to R is given as follows:

$$K_{RSR}(m, \overline{\sigma}_{RS}, \overline{\sigma}_{RSR}) = \int_{\omega} f(\omega \,|m, \overline{\sigma}_{RS}) K_{SR}(m \,|\omega, \overline{\sigma}_{RSR}).$$

In other words R uses the information about the state that is revealed by m to evaluate how kind she perceives S as being by sending message m. The sign of K_{SRS} has an interpretation analogous to that of K_{RSR}.

We may now define utilities, which are composed of two elements: material payoffs and reciprocity payoffs. The expected utility of S at information set ω given that she sends message m, conditional on beliefs $\{\overline{\sigma}_{SR}, \overline{\sigma}_{SRS}\}$ is given by:

$$U_S(m \,|\omega, \overline{\sigma}_{SR}, \overline{\sigma}_{SRS}) = \underbrace{E\pi_S \,(m \,|\omega, \overline{\sigma}_{SR})}_{\text{material payoff}} + \underbrace{\gamma K_{SR}(m \,|\omega, \overline{\sigma}_{SR}) K_{SRS}(\overline{\sigma}_{SR}, \overline{\sigma}_{SRS})}_{\text{reciprocity payoff}},$$

where $\gamma \geq 0$ is S's sensitivity to reciprocity. If $\gamma = 0$, then utility equals the material payoff. Notice that the product of K_{SR} and K_{SRS} enters the utility function; the implied sign-matching property captures reciprocity incentives. For example, suppose S perceives R as kind ($K_{SRS} > 0$), then S's utility increases if she is kind to R ($K_{SR} > 0$), but decreases if she is unkind to R ($K_{SRS} < 0$).

The expected utility of R at information set m given that she picks action a, conditional on beliefs $\{\overline{\sigma}_{RS}, \overline{\sigma}_{RSR}\}$, is given by:

$$U_R(a\,|m, \overline{\sigma}_{RS}, \overline{\sigma}_{RSR}) = \underbrace{E\pi_R\,(a\,|m, \overline{\sigma}_{RS})}_{\text{material payoff}} + \underbrace{\gamma K_{RS}(a\,|m, \overline{\sigma}_{RS})K_{RSR}(\overline{\sigma}_{RS}, \overline{\sigma}_{RSR})}_{\text{reciprocity payoff}},$$

where $\gamma \geq 0$ is R's sensitivity to reciprocity. Reciprocity incentives are represented analogously to how they appear in U_S.

Appending utility functions U_S and U_R to the cheap talk game described in Section 1.1 (with its implied type structure) gives a Bayesian psychological game (see Genakoplos et al. (1989) and Battigalli and Dufwenberg (2009) for formal definitions of psychological games of complete information; and Battigalli and Dufwenberg (section 6.2, 2009) and Attanasi et al. (section 3, 2016) for formal definitions of Bayesian psychological games).

Solution concept: Following Attanasi et al. (2016) we apply Perfect Bayesian Equilibrium (see also Battigalli et al. 2018). Such an equilibrium is a strategy profile (σ_S, σ_R) and a set of beliefs that satisfy the following. First, beliefs are consistent with strategies. Second, beliefs are updated via Bayes' rule whenever possible. Third, each strategy is sequentially rational given beliefs. Sequential rationality of S's strategy means that at each information set ω, any message picked by S with positive probability maximises $U_S(m\,|\omega)$. Sequential rationality of R's strategy means that at each information set m, any action picked by R with positive probability maximises $U_R(a\,|m)$.

Formally, it must be true that if m^* is sent with positive probability given ω, then

$$U_S(m^*\,|\omega, \sigma_R, \sigma_S) = \arg\max_{m \in M} U_S(m\,|\omega, \sigma_R, \sigma_S).$$

Furthermore, if a^* is picked with positive probability given message m, then it must be true that

$$U_R(a^*\,|m, \sigma_S, \sigma_R) = \arg\max_{a \in A} U_R(a\,|m, \sigma_S, \sigma_R).$$

Note that when $\gamma = 0$ the definition collapses to that of Perfect Bayesian Equilibrium for the material game.

2 Main Analysis

2.1 Model

We shall consider the following discrete cheap talk model that falls into the general environment outlined in Section 1.1. There are two possible states of the world, 0 and h, with prior distribution $\alpha, 1-\alpha$. S knows the state and can communicate with R via one stage of cheap talk messaging. The action space of S is $\{m_0, m_1\}$. The action space of R is $[0, h]$. The material payoff function of R is $\pi_R(a, \omega) = -|a - \omega|$, thus R's material payoff is higher the closer her action is to the state. The material payoff function of S is given as follows:

$$\pi_S(a, \omega) = \begin{cases} -|a - 0| \text{ if } \omega = 0, \\ -\left|a - a^*_{S,h}\right| \text{ if } \omega = h, \end{cases}$$

where $a^*_{S,h} \in [0, h)$. That is, S has identical material incentives to R if the state is 0, but prefers a lower action than R if the state is h. By increasing $a^*_{S,h}$, we make the material incentives of S and R more aligned. We call $\left|h - a^*_{S,h}\right|$ the bias of S. Assuming $a^*_{S,h} < \frac{h}{2}$, a key question is whether there exist equilibria in which S truthfully reveals to R that $\omega = h$. Indeed, assuming purely materially driven players, it would be against S's material interest to do so.

The following scenarios would match the model. Two individuals face a defendant who may be innocent or guilty, one of the two knowing the truth. While both agree that an innocent defendant ought to be released, they disagree on the optimal extent of punishment if the defendant is guilty, the informed party being (potentially very significantly) more lenient. Alternatively, consider an entrepreneur who knows whether or not her company's profits are high. She faces uninformed investors whose favoured policy conditional on disappointing results she considers excessively cautious. Other examples could be taken from doctor-patient, parent-child or friendship relations.

Applying the methodology outlined in Section 1.2, we now identify the effect of reciprocity preferences on truth-telling by analysing the equilibria of the Bayesian psychological game.

2.2 Positive truth-telling equilibria

In what follows we consider *truth-telling equilibria* (TTE) in which S sends message m_0 if the state is 0 and m_1 if it is h.[3] We distinguish between TTE featuring $a(m_1) \geq a(m_0)$, labelled *positive TTE*, and equilibria featuring $a(m_1) < a(m_0)$, *negative TTE*.

We first briefly recall the trivial conditions under which there exists a truth-telling equilibrium under pure material preferences, i.e. given $\gamma = 0$.

Observation 1 *Given $\gamma = 0$, there exists a TTE if and only if $a^*_{S,h} \geq \frac{h}{2}$. It features $a(m_0) = 0$ and $a(m_1) = h$.*

Proof: Trivial and thus omitted. ■

Assuming no reciprocity concerns, a TTE exists if and only if agents' material payoffs are sufficiently aligned. If there exists one, it is a positive TTE. In any TTE, R picks her material payoff maximising action given each message, and this action equals the state implied by a message. Second, conditional on such R-responses, truth-telling is optimal for S in both states if and only if $a^*_{S,h} \geq \frac{h}{2}$.

Next, we characterise the set of positive TTE given $\gamma > 0$. We first provide a lemma showing that in all positive TTE, R never chooses an inefficient action conditional on ω.

Lemma 1 *For all $\gamma > 0$, all positive TTE feature $a(m_0) = 0$ and $a(m_1) \geq a^*_{S,h}$.*

Proof: See appendix. ■

Consider a positive TTE. Conditional on m_0 (and the state thus being 0), R perceives S as exhibiting zero kindness (i.e. $K_{RSR}(m_0, \overline{\sigma}_{RS}, \overline{\sigma}_{RSR}) = 0$). Indeed, either m_1 is inefficient at $\omega = 0$ (if $a(m_1) > a(m_0)$) or R believes that S believes that R's material payoff is independent of m (if $a(m_1) = a(m_0)$). Given $K_{RSR}(m_0, .) = 0$, R has no reciprocity incentives, and thus chooses $a(m_0) = 0$ to maximise her material payoff.

Conditional on m_1 (and thus state h), R receives a weakly lower material payoff from $a(m_0)$ than $a(m_1)$ (as we assume $a(m_1) \geq a(m_0)$). Consider S's material payoffs to identify

[3] Equilibria where S sends message m_1 if the state is 0 and m_0 if the state is h are equivalent. We thus focus on TTE as defined in the text throughout.

whether R ever chooses an inefficient action. If S obtains a strictly lower payoff from $a(m_0)$ than $a(m_1)$ then m_0 is inefficient, and thus $K_{RSR}(m_1, .) = 0$. Then R's utility equals her material payoff, and choosing any a less than $a^*_{S,h}$ thus gives R a lower utility than $a^*_{S,h}$. If S obtains a weakly higher material payoff from $a(m_0)$ than $a(m_1)$ then no message is inefficient and thus $K_{RSR}(m_1, .) > 0$. Given m_1, R knows the state (h) and both agents derive a strictly higher material payoff from $a^*_{S,h}$ than any a less than $a^*_{S,h}$. Hence both R's material and reciprocity incentives imply that $a^*_{S,h}$ yields a higher utility than any lower action. Thus an inefficient action is never chosen in a positive TTE.

Our first proposition examines whether a positive TTE exists given reciprocity concerns ($\gamma > 0$) and relatively high material preference alignment ($a^*_{S,h} \geq \frac{h}{2}$, for which we know that a TTE exists even with $\gamma = 0$, by Observation 1). We find that reciprocity incentives can preclude the existence of a positive TTE.

Proposition 1 *For all $\gamma > 0$, given $a^*_{S,h} \in [\frac{h}{2}, h)$, there exists a positive TTE iff $\gamma \leq \frac{2\left(2a^*_{S,h} - h\right)}{h(1-\alpha)\left(h - a^*_{S,h}\right)}$. If there exists one it is unique and features $a(m_0) = 0$ and $a(m_1) = h$.*

Proof: See appendix. ■

If material incentives are sufficiently aligned ($a^*_{S,h} \geq \frac{h}{2}$) but reciprocity concerns are high $\left(\gamma > \frac{2\left(2a^*_{S,h} - h\right)}{h(1-\alpha)\left(h - a^*_{S,h}\right)}\right)$ then no positive TTE exists. We know from Observation 1 that here, material incentives motivate truth-telling, so reciprocity incentives must motivate lying. To see why this is true, consider first R's incentives. Given Lemma 1 and given low bias ($a^*_{S,h} \geq \frac{h}{2}$), in each state one of S's messages is conditionally inefficient. R thus perceives S's kindness as zero and thus simply maximises her material payoff in each state ($a(m_0) = 0$ and $a(m_1) = h$). Given this, consider S's reciprocity incentives at $\omega = h$. First, S perceives R as unkind. R's kindness after m_0 is zero and R is unkind after m_1 as R chooses $a = h$, giving S her lowest material payoff among the set of conditionally efficient actions ($a \in [a^*_{S,h}, h]$). Second, S can reciprocate R's unkindness by deviating to m_0, thereby inducing R to choose action 0 despite $\omega = h$, thus reducing R's material payoff. Such a deviation also reduces S's material payoff, so S only deviates if she cares

enough about reciprocity, i.e. if $\gamma > \frac{2\left(2a^*_{S,h}-h\right)}{h(1-\alpha)\left(h-a^*_{S,h}\right)}$. Hence greater reciprocity concerns here preclude truth-telling.

Note the intuitive comparative statics of the identified critical value of γ. A lower α means a higher ex-ante chance of the state in which preferences are misaligned and R is unkind. When α decreases, S thus perceives R as more unkind. In consequence, a lower α reduces the critical value of γ and makes truth-telling more difficult to achieve. In contrast, more material preference alignment (higher $a^*_{S,h}$) mechanically reduces S's perception of R's unkindness given $a(m_1) = h$, thus increasing the critical value of γ.

Having characterised outcomes given $a^*_{S,h} \geq \frac{h}{2}$, we now concentrate on the more interesting case of $a^*_{S,h} < \frac{h}{2}$, where truth-telling was impossible absent reciprocity concerns.

Proposition 2 ***Positive truth-telling equilibrium***

For all $\gamma > 0$,

(a) *Given $a^*_{S,h} \in \left[0, \frac{h}{3}\right]$, there exists a positive TTE iff $\gamma \in \left[\frac{2(3-2\alpha)}{2a^*_{S,h}+(1-\alpha)(h+a^*_{S,h})}, \frac{1}{a^*_{S,h}}\right]$. If there exists one, it is unique and features $a(m_1) = \frac{2}{\gamma}$.*

(b) *Given $a^*_{S,h} \in \left(\frac{h}{3}, \frac{h}{2}\right)$, there exists no positive TTE, whatever the value of γ.*

Proof: See appendix. ■

Unlike our previous result, this result demonstrates how reciprocity can foster information transmission. When truth-telling is impossible with pure material payoffs ($a^*_{S,h} < \frac{h}{2}$), it becomes feasible with reciprocity concerns if the bias is high enough and reciprocity concerns are intermediate. In this equilibrium S is kind to R by revealing information that is valuable to R (but could be detrimental to S) and R reciprocates the favour by skewing her action choice towards S's materially optimal action. Figure 1 illustrates existence conditions for positive TTE via a parameterised example.

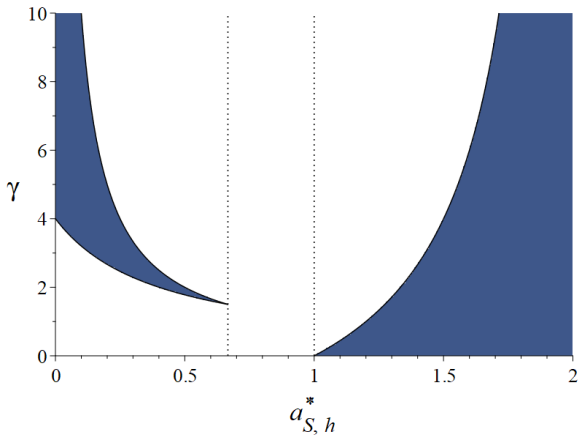


Figure 1: Existence of a positive TTE.

Notes: For $h = 2$ and $\alpha = \frac{1}{2}$, the shaded areas depict values of $\{a^*_{S,h}, \gamma\}$ such that a (unique) positive TTE exists.

The example in Figure 1 illustrates both Propositions 1 and 2. When the bias is low enough ($a^*_{S,h} \geq \frac{h}{2}$), a positive TTE exists if reciprocity concerns are low $\left(\gamma \leq \frac{2\left(2a^*_{S,h}-h\right)}{h(1-\alpha)\left(h-a^*_{S,h}\right)}\right)$ (Prop. 1). When the bias is moderate ($a^*_{S,h} \in (\frac{h}{3}, \frac{h}{2})$), a positive TTE does not exist (Prop 2.(b)).[4] When bias is high enough ($a^*_{S,h} \leq \frac{h}{3}$), Proposition 2(a) states that a positive TTE exists if reciprocity sensitivity is intermediate, i.e

$$\gamma \in \left[\frac{2(3-2\alpha)}{2a^*_{S,h} + (1-\alpha)(h + a^*_{S,h})}, \frac{1}{a^*_{S,h}}\right].$$

For more intuition behind Proposition 2, we consider each of the two outlined intervals of $a^*_{S,h}$ in turn.

[4]One might be concerned that no PBE exists for such parameters. Indeed, for dynamic games of complete information, Dufwenberg and Kirchsteiger (2004) note that a sequential equilibrium may not exist when Rabin's definition of an efficient strategy is used (see pp. 288-9, Dufwenberg and Kirchsteiger (2004)). We use Rabin's definition, but a PBE always exists in our cheap talk game. It is straightforward to show that a "babbling" PBE always exists: For example, where S randomizes uniformly across messages and R chooses a to maximise her material payoff given her prior.

We leave it for future work to determine whether equilibrium existence is less problematic with Rabin's efficiency definition in dynamic games of incomplete information more generally.

For $a^*_{S,h} \in (\frac{h}{3}, \frac{h}{2})$, Observation 1 establishes that material incentives alone cannot sustain truth-telling. For reciprocity incentives to encourage truth-telling, sufficient mutual kindness needs to be involved. This is however not feasible for the following reason. In any putative positive TTE, R's action after m_1 must be skewed towards S's materially preferred action in order for R to be kind to S. However, since the bias is relatively low, the kind action $a(m_1) = \frac{2}{\gamma}$ chosen by R implies that both agents' material payoffs are higher following m_1 than m_0 when $\omega = h$. In other words, m_0 becomes conditionally inefficient when $\omega = h$. But then R's perception of S's kindness from m_1 is zero, and thus R has no reciprocity incentive to be kind to S. In other words, in the putative positive TTE, R is so kind in response to m_1 that she de facto prohibits S from being kind when $\omega = h$. One might speak of a zero-sum game in kindness or of a self-defeating kindness dynamic. R has a commitment (self-control) problem: In terms of inducing information revelation by S, she would benefit if she could commit to less generosity in response to the favour embedded in m_1.

Consider now $a^*_{S,h} \in [0, \frac{h}{3}]$. Given the above intuition, it should be clear why sufficient bias can facilitate the existence of a positive TTE: It alleviates R's commitment problem, ensuring that R is never so kind that she precludes S from being kind. Note that a positive TTE requires that γ is intermediate. If it is too high, then the self-defeating mechanism described above reemerges. If γ is too low, material incentives dominate and bias is too high to sustain truth-telling. Note that increasing the probability of h reduces the lower bound of the permitted interval of values of γ.[5] For a given kindness of R conditional on m_1, the higher the prior probability of h, the more likely the state in which R is kind and thus the higher S's perception of R's kindness. In consequence, a lower reciprocity concern is sufficient to motivate S to truth-tell.

2.3 Negative truth-telling equilibria

We now turn to negative TTE. These feature the counterintuitive relation $a(m_1) < a(m_0)$. Recall that both agents' materially optimal decision rules specify a higher action for $\omega = h$

[5] Note that $d\frac{2(3-2\alpha)}{2a^*_{S,h}+(1-\alpha)(h+a^*_{S,h})}/d\alpha > 0 \Rightarrow a^*_{S,h} < \frac{h}{3}$, which is true for the case under consideration.

than for $\omega = 0$.

Our next result demonstrates that with reciprocity, negative TTE may exist.[6] Let

$$\begin{aligned} \varphi &= \frac{2\alpha + (1-\alpha)h\gamma + (1+3\alpha)\gamma a^*_{S,h}}{2\gamma(1+2\alpha)}, \\ \text{and } \eta &= 3(1-\alpha)\left(h - a^*_{S,h}\right) - 2\alpha h. \end{aligned}$$

Proposition 3 ***Negative truth-telling equilibrium***

(a) *There exists a negative TTE featuring $a(m_0) = h$ and $a(m_1) = 0$ if and only if $a^*_{S,h} > \frac{h}{2}$ and $\gamma \geq \frac{1}{2a^*_{S,h} - h}$.*

(b) *If $\alpha < \frac{1}{2}$, then for every $\widetilde{a} \in \left[\max\left\{\frac{1}{\gamma}, \varphi\right\}, a^*_{S,h}\right]$, there exists a negative TTE featuring $a(m_0) = \widetilde{a}$ and $a(m_1) = a(m_0) - \frac{1}{\gamma}$.*

(c) *There exists a negative TTE featuring $a(m_0) = h$ and $a(m_1) = h - \frac{2}{\gamma}$ if and only if either $\min\left\{\frac{2}{h}, \frac{1}{h - a^*_{S,h}}\right\} < \gamma < \frac{2}{h - a^*_{S,h}}$ and $\gamma\eta \leq 2(1-2\alpha)$ or $\gamma \geq \frac{2}{h - a^*_{S,h}}$.*

(d) *All negative TTE feature $\{a(m_0), a(m_1)\}$ as in either (a), (b) or (c).*

Proof: See online appendix. ■

Negative TTE feature gift-exchange in the germanic sense of the word gift, i.e. poison. Such equilibria exhibit what could be termed a self-fulfilling prophecy of conflict. In equilibrium, R responds to S's information in a way that is hurtful both to herself and to S. The latter responds by being unkind, which involves truth-telling given R's paradoxical response to information. Notice that here, telling the truth is *not* kind. If S wanted to be kind to R, she should tell a "benevolent lie".

As negative TTE do not exist for $\gamma = 0$ (Observation 1), agents' reciprocity incentives must drive truth-telling. Notice that the existence conditions in Proposition 3 only hold if γ is sufficiently high.[7]

[6] For clarity, we provide only sufficient conditions for the existence of negative TTE in the main text (Proposition 3). Necessary and sufficient conditions are stated in Proposition 3* in the online appendix. Proposition 3 is a trivial corollary of Proposition 3*. The qualitative features highlighted following Proposition 3 also hold for Proposition 3*.

[7] For (a), we require $\gamma \geq \frac{1}{2a^*_{S,h} - h}$. For (b), we require $\gamma \geq \frac{1}{h}$, otherwise the equilibrium would need $a(m_0) > h$, which is impossible. For (c), we require $\gamma > \min\left\{\frac{2}{h}, \frac{1}{h - a^*_{S,h}}\right\}$.

Interestingly, a lower bias facilitates the existence of negative TTE. Consider the equilibrium in Proposition 3(a). This only occurs when $a^*_{S,h} > \frac{h}{2}$ and the higher $a^*_{S,h}$, the lower the requirement on γ. To see why, suppose that the bias is very *high*, i.e. $a^*_{S,h}$ is small. If R learns that $\omega = h$, there is no action that gives both players a low material payoff. Thus the poisonous gift-exchange we described earlier is impossible. The effect of bias is analogous for the negative TTE identified in parts (b) and (c) of the result.[8]

Recall that positive TTE did not exist when $a^*_{S,h} \in (\frac{h}{3}, \frac{h}{2})$ (Proposition 2). Proposition 3(c) in contrast establishes that a negative TTE can exist for $a^*_{S,h} \in (\frac{h}{3}, \frac{h}{2})$, provided $\gamma \geq \frac{2}{h - a^*_{S,h}}$. The reason why a positive TTE could not exist for $a^*_{S,h} \in (\frac{h}{3}, \frac{h}{2})$ was that after m_1, R's reciprocation negated S's kindness, in turn killing R's incentive to be kind. By contrast, the mutual unkindness sustaining negative TTE is not self-defeating. The equilibrium considered in Proposition 3(c) features $a(m_0) = h$ and $a(m_1) = h - \frac{2}{\gamma}$. As γ increases, R continually increases here unkindness after m_1 without converging to a point where m_1 is not unkind anymore, which would require $a(m_1) \geq a(m_0)$.

The fact that negative TTE exist for intermediate bias (in contrast to positive TTE) might be deemed a virtue. The negative gift-exchange on which these rely is however very costly in material terms.

2.4 Welfare in truth-telling equilibria

Having fully characterised the set of TTE, we now provide a statement concerning welfare in TTE, which we define as the sum of ex-ante material payoffs. Corollary 1 follows from Propositions 1-2 and 3* (in the online appendix).

Corollary 1 *If both a positive and negative TTE exist, then welfare is strictly higher in the positive TTE.*

Proof: See appendix. ■

[8] For part (b), note that since $d\varphi / da^*_{S,h} < 1$, the interval $\left[\max\left\{\frac{1}{\gamma}, \varphi\right\}, a^*_{S,h}\right]$ is larger, and a larger range of $a(m_0)$ occur in equilibrium. For (c), a high $a^*_{S,h}$ increases the interval of reciprocity sensitivities satisfying $\min\left\{\frac{2}{h}, \frac{1}{h - a^*_{S,h}}\right\} < \gamma < \frac{2}{h - a^*_{S,h}}$ and also increases the set of parameter values satisfying $\gamma\eta \leq 2(1 - 2\alpha)$ since $d\eta / da^*_{S,h} < 0$.

The result is very intuitive given that positive TTE make a more efficient use of information from all parties' perspective.

3 Alternative environments

In this section we consider whether reciprocity can motivate truth-telling in three important, alternative environments. The first is a setting where both state and message spaces are continuous. The second assumes an alternative conception of reciprocity preferences closely following Dufwenberg and Kirchsteiger (2004). The key idea is that in constructing kindness reference points, the set of efficient strategies is now defined independently of beliefs. In the third setup, R has reciprocity preferences while S is motivated either solely by material payoffs or by an unconditional desire to be kind (which may be relevant for understanding how a socially responsible firm (S) interacts with a reciprocal consumer (R)).

3.1 A continuous model

Suppose that the state ω is drawn from a distribution F with pdf f which has full support on the domain Ω. The domain Ω can be bounded, in which case we denote it $[\underline{\omega}, \overline{\omega}]$, or unbounded $(-\infty, +\infty)$. The message set is $M = \Omega$. The action set of R is $\mathbb{R}$.

We consider two different material payoff functions. In the *linear case*, the material payoff function of S is $\pi_S(a, \omega) = -|a - (\omega + b)|$, for $b > 0$, while that of R is $\pi_R(a, \omega) = -|a - \omega|$. In the *quadratic case*, material payoff functions are $\pi_S(a, \omega) = -(a - (\omega + b))^2$, for $b > 0$, and $\pi_R(a, \omega) = -(a - \omega)^2$. Note that in both cases, given information I, R's ideal action is $E[\omega | I]$ and S's ideal action is $E[\omega | I] + b$.

Given material payoffs and the structure of the game, it is straightforward to define kindness, perceived kindness and hence reciprocity payoffs as described in Section 1.2. In addition to material and reciprocity components to preferences, we assume that S is lying averse. As we shall see shortly, in a simple continuous model neither material nor reciprocity preferences motivate perfect truth-telling. We thus need this additional (though

potentially infinitesimal) incentive to support full truth-telling equilibria and be able to study their comparative statics properties w.r.t. bias and γ.[9]

We focus on truth-telling equilibria (TTE) featuring $m = \omega$ for every ω, which is w.l.o.g within the class of TTE. We model lying costs as follows. S dislikes to mislead R and incurs a direct psychological cost from doing so (Kartik 2009; Gneezy et al. 2018[10]**)**. Given $\overline{\sigma}_{SRS}$, define the sincere message $m(\omega, \overline{\sigma}_{SRS})$ as the message that minimises $E(\omega \,|m, \overline{\sigma}_{SRS}) - \omega$ among all existing messages (we assume uniqueness, which will hold true). Note that in the studied equilibria, $E(\omega \,|m, \overline{\sigma}_{SRS})$ is pinned down by Bayes rule for any m. Given ω, we say that S misleads R if she sends $m' \neq m(\omega, \overline{\sigma}_{SRS})$. Define

$$\Delta(m', \overline{\sigma}_{SRS}) = \left| E(\omega \,|m', \overline{\sigma}_{SRS}) - E(\omega \,|m(\omega, \overline{\sigma}_{SRS}), \overline{\sigma}_{SRS}) \right|,$$

which thus measures the distance between the belief induced by the non-misleading message and the message actually sent by S. If S misleads R by sending m', she bears a psychological cost $\tau\Delta(m', \overline{\sigma}_{SRS})$, where $\tau > 0$. The expected utility function of S at information set ω given that she sends message m, conditional on beliefs $\{\overline{\sigma}_{SR}, \overline{\sigma}_{SRS}\}$, is thus given by:

$$\begin{aligned} & U_S(m \,|\omega, \overline{\sigma}_{SR}, \overline{\sigma}_{SRS}) \\ = \;& E\pi_S(m \,|\omega, \overline{\sigma}_{SR}) + \gamma K_{SR}(m \,|\omega, \overline{\sigma}_{SR}) K_{SRS}(\overline{\sigma}_{SR}, \overline{\sigma}_{SRS}) - \tau\Delta(m, \overline{\sigma}_{SRS}). \end{aligned}$$

The expected utility function maximised by R and the equilibrium concept are as defined in Section 1.2. We start by identifying conditions under which a TTE exists in the absence of reciprocity payoffs.

Observation 2 ***TTE without reciprocity preferences***

Let $\gamma = 0$.

***(a)** With linear material payoffs, there exists a TTE if and only if $\tau \geq 1$.*

***(b)** With quadratic material payoffs, there exists a TTE if and only if $\tau \geq 2b$.*

[9]The asymmetry in preferences between S and R may seem ad hoc. However, it would be redundant to introduce lying-aversion for R since she does not communicate (see Attanasi et al. (2016) for more examples of role-dependent preferences).

[10]Given our focus on truth-telling equilibria, *perceived cheating aversion* (Dufwenberg and Dufwenberg 2018) would not affect S's incentives as R believes S tells the truth in such equilibria.

Proof: See online appendix. ■

In the absence of reciprocity, a truth-telling equilibrium trivially exists if S is sufficiently lying averse. Since R learns the state in a TTE she chooses the action that matches the state. Given this, S clearly has no incentive to lie downwards, which both hurts her material payoff and implies a lying cost. The same is true if S lies sufficiently high upwards (inducing any action $a \geq \omega + 2b$). For messages that induce actions in $(\omega, \omega + 2b)$, S receives a strictly higher material payoff relative to truth-telling, but if she is sufficiently lying averse (τ sufficiently high), the lying cost dominates the material gain. Note that with quadratic material payoffs, the larger S's bias is (higher b), the larger the lying aversion required for a TTE to exist.

We now examine how reciprocity incentives affect the existence of TTE. We focus on a natural subclass of TTE which we label *simple truth-telling equilibria*. In these, S always sends $m = \omega$ and for any m, R picks action $m + c$, for some $c \in [0, b]$. For any announced value of the state, R picks a compromise action located between the two parties' preferred actions, and the degree of the compromise is constant across states.

Proposition 4 ***Linear material payoff functions***

Assume linear material payoff functions. If there exists a simple TTE, then it features either $c = 0$ *or* $c = \frac{b}{2} - \frac{1}{\gamma} > 0$.

(a) *A simple TTE featuring* $c = 0$ *exists if and only if* $\gamma \leq \gamma^*(b) = \frac{2}{b}$ *and* $\tau \geq \frac{b}{2}\gamma + 1$.

(b) *A simple TTE featuring* $c = \frac{b}{2} - \frac{1}{\gamma} \in (0, b)$ *exists if and only if* $\gamma > \gamma^*(b) = \frac{2}{b}$ *and* $\tau \geq 2$.

Proof: See online appendix. ■

A noteworthy implication of the above is that it is strictly more difficult to achieve truth-telling with reciprocity incentives ($\gamma > 0$) than without ($\gamma = 0$), if we limit ourselves to simple TTE. Indeed, if $\gamma = 0$ truth-telling is feasible with $\tau \geq 1$ (Observation 2(a)) while if $\gamma > 0$, a simple TTE requires $\tau > 1$, whichever putative simple TTE one considers (Proposition 4).

The intuition is that in a simple TTE with $c \in [0, b]$ neither mutual kindness nor mutual unkindness are possible given $\gamma > 0$. To understand why this is the case, reason as

follows. First, given ω, choosing (either implicitly via sending a message in S's case, or explicitly in R's case) any action outside of $[\omega, \omega + b]$ is inefficient. Within the efficient interval $[\omega, \omega + b]$, S and R play a zero-sum game in material payoffs, with S's material payoff increasing in a and R's decreasing in a. Thus in equilibrium, if $c \neq \frac{b}{2}$ then one of the players is kind and the other is unkind. If instead $c = \frac{b}{2}$, both players exhibit zero kindness.

In the simple truth-telling equilibria identified in Proposition 4, R is unkind to S and S is kind to R. The kindness-neutral action of R would be $a(m) = m + \frac{b}{2}$, but R takes action m in (a) and instead $m + \frac{b}{2} - \frac{1}{\gamma} < m + \frac{b}{2}$ in (b). As to S's kindness, in both equilibria $K_{RSR} = -c - \left(\frac{-b}{2}\right)$, yielding $K_{RSR} = \frac{b}{2} > 0$ in (a) and $K_{RSR} = \frac{1}{\gamma} > 0$ in (b).

Given the asymmetry in kindness, reciprocity incentives actually hinder rather than motivate truth-telling, as S has an incentive to deviate to $m' > \omega$ in order to reciprocate R's unkindness. S has to overcome not only material incentives to lie but also psychological incentives relating to the reciprocation of R's unkindness. Thus a higher lying aversion sensitivity τ is needed to support the simple TTE if $\gamma > 0$ than if $\gamma = 0$. This echoes a main insight of Propositions 1 and 2, which is that information transmission can be hurt by a greater γ.

We attach a final remark on the effect of bias b in Proposition 4(b). Note that the critical value $\gamma^*(b)$ is interestingly decreasing in b. Given $c = \frac{b}{2} - \frac{1}{\gamma}$, increasing b means that the condition $c \geq 0$ can be guaranteed for a lower minimal γ.

We now consider the case of quadratic losses.

Proposition 5 ***Quadratic material payoff functions***

Assume quadratic material payoff functions.

(a) *Given b and $\tau > 2b$, a simple TTE featuring $c = 0$ exists if γ is sufficiently small.*

(b) *Given b and $\tau < 2b$, there is a finite $\gamma^*(b) = \frac{4}{b^2}$ such that a simple TTE featuring $a(m) = m + c$, where $c \in [0, b]$, exists if and only if γ is sufficiently close to $\gamma^*(b)$. If such an equilibrium exists, it is unique and features $c \in (0, b)$.*

Proof: See online appendix. ■

When replacing linear with quadratic material payoff functions, some insights remain

(part (a)) while new insights emerge (part (b)). Part (a) shows that when lying aversion alone is sufficient to motivate truth-telling ($\tau > 2b$, Observation 2(b)), reciprocity incentives once again hinder rather than foster simple TTEs. The intuition is similar to the one seen previously and relates to the difficulty of achieving mutual kindness or mutual unkindness (e.g. Proposition 4).

Part (b) is qualitatively different from Proposition 4 in that it shows that when truth-telling is impossible without reciprocity ($\tau < 2b$), reciprocity incentives can facilitate existence of simple TTEs. As in our binary model (see Proposition 2(a)), reciprocity concerns must however be intermediate (γ close enough to $\frac{4}{b^2}$). Truth-telling can be achieved with arbitrarily low lying aversion as $\gamma \to \frac{4}{b^2}$.

A fundamental reason why reciprocity can incentivise truth-telling with quadratic material payoffs is that concavity (in contrast to linearity) of the loss functions implies that truth-telling can be mutually kind.[11] As with linear payoffs, the set of efficient actions conditional on ω remains $[\omega, \omega + b]$, corresponding to a simple TTE with $c \in [0, b]$. However with quadratic payoffs, $c = \frac{b}{2}$ no longer implies zero-kindness. To see why, consider R's choice of action. In a putative simple TTE, the material payoff that R could impose on S by choosing $a = \omega + c$ with $c = 0$ is very low (due to concavity). Given this, there is $\underline{c} = b\left(\frac{\sqrt{2}-1}{\sqrt{2}}\right) < \frac{b}{2}$ such that R is kind to S as long as $c \in (\underline{c}, b]$. Analogously, there exists some $\overline{c} = \frac{b}{\sqrt{2}} > \frac{b}{2}$ such that S is kind to R provided that $c \in (0, \overline{c}]$. In other words, mutual kindness is possible in a simple TTE if and only if $c \in (\underline{c}, \overline{c})$. It is this mutual kindness that incentivises non-deviation from truth-telling in a simple TTE. For each b, γ, it can be shown that there is a unique value of c (denoted $c^*(b, \gamma)$) that may constitute a simple TTE. The value $c^*(b, \gamma)$ is the solution to a fixed point problem: Given a perceived kindness of S computed on the basis of $c = c^*(b, \gamma)$, R's best response to m is indeed $a^*(m) = m + c^*(b, \gamma)$. Note that $c^*(b, \gamma)$ does not have a simple closed form and is thus not stated explicitly.

Incentives change as a function of γ in the following way. As γ increases, $c^*(b, \gamma)$ increases monotonically, starting from 0 and tending to $\frac{b}{\sqrt{2}}$. A first implication is that

[11] Given the critical role concavity plays in driving Proposition 5(b), the result is presumably an instance of a general finding that reciprocity can motivate truth-telling for general concave material payoffs.

S is always kind in a simple truth-telling equilibrium. What changes as γ changes is the kindness of R: There is a threshold value of γ (call it γ_1) such that R is kind in the putative simple TTE if and only if $\gamma > \gamma_1$. We now describe S's incentives as γ increases, in the putative simple TTE. For γ small, R is unkind, so that both material and reciprocity payoffs incentivise S to deviate to $m' > \omega$. As γ crosses γ_1 from below, R becomes kind, which now generates a trade-off for S. Material payoffs incentivise a deviation to $m' > \omega$ while reciprocity payoffs incentivise a deviation to $m' < \omega$ (effectively a benevolent lie). As γ increases and $c^*(b, \gamma)$ increases in consequence, the first incentive continuously weakens while the second continuously strenghtens. For some critical value ($\gamma^*(b) = \frac{4}{b^2}$), the two effects exactly counterbalance each other and S would be willing to truth-tell even with no lying cost. In the limit for $\gamma \to \infty$, S's material payoff in the simple TTE converges to $-\left(b - \frac{b}{\sqrt{2}}\right)^2$, R's kindness converges to a positive constant and S's kindness conditional on truth-telling converges to 0. Trivially, beyond some threshold value of γ, S favours deviating to $m' < \omega$. The interpretation is that S's drive to reciprocate R's kindness becomes so strong that she favors $m' < \omega$ with the aim of tricking R into making a smaller concession, thereby increasing R's material payoff "against her will". S's good intentions thus hurt her ability to commit to truth-telling.

We now comment on the role of bias b. Note that $\gamma^*(b)$ is decreasing in b, meaning that more bias decreases the minimal value of γ required to ensure a simple TTE. The intuition, as in the binary model, is that a larger b inflates the size of the cake to be shared (and thus of potential gifts), thereby magnifying the relative role of reciprocity concerns w.r.t. that of material payoffs. To the extent that we consider large values of γ to be unrealistic, this provides a sense in which perfect truth-telling (in the form of simple TTE) is easier to achieve, the higher S's bias. Finally, note that the distribution of ω plays no role in our condition. The reason is that in simple TTE, the kindness of players is the same in all states. This stands in contrast to the binary model examined in the previous section.

3.2 An alternative kindness reference point

We now return to the binary model studied in Sections 1-2 to examine the implications of an alternative conception of reciprocity preferences. When defining the kindness ref-

erence point (see equitable payoff in Section 1.2), we assumed that agents use the set of efficient strategies *conditional on their beliefs*. This approach has been used by several authors modelling intention-based reciprocity (Rabin 1993; Netzer and Schmutzler 2014; Bierbrauer and Netzer 2016).

An alternative approach to defining the kindness reference point proposed by Dufwenberg and Kirchsteiger (2004) is to use the set of efficient strategies *independent of beliefs* (see Dufwenberg and Kirchsteiger 2004 and Dufwenberg and Kirchsteiger 2018 for arguments in support of this approach).

Given the lack of empirical evidence on reciprocity preferences in games of incomplete information, we leave it to future research to determine which reference point is more empirically relevant. For completeness, we now characterise the implications of defining efficient strategies independently of beliefs.

First consider the efficiency of S's messages. The impact of a message on payoffs is always dependent on how R reacts to the message. Since the set of efficient strategies is not conditioned on (R's belief about) S's belief about R's strategy, neither message is Pareto dominated. Thus all of S's strategies are efficient.

Similarly, all of R's strategies are efficient. Note that R's set of efficient strategies is not conditioned on (S's belief about) R's beliefs about S's strategy and thus not conditioned on R's updated belief about the state after messages. An inefficient strategy would have to specify taking an action that is Pareto dominated by another action in both states. Since there is no such action, all of R's strategies are efficient.

Given this definition of efficient strategies, we obtain the following characterisation of the set of positive TTE.

Proposition 6 ***Positive TTE with belief-independent efficiency***

(a) *There exists a positive TTE featuring* $a(m_0) = 0$ *and* $a(m_1) = a^*_{S,h}$ *iff* $\gamma \geq \frac{2}{a^*_{S,h}}$.

(b) *There exists a positive TTE featuring* $a(m_0) = 0$ *and* $a(m_1) = \frac{2}{\gamma}$ *iff* $\gamma \in \left(\frac{2}{h}, \frac{2}{a^*_{S,h}}\right)$, *and either*

(i) $a^*_{S,h} \geq \frac{h}{2}$ *and* $\alpha\frac{h}{2} + (1-\alpha)\left(\frac{3a^*_{S,h}}{2} - \frac{2}{\gamma}\right) \geq \frac{1}{\gamma} - a^*_{S,h}$, *or,*

(ii) $a^*_{S,h} < \frac{h}{2}$ *and* $\alpha\frac{h}{2} + (1-\alpha)\left(\frac{h+a^*_{S,h}}{2} - \frac{2}{\gamma}\right) \geq \frac{1}{\gamma} - a^*_{S,h}$.

(c) *There exists a positive TTE featuring* $a(m_0) = 0$ *and* $a(m_1) = h$ *iff* $\gamma \in (0, \frac{2}{h}]$ *and either*

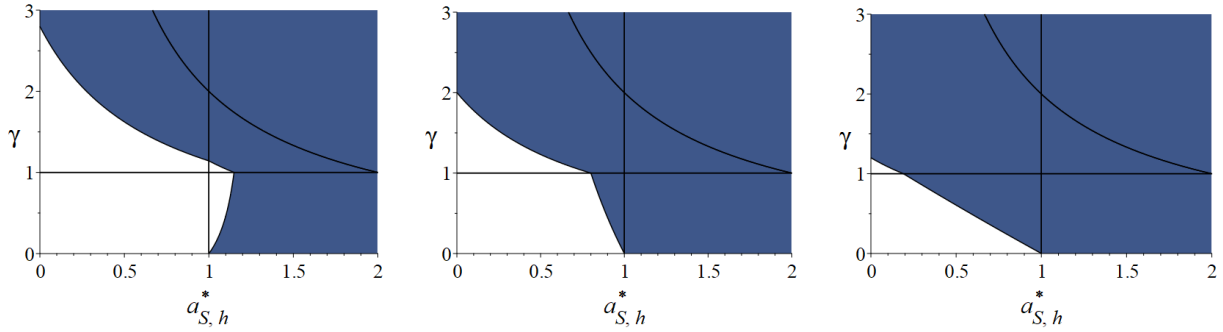
(i) $a^*_{S,h} \geq \frac{h}{2}$ *and* $\gamma\left(\alpha\frac{h}{2} + (1-\alpha)\left(\frac{3a^*_{S,h}}{2} - h\right)\right) \geq 1 - \frac{2a^*_{S,h}}{h}$, *or,*

(ii) $a^*_{S,h} < \frac{h}{2}$ *and* $\gamma\left(\alpha\frac{h}{2} - (1-\alpha)\left(\frac{h-a^*_{S,h}}{2}\right)\right) \geq 1 - \frac{2a^*_{S,h}}{h}$.

(d) *There exist no other positive TTE.*

Proof: See online appendix. ■

To help understand Proposition 6, consider the figures below which illustrate the existence conditions via a parameterised example.



(a) $\alpha = 0.1$ (b) $\alpha = 0.5$ (c) $\alpha = 0.9$

Figure 2: Existence of a positive TTE with belief-independent efficiency.

Notes: For $h = 2$ the shaded areas depict values of $\{a^*_{S,h}, \gamma\}$ such that a positive TTE exists. The value of α differs across panel (a)-(c) as labelled.

Notice that the conditions for existence of a positive TTE are mostly qualitatively different from those we identified with the belief-dependent definition of efficient strategies (Propositions 1 and 2).

We had found that a greater reciprocity concern was often detrimental for the existence of positive TTE (Propositions 1 and 2(a)). That is typically not the case here. Indeed, notice that by Proposition 6(a), for any material preference alignment, there exists a positive TTE if $\gamma \geq \frac{2}{a^*_{S,h}}$, which holds if γ is sufficiently high. For lower reciprocity concerns ($\gamma \in \left(0, \frac{2}{a^*_{S,h}}\right)$, parts (b) and (c)), most of the existence conditions are more likely to hold if γ is higher.

The intuition for the difference in the effect of γ is that the belief-independent efficiency definition allows actions and messages to be perceived as kinder than they are under the belief-dependent efficiency definition. This creates the possibility of truth-telling underpinned by mutual kindness. To illustrate, recall Proposition 2(a) which used the belief-dependent efficiency definition. For $a^*_{S,h} < \frac{h}{3}$, if γ is sufficiently high, then in response to truth-telling R has an incentive to take an action that is so kind that it makes S's untruthful message at $\omega = h$ inefficient. This implies that it is impossible for S to actually be kind to R and that mutual kindness cannot incentivise truth-telling as it is self-defeating. With a belief-independent efficiency definition, this logic breaks down (m_0 does not become inefficient at $\omega = h$) and thus truth-telling supported by mutual kindness is feasible.

Despite the above, even with the belief-independent efficiency definition, greater reciprocity concerns can preclude truth-telling. This is easily seen in Figure 2, panel (a). Consider $a^*_{S,h} \in (1, \frac{54}{47}]$. There exist two critical values of γ, given by $0 < \gamma_1(a^*_{S,h}) < \gamma_2(a^*_{S,h})$, such that a positive TTE exists iff $\gamma \leq \gamma_1(a^*_{S,h})$ or $\gamma \geq \gamma_2(a^*_{S,h})$. Here, as in our main analysis, increased reciprocity concerns can hurt the existence of positive TTE.

To understand the intuition for the non-monotonicity, first note that S is kind to R in each of the equilibria identified in Proposition 6 as $a(m_1) \geq a^*_{S,h}$ and we are considering $a^*_{S,h} > \frac{h}{2}$. Second, note that R is kind to S only if $\gamma \geq \gamma_2(a^*_{S,h})$. Finally, consider each of the regions of γ defined by $\gamma_1(a^*_{S,h})$ and $\gamma_2(a^*_{S,h})$. For $\gamma \leq \gamma_1(a^*_{S,h})$, R is unkind to S, however since γ is very low, S follows her material interests to truth-tell. For $\gamma \in \left(\gamma_1(a^*_{S,h}), \gamma_2(a^*_{S,h})\right)$, S is sufficiently concerned about reciprocating R's unkindness that she deviates to m_0 in state h, so as to give R a lower payoff and thus be unkind to R. For $\gamma \geq \gamma_2(a^*_{S,h})$, R is now kind to S. The latter has no incentive to deviate from truth-telling, mutual kindness is feasible and truth-telling involves gift-exchange.

Finally, in our main results we found that less bias (higher $a^*_{S,h}$) could prevent truth-telling (Proposition 2). With the belief-independent efficiency definition, more preference alignment only ever encourages the existence of positive TTE (Proposition 6). To see this, note that in part (a), the higher $a^*_{S,h}$, the lower the γ needed for existence of that type of TTE. In parts (b) and (c) it is straightforward to see that the existence conditions are more likely to hold the higher is $a^*_{S,h}$.

For the intuition behind the different effect of the bias, recall that with the belief-dependent efficiency definition, more preference alignment meant that actions/messages became inefficient, such that agents stopped perceiving each other as kind, thereby impeding truth-telling. However, with a belief-independent efficiency definition, such an issue does not arise and thus reciprocity incentives work in the same direction as material incentives with regards preference alignment.

3.3 One-sided reciprocity

In this section, we again revisit the model studied in our main analysis. Both S and R having reciprocity preferences seems reasonable for situations where the agents are to some extent peers (e.g. work colleagues) differing only in information and material interests. Often times however, cheap talk-like situations involve more vertical relationships (e.g. a firm that knows the quality of the product that it sells to an uninformed consumer; an employer who knows the potential gains from a project talking to an employee who does not). For such instances, while it has long been argued that R may have reciprocity concerns (e.g. efficiency wages (Akerlof 1982) or ethically minded consumers), it is in contrast less obvious that S (e.g. a firm) would have such concerns.

If S only cares about material payoffs and R has reciprocity preferences, the set of equilibria are qualitatively very similar to those where both players only care about material payoffs (see Observation 1).[12] In other words one-sided reciprocity generically does not facilitate truth-telling. This echoes results found in complete information settings on the impossibility of gift-exchange with one-sided reciprocity (Netzer and Schmutzler 2014).

Even if S is not reciprocal, she might not be maximising material payoffs in all situations. Continuing the firm example, S may care about corporate social responsibility. That is, while it does not want to reciprocate consumers' behaviour, it wants to be kind to consumers, R. We now examine TTE in such a context.

[12]The full result and its proof is found in the online appendix. Briefly, equilibria are identical to those in Observation 1, except that there also exists a set of knife-edge TTE where $\gamma = \frac{1}{a^*_{S,h}}$ and $a^*_{S,h} < \frac{h}{2}$.

Consider a model identical to our main analysis except now

$$U_S(m\,|\omega,\overline{\sigma}_{SR},\overline{\sigma}_{SRS}) = \underbrace{E\pi_S\,(m\,|\omega,\overline{\sigma}_{SR})}_{\text{material payoff}} + \underbrace{\gamma K_{SR}(m\,|\omega,\overline{\sigma}_{SR})}_{\text{kindness payoff}},$$

where for S, γ is interpreted as how much she cares about being kind.

Proposition 7 ***Positive TTE with one-sided reciprocity***

(a) *There exists a positive TTE featuring* $a(m_0) = 0$ *and* $a(m_1) = h$ *iff either*

(i) $a^*_{S,h} \geq \frac{h}{2}$, *or,*

(ii) $a^*_{S,h} \in \left[\frac{h}{2} - 1, \frac{h}{2}\right]$ *and* $\gamma \in \left[1 - \frac{2a^*_{S,h}}{h}, \frac{2}{h}\right]$.

(b) *There exists a positive TTE featuring* $a(m_0) = 0$ *and* $a(m_1) = \frac{2}{\gamma}$ *iff* $a^*_{S,h} < \frac{h}{2}$ *and* $\gamma \in \left[\frac{1}{a^*_{S,h}+1}, \frac{1}{a^*_{S,h}}\right]$.

(c) *There exist no other positive TTE.*

The figure below illustrates the existence conditions.

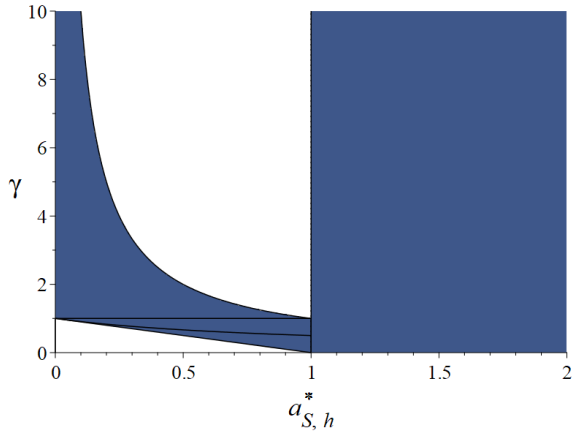


Figure 3: Existence of a positive TTE where R is reciprocal and S is kind.

Notes: For $h = 2$ the shaded areas depict values of $\{a^*_{S,h}, \gamma\}$ such that a positive TTE exists when R has reciprocity preferences and S has a desire to be kind.

Broadly speaking, truth-telling via a positive TTE is easier to achieve in this model than in our baseline model. This is because S does not deviate from truth-telling when R is unkind to S. For example, when $a^*_{S,h} \geq \frac{h}{2}$, R is unkind to S by choosing action h in state

h and thus in our main analysis S had an incentive to deviate from truth-telling. Clearly this mechanism is irrelevant when S does not care about reciprocity.

Despite this difference, it is striking that for $a^*_{S,h} < \frac{h}{2}$, some of the principal qualitative features of our main results hold even in this setting. First, note that a higher γ can be detrimental to truth-telling. The intuition is once again 'self-defeating kindness', where R is so kind in response to S's kindness that she prevents S from being kind to her, thus precluding gift-exchange via truth-telling. Second, note that more material preference misalignment can be beneficial to truth telling. As before, this is because an agent's ability to be kind to her co-agent is greater if material preferences are more misaligned, the potential size of gifts being larger, and material preference misalignment acting as a commitment device for R.

3.4 Comparisons with Rabin (1993) and Dufwenberg and Kirchsteiger (2004)

While we are the first to study the effect of reciprocity in cheap-talk environments, the effect of reciprocity has been studied in other games. To better understand the significance of our results in this section we examine whether analogues of our three main insights are found in other games where players have reciprocity preferences.

Our first main insight was that with reciprocity truth-telling can becomes feasible in a cheap-talk game even if material preference misalignment is high (e.g. Proposition 2). More generally, one can think of this as reciprocity implying the existence of more efficient equilibria underpinned by gift-exchange. While this insight is novel to the cheap-talk literature, the broader idea has been demonstrated in several different contexts. For instance, Dufwenberg and Kirchsteiger (2004) demonstrate how reciprocity can lead to cooperation in a sequential prisoner's dilemma.

Our second main insight was that higher reciprocity concerns can be detrimental to truth-telling (e.g. Proposition 1). More generally, one can think of this as greater reciprocity concerns preventing the existence of efficient equilibria. A typical finding in the literature is that higher payoff equilibria exist when reciprocity concerns are high. For instance, Rabin (1993) examines several normal form games under reciprocity and demon-

strates that as the relative importance of reciprocity incentives increases (i.e. X decreases in his notation), equilibria where players maximise each others' material payoffs eventually exist. Dufwenberg and Kirchsteiger (2004) demonstrate in a variety of classic games (e.g. the centipede game and the sequential prisoners' dilemma) that higher payoff equilibria exist only if reciprocity concerns are sufficiently important (i.e. sufficiently high Y in their notation). As compared to these papers, the insight that greater reciprocity concerns can preclude the existence of more efficient equilibria is thus novel.

Our final main insight was that greater material preference misalignment can facilitate the existence of TTE. How best to map interest alignment into other games is not always obvious. In Rabin (1993, p. 1293) studies a employment relationship model where γ is the cost the worker incurs if he exerts high effort and R is the revenue that the employer gets in this case. These parameters can be thought of as capturing material preference alignment. Rabin finds that a "positive fairness equilibrium" (i.e. one involving mutual kindness) is only possible if R and γ are sufficiently low, in other words if material preferences are sufficiently well aligned.

4 Conclusion

We studied whether reciprocity preferences affect interaction in cheap talk games. Our analysis suggests they indeed can, by generating instances where interaction takes the form of gift-exchange. Two counterintuitive results were observed in both discrete and continuous environments and with alternative preference assumptions. First, too great a concern for reciprocity can hurt information transmission. Second, a greater *mis*alignment of material preferences can improve information transmission.

Future work ought to consider important aspects left unaddressed. First, we did not examine noisy communication, in the form of either randomized or partitional communication strategies. Do our basic qualitative insights hold once we consider such equilibria? Second, do our counterintuitive comparative statics w.r.t. γ and incentive misalignment arise in consequence of incomplete information, or do they rather stem from other features of the game? Third, to the best of our knowledge, while there exists much ex-

perimental and empirical work studying gift-exchange with physical items or monetary transfers, there is no such work examining whether communication can form the basis of gift-exchange.

Finally, we alluded to the relevance of our one-sided reciprocity model for a socially responsible firm interacting with a reciprocal consumer. One could introduce this assumption into standard industrial organisation models and study how it affects competition and equilibrium outcomes.

Appendix: Proofs

4.1 Proof of Lemma 1

In this proof we consider R's incentives in a positive TTE. First examine R's best response after m_0. Note that conditional on m_0, $K_{RSR}(m_0, \sigma_S, \sigma_R) = 0$. To see this, recall first that

$$K_{RSR}(m, \sigma_S, \sigma_R) = \sum_{\omega} P(\omega \,|\sigma_S, m) K_{SR}(m \,|\omega, \sigma_R). \tag{1}$$

Now, note that $P(0 \,|\sigma_S, m_0) = 1$ in the considered truth-telling equilibrium. If $a(m_1) > a(m_0)$, message m_1 is conditionally inefficient given state 0; if $a(m_1) = a(m_0)$, then neither message is conditionally inefficient. In both cases, it holds true that

$$K_{SR}(m_0 \,|0, \sigma_R) = -a(m_0) - \frac{(-a(m_0) - a(m_0))}{2} = 0.$$

Substituting $P(0 \,|\sigma_S, m_0) = 1$ and $K_{SR}(m_0 \,|0, \sigma_R)$ into (1) gives $K_{RSR}(m_0, \sigma_S, \sigma_R) = 0$. This implies $U_R(a \,|m_0)$ reduces to R's material payoff function, which is strictly decreasing in a, so R's optimal action after m_0 is 0.

We now examine R's best response after m_1. Note that $K_{RSR}(m_1, \sigma_S, \sigma_R) \geq 0$. To see this, use (1) and first note that $P(h \,|\sigma_S, m_1) = 1$. Note second that if $\left|a(m_0) - a^*_{S,h}\right| \leq \left|a(m_1) - a^*_{S,h}\right|$ (meaning that no message is conditionally inefficient given $\omega = h$), then it holds true that

$$K_{SR}(m_1 \,|h, \sigma_R) = -(h - a(m_1)) - \frac{-(h - a(m_1)) - (h - a(m_0))}{2} \geq 0.$$

If instead $\left|a(m_0) - a^*_{S,h}\right| > \left|a(m_1) - a^*_{S,h}\right|$ (meaning that m_0 is conditionally inefficient given $\omega = h$), then it holds true that

$$K_{SR}(m_1 \,|h, \sigma_R\,) = -(h - a(m_1)) - \frac{-(h - a(m_1)) - (h - a(m_1))}{2} = 0.$$

Note also that conditional on m_1, any action strictly below $a^*_{S,h}$ is conditionally inefficient. The kindness of R conditional on m_1, denoted $K_{RS}(a\,|m_1, \sigma_S\,)$, is thus given by:

$$\begin{aligned} K_{RS}(a\,|m_1, \sigma_S\,) &= E\pi_S(a\,|m_1, \sigma_S\,) - \frac{\max\limits_{a \notin \Sigma^{A,-}(m_1, \sigma_S)} E\pi_S(a\,|m_1, \sigma_S\,) + \min\limits_{a \notin \Sigma^{A,-}(m_1, \sigma_S)} E\pi_S(a\,|m_1, \sigma_S\,)}{2} \\ &= -\left|a - a^*_{S,h}\right| - \left[\frac{-(h - a^*_{S,h}) - (a^*_{S,h} - a^*_{S,h})}{2}\right] = -\left|a - a^*_{S,h}\right| - \frac{-(h - a^*_{S,h})}{2}. \end{aligned}$$

Consequently, R's reciprocity payoff in $U_R(a\,|m_1\,)$, denoted

$$\gamma K_{RS}(a\,|m_1, \sigma_S\,) K_{RSR}(m_1, \sigma_S, \sigma_R),$$

is either 0 or strictly monotonically increasing in a for $a < a^*_{S,h}$. Given R's material payoff in $U_R(a\,|m_1\,)$ is trivially increasing in a, we may conclude that R's optimal action after m_1 is weakly larger than $a^*_{S,h}$. ■

4.2 Proof of Proposition 1

Assume throughout that $a^*_{S,h} \geq \frac{h}{2}$. By Lemma 1 we know that $a(m_0) = 0$ and $a(m_1) \geq a^*_{S,h}$. We shall thus solve the game backwards, first considering R's incentives at following m_1 and then S's incentives at each of her information sets.

Incentives of R: Consider R's incentives following m_1. The relationship between $\left|a(m_1) - a^*_{S,h}\right|$ and $\left|a^*_{S,h} - a(m_0)\right|$ determines whether any of S's messages are conditionally inefficient given $\omega = h$. Using $a(m_0) = 0$ and $a(m_1) \geq a^*_{S,h}$ (Lemma 1), we thus compare $a(m_1) - a^*_{S,h}$ and $a^*_{S,h}$. Suppose that $a(m_1) - a^*_{S,h} > a^*_{S,h}$. This simplifies to $\frac{a(m_1)}{2} > a^*_{S,h}$, however this contradicts our assumption that $a^*_{S,h} \geq \frac{h}{2}$ given that $a(m_1) \leq h$. Thus it must be that $a(m_1) - a^*_{S,h} \leq a^*_{S,h}$ (put differently, $\frac{a(m_1)}{2} \leq a^*_{S,h}$). In this case, m_0 is conditionally inefficient given $\omega = h$. Since m_0 is conditionally inefficient given state h, we have:

$$K_{RSR}(m_1) = -(h - a(m_1)) - \left(\frac{-\,(h - a(m_1)) - (h - a(m_1))}{2}\right) = 0.$$

It follows that $U_R(a\,|m_1\,) = -(h-a)$. Clearly, R's optimal response equals h. Note that $a(m_1) = h$ and $\frac{a(m_1)}{2} \leq a^*_{S,h}$ are consistent with our initial assumption $a^*_{S,h} \geq \frac{h}{2}$.

Incentives of S: First consider S's incentives following $\omega = h$, given $a(m_1) = h$.

We first identify how kind S perceives R as being. Following a message of m_0, the kindness of R to S is zero since all actions other than 0 are conditionally inefficient given the state $\omega = 0$. Following a message of m_1, the kindness of R to S is:

$$K_{RS}(a\,|m_1\,) = -\left(h - a^*_{S,h}\right) - \left(\frac{0-(h-a^*_{S,h})}{2}\right) = \frac{1}{2}a^*_{S,h} - \frac{1}{2}h < 0,$$

where the final inequality follows from $a^*_{S,h} < h$. Given the common prior, S's perception of R's kindness equals $(1-\alpha)\,\frac{1}{2}\left(a^*_{S,h} - h\right) < 0$.

Now consider S's kindness from sending each message at $\omega = h$. The kindness of sending m_0 is $-h - \frac{(-0-0)}{2} = -h$ and that of sending m_1 is $-0 - \frac{(-0-0)}{2} = 0$. Summarising, it follows that at $\omega = h$, S's utility from sending each message is given respectively by

$$U_S(m_1\,|h\,) = -(h - a^*_{S,h}) + \gamma\,(0)\,(1-\alpha)\,\frac{1}{2}\left(a^*_{S,h} - h\right)$$

and

$$U_S(m_0\,|h\,) = -a^*_{S,h} + \gamma\,(-h)\,(1-\alpha)\,\frac{1}{2}\left(a^*_{S,h} - h\right).$$

Note that $U_S(m_1\,|h\,) \geq U_S(m_0\,|h\,)$ is equivalent to

$$-(h - a^*_{S,h}) + \gamma\,(0)\,(1-\alpha)\,\frac{1}{2}\left(a^*_{S,h} - h\right) \geq -a^*_{S,h} + \gamma\,(-h)\,(1-\alpha)\,\frac{1}{2}\left(a^*_{S,h} - h\right)$$

which rewrites as $\gamma \leq \frac{2\left(2a^*_{S,h}-h\right)}{h(1-\alpha)\left(h-a^*_{S,h}\right)}$.

Finally, consider S's incentives at $\omega = 0$. Note that S's perception of R's kindness remains $(1-\alpha)\,\frac{1}{2}\left(a^*_{S,h} - h\right) < 0$. S's kindness to R if sending m_0 is $0 - \frac{(-0-0)}{2} = 0$, and that of sending m_1 is $-(h-0) + \frac{0+0}{2} = -h$. It follows that at $\omega = 0$, S's utility from sending each message is given by

$$U_S(m_0\,|0\,) = -(0-0) + \gamma\,(0)\,(1-\alpha)\,\frac{1}{2}\left(a^*_{S,h} - h\right)$$

and

$$U_S(m_1\,|0\,) = -h + \gamma\,(-h)\,(1-\alpha)\,\frac{1}{2}\left(a^*_{S,h} - h\right).$$

Note that $U_S(m_0\,|0\,) \geq U_S(m_1\,|0\,)$ is equivalent to

$$-(0-0)+\gamma\,(0)\,(1-\alpha)\,\frac{1}{2}\left(a^*_{S,h}-h\right) \geq -h+\gamma\,(-h)\,(1-\alpha)\,\frac{1}{2}\left(a^*_{S,h}-h\right)$$

which rewrites as $\frac{2}{(1-\alpha)\left(h-a^*_{S,h}\right)} \geq \gamma$.

Note that $\frac{2}{(1-\alpha)\left(h-a^*_{S,h}\right)} > \frac{2\left(2a^*_{S,h}-h\right)}{h(1-\alpha)\left(h-a^*_{S,h}\right)}$ simplifies to $h > a^*_{S,h}$, which is true, and thus $U_S(m_0\,|0\,) \geq U_S(m_1\,|0\,)$ is implied by $U_S(m_1\,|h\,) \geq U_S(m_0\,|h\,)$. ■

4.3 Proof of Proposition 2

Assume throughout that $a^*_{S,h} < \frac{h}{2}$ and that $a(m_0) = 0$ and $a(m_1) \geq a^*_{S,h}$ (by Lemma 1). Given these assumptions we shall solve the game backwards, first considering R's incentives at each of her information sets, and then S's incentives at each of her information sets.

Incentives of R: First consider R's incentives following m_0. Note that either message m_1 is conditionally inefficient given $\omega = 0$ (if $a(m_1) > 0$) or neither message is conditionally inefficient given $\omega = 0$ (if $a(m_1) = 0$). In either case, R's perception of S's kindness given m_0 is zero, i.e. $K_{RSR}(m_0) = 0$. It follows that $U_R(a\,|m_0\,) = -a$, so that R's best response to m_0 is trivially 0.

Now consider R's incentives following m_1. The relationship between $\left|a(m_1) - a^*_{S,h}\right|$ and $\left|a^*_{S,h} - a(m_0)\right|$ determines whether any of S's messages are conditionally inefficient given $\omega = h$. If $\frac{a(m_1)}{2} \geq a^*_{S,h}$ (call this *Case 1*), then $a(m_1) - a^*_{S,h} \geq a^*_{S,h}$, so $a(m_0)$ is closer to $a^*_{S,h}$ than $a(m_1)$. It follows that no message is conditionally inefficient given $\omega = h$. If instead $\frac{a(m_1)}{2} < a^*_{S,h}$ (call this *Case 2*), then m_0 is conditionally inefficient given $\omega = h$. Consider R's incentives after m_1 for each of the two cases we have identified.

Case 1: In this case, it holds true that $\frac{a(m_1)}{2} \geq a^*_{S,h}$. Given that no message is conditionally inefficient at $\omega = h$, R's perception of S's kindness given m_1 is:

$$K_{RSR}(m_1) = -(h - a(m_1)) - \left(\frac{-(h-a(m_1)) - (h-0)}{2}\right) = \frac{1}{2}a(m_1) > 0.$$

Thus S is perceived to be kind to R. After m_1, any action $a < a^*_{S,h}$ is conditionally inefficient. Given that $K_{RSR}\,(m_1) > 0$, it follows that the utility maximizing response of R to

m_1 satisfies $a \geq a^*_{S,h}$, since choosing $a < a^*_{S,h}$ reduces both R's material and reciprocity payoffs. The kindness of R towards S when picking some action $a \geq a^*_{S,h}$ after receiving m_1 is given by:

$$K_{RS}(a\,|m_1) = -\left(a - a^*_{S,h}\right) - \left(\frac{0 - (h - a^*_{S,h})}{2}\right) = \frac{1}{2}a^*_{S,h} - a + \frac{1}{2}h.$$

It follows that R's utility after m_1, subject to $a \geq a^*_{S,h}$, is given by:

$$U_R(a\,|m_1) = -(h-a) + \gamma\left(\frac{1}{2}a(m_1)\right)\left(\frac{1}{2}a^*_{S,h} - a + \frac{1}{2}h\right).$$

To identify the optimal action a^* satisfying $a^* \geq a^*_{S,h}$, we examine the following derivative:

$$\frac{\partial U_R(a\,|m_1)}{\partial a} = 1 - \frac{1}{2}a(m_1)\gamma.$$

There are now 3 subcases to consider:

- *Subcase 1.a)*: Here, $1 - \frac{1}{2}a(m_1)\gamma > 0$ in which case $a^* = h$. The implied candidate equilibrium features $a(m_1) = h$ and requires $\gamma < \frac{2}{h}$. Note that $a(m_1) = h$ is compatible with the assumption that $\frac{a(m_1)}{2} \geq a^*_{S,h}$. Thus this subcase may potentially correspond to a positive TTE.

- *Subcase 1.b)*: Here, $1 - \frac{1}{2}a(m_1)\gamma < 0$ in which case $a^* = a^*_{S,h}$. Here, the implied candidate equilibrium features $a(m_1) = a^*_{S,h}$. This however contradicts the assumption that $\frac{a(m_1)}{2} \geq a^*_{S,h}$. Thus this subcase cannot correspond to a positive TTE.

- *Subcase 1.c)*: Here, $1 - \frac{1}{2}a(m_1)\gamma = 0$, in which case any $a \geq a^*_{S,h}$ is optimal. Note that the stated equality is equivalent to $a(m_1) = \frac{2}{\gamma}$. Combining this equality with $\frac{a(m_1)}{2} \geq a^*_{S,h}$, the implied candidate equilibrium thus requires $\frac{1}{\gamma} \geq a^*_{S,h}$. Since the optimal action must be within the action set, we require $a(m_1) = \frac{2}{\gamma} \leq h$. This is in turn equivalent to $\frac{2}{h} \leq \gamma$. Note that the condition, as it turns out, is always satisfied for the γ permitted in the Proposition (i.e. $\gamma \geq \frac{2(3-2\alpha)}{2a^*_{S,h}+(1-\alpha)(h+a^*_{S,h})}$). To prove this, we show that $\gamma \geq \frac{2(3-2\alpha)}{2a^*_{S,h}+(1-\alpha)(h+a^*_{S,h})}$ implies $\frac{2}{h} \leq \gamma$. Consider the value of $\frac{2(3-2\alpha)}{2a^*_{S,h}+(1-\alpha)(h+a^*_{S,h})}$ at the maximum value of $a^*_{S,h}$ allowed $\left(\text{given by } \overline{a}^*_{S,h} = \frac{h}{3}\right)$, and show that $\frac{2(3-2\alpha)}{2a^*_{S,h}+(1-\alpha)(h+a^*_{S,h})} > \frac{2}{h}$ (call this *Property I*). Given that $\frac{2(3-2\alpha)}{2a^*_{S,h}+(1-\alpha)(h+a^*_{S,h})}$ is decreasing in $a^*_{S,h}$ until $\overline{a}^*_{S,h}$, proving this property is sufficient for our

purpose. Recall that at $a^*_{S,h} = \overline{a}^*_{S,h}$ it holds true that $\frac{2(3-2\alpha)}{2a^*_{S,h}+(1-\alpha)(h+a^*_{S,h})} = \frac{1}{a^*_{S,h}}$. We now prove Property I. Note that $\frac{1}{a^*_{S,h}}$, evaluated at $a^*_{S,h} = \frac{h}{3}$, equals $\frac{3}{h}$, which is trivially larger than $\frac{2}{h}$. Overall then, if $\frac{1}{\gamma} \geq a^*_{S,h}$, then $a(m_1) = \frac{2}{\gamma}$ may be part of a positive TTE.

Case 2: In this case, it holds true that $\frac{a(m_1)}{2} < a^*_{S,h}$. Here, m_0 is conditionally inefficient given state h. We thus have:

$$K_{RSR}(m_1) = -(h - a(m_1)) - \left(\frac{-(h - a(m_1)) - (h - a(m_1))}{2} \right) = 0.$$

It follows that $U_R(a\,|m_1\,) = -(h-a)$. Clearly, R's optimal response equals h. However if $a(m_1) = h$ and $\frac{a(m_1)}{2} < a^*_{S,h}$, then $\frac{h}{2} < a^*_{S,h}$ which contradicts the assumption at the start of the proof that $a^*_{S,h} < \frac{h}{2}$. Thus an equilibrium cannot involve $\frac{a(m_1)}{2} < a^*_{S,h}$.

Incentives of S: First consider S's incentives following $\omega = h$. Consider the two relevant cases inherited from our analysis of R's incentives, namely subcase 1.a) and subcase 1.c).

Subcase 1.a): Here, it holds true that $a(m_1) = h$ and $\gamma < \frac{2}{h}$. We first identify how kind S perceives R as being. Following a message of m_0, the kindness of R to S is zero since all actions other than 0 are conditionally inefficient given the state $\omega = 0$. Following a message of m_1, the kindness of R to S is:

$$K_{RS}(a\,|m_1\,) = -\left(h - a^*_{S,h}\right) - \left(\frac{0 - (h - a^*_{S,h})}{2} \right) = \frac{1}{2}a^*_{S,h} - \frac{1}{2}h < 0,$$

where the final inequality follows from $a^*_{S,h} < h$. Given the common prior, S's perception of R's kindness equals $(1-\alpha)\,\frac{1}{2}\left(a^*_{S,h} - h\right) < 0$.

Now consider S's kindness from sending each message at $\omega = h$. The kindness of sending m_0 is $-h - \frac{(-h-0)}{2} = -\frac{h}{2}$ and that of sending m_1 is $-0 - \frac{(-h-0)}{2} = \frac{h}{2}$. Summarizing, it follows that at $\omega = h$, S's utility from sending each message is given respectively by

$$U_S(m_1\,|h\,) = -(h - a^*_{S,h}) + \gamma \left(\frac{h}{2}\right)(1-\alpha)\,\frac{1}{2}\left(a^*_{S,h} - h\right)$$

and

$$U_S(m_0\,|h\,) = -a^*_{S,h} + \gamma \left(-\frac{h}{2}\right)(1-\alpha)\,\frac{1}{2}\left(a^*_{S,h} - h\right).$$

Note that m_0 gives S both higher material and reciprocity payoffs than m_1, so that S deviates from truth-telling and sends m_0 at $\omega = h$. Hence $a(m_1) = h$ cannot be a part of a positive TTE.

Subcase 1.c): Here, it holds true that $a(m_1) = \frac{2}{\gamma}$ and $\frac{1}{\gamma} \geq a^*_{S,h}$. To determine whether S deviates, we first identify how kind S perceives R as being. The kindness of R given m_1 is given by

$$-\left(a(m_1) - a^*_{S,h}\right) - \frac{0 - (h - a^*_{S,h})}{2} = \frac{1}{2}a^*_{S,h} - a(m_1) + \frac{1}{2}h.$$

To see this, note that any action outside $[a^*_{S,h}, h]$ is conditionally inefficient given m_1. The kindness of R given m_0 is on the other hand 0. To see this, note that conditional on m_0, any action other than 0 is conditionally inefficient. It follows that S's perception of R's kindness is $(1-\alpha)\left(\frac{1}{2}a^*_{S,h} - a(m_1) + \frac{1}{2}h\right)$.

The kindness of S if sending m_1 is

$$K_{SR}(m_1 \,|h) = -(h - a(m_1)) - \left(\frac{-(h - a(m_1)) - (h - 0)}{2}\right) = \frac{1}{2}a(m_1).$$

We thus have

$$U_S\left(m_1 \,|h\right) = -\left(a(m_1) - a^*_{S,h}\right) + \gamma(1-\alpha)\left(\frac{1}{2}a(m_1)\right)\left(\frac{1}{2}h - a(m_1) + \frac{1}{2}a^*_{S,h}\right).$$

The kindness of S if sending m_0 is $-\frac{1}{2}a(m_1)$. Recall indeed that no message is conditionally inefficient at h, so that

$$K_{SR}(m_0 \,|h) = -(h - 0) - \left(\frac{-(h - 0) - (h - a(m_1))}{2}\right) = -\frac{1}{2}a(m_1).$$

We thus have

$$U_S\left(m_0 \,|h\right) = -a^*_{S,h} - \gamma\left(\frac{1}{2}a(m_1)\right)(1-\alpha)\left(\frac{1}{2}a^*_{S,h} - a(m_1) + \frac{1}{2}h\right).$$

Now simply examine $U_S\left(m_1 \,|h\right) - U_S\left(m_0 \,|h\right)$ as a function of underlying parameters:

$$\begin{aligned} & U_S\left(m_1 \,|h\right) - U_S\left(m_0 \,|h\right) \\ = \; & -\left(\frac{2}{\gamma} - a^*_{S,h}\right) + \gamma(1-\alpha)\left(\left(\frac{1}{2}\frac{2}{\gamma}\right)\left(\frac{1}{2}h - \frac{2}{\gamma} + \frac{1}{2}a^*_{S,h}\right)\right) \\ & + a^*_{S,h} + \gamma\left(\frac{1}{2}\frac{2}{\gamma}\right)(1-\alpha)\left(\frac{1}{2}a^*_{S,h} - \frac{2}{\gamma} + \frac{1}{2}h\right). \end{aligned}$$

Note that $U_S(m_1|h) - U_S(m_0|h) \geq 0$ is equivalent to:

$$\gamma \geq \frac{2(3-2\alpha)}{2a^*_{S,h} + (1-\alpha)(h + a^*_{S,h})}.$$

Thus S does not deviate from truth-telling at $\omega = h$ if the above inequality holds.

Note that the above gives a lower bound on γ for incentive compatibility of S. Recall that for subcase 1.c) we identified an upper bound on γ implied by R's incentive compatibility conditions. The condition read $\frac{1}{\gamma} \geq a^*_{S,h} \Rightarrow \frac{1}{a^*_{S,h}} \geq \gamma$. To find the cut-off value the two conditions obtained for γ are simultaneously satisfied, simply solve

$$\frac{2(3-2\alpha)}{2a^*_{S,h} + (1-\alpha)(h + a^*_{S,h})} = \frac{1}{a^*_{S,h}}$$

for $a^*_{S,h}$, this yields $a^*_{S,h} = \frac{h}{3}$. Thus the candidate positive TTE requires $a^*_{S,h} \in [0, \frac{h}{3}]$.

Finally, consider S's incentives at $\omega = 0$. Note that in any putative equilibrium within the class that we consider here, S perceives R as kind and $a(m_0) = 0$. Note furthermore that the kindness of S is 0 if sending m_0 and negative if sending m_1. To see this, note that m_1 is conditionally inefficient given $\omega = 0$. It trivially follows that m_0 gives S both a higher material and a higher reciprocity payoff than m_1. Thus S does not deviate from truth-telling at $\omega = 0$. ■

4.4 Proof of Corollary 1

For $\omega = 0$, it is trivial to see that agents' material payoffs are zero in a positive TTE, while they are strictly negative in a negative TTE.

For $\omega = h$, reason as follows to see that the sum of agents' material payoffs is no higher in a negative TTE than in a positive TTE. Let $W^+(\omega)$ denote the sum of agents' material payoffs in state ω in a positive TTE and $W^-(\omega)$ denote that in a negative TTE. Note that $W^-(h) \in \{-\left(h + a^*_{S,h}\right), -\left(h + a^*_{S,h} - 2a(m_1)\right), -\left(h - a^*_{S,h}\right), -\left(a^*_{S,h} + \frac{4}{\gamma} - h\right)\}$ and $W^+(h) = -\left(h - a^*_{S,h}\right)$. It is trivial that $W^+(h) \geq \max\left\{-\left(h + a^*_{S,h}\right), -\left(h - a^*_{S,h}\right)\right\}$. To see that $W^+(h) \geq -\left(a^*_{S,h} + \frac{4}{\gamma} - h\right)$, note that this value of W^- refers to the TTE in Proposition 3(c) where $\gamma < \frac{2}{h - a^*_{S,h}}$, thus $\gamma > \min\left\{\frac{2}{h}, \frac{1}{h - a^*_{S,h}}\right\}$ hence $-\left(a^*_{S,h} + \frac{4}{\gamma} - h\right) < \max\{3(h -$

$a^*_{S,h}), h + a^*_{S,h}\}$. To see that $W^+(h) \geq -\left(h + a^*_{S,h} - 2a(m_1)\right)$, note that this value of W^- refers to the TTE in Proposition 3(b) for cases where $a(m_1) < a^*_{S,h}$, hence $-\left(h + a^*_{S,h} - 2a(m_1)\right) < -\left(h - a^*_{S,h}\right)$. ■

References

[1] Abeler, J., Nosenzo, D. and C. Raymond (2018) "Preferences for truth telling" *Econometrica* (forthcoming).

[2] Akerlof, G. A. (1982) "Labor contracts as partial gift exchange" *Quarterly Journal of Economics* 97(4): 543-569.

[3] Attanasi, G., Battigalli, P. and E. Manzoni (2016) "Incomplete information models of guilt aversion in the trust game" *Management Science* 62(3): 648-667.

[4] Bartling, B. and N. Netzer (2016) "An externality-robust auction: Theory and experimental evidence" *Games and Economic Behavior* 97: 186-204.

[5] Battigalli, P. and M. Dufwenberg (2009) "Dynamic psychological games" *Journal of Economic Theory* 144: 1-35.

[6] Battigalli, P., Charness, G. and M. Dufwenberg (2013) "Deception: The role of guilt" *Journal of Economic Behavior and Organization* 93: 227-232.

[7] Battigalli, P. Corrao, R. and M. Dufwenberg (2018) "Incorporating belief-dependent motivation in games" mimeo.

[8] Bierbrauer, F. and N. Netzer (2016) "Mechanism design and intentions" *Journal of Economic Theory* 163: 557-603.

[9] Bierbrauer, F., Ockenfels, A., Pollak, A. and D. Ruckert (2017) "Robust mechanism design and social preferences" *Journal of Public Economics* 149: 59-80.

[10] Cai, H. and J. Wang (2006) "Overcommunication in strategic information transmission games" *Games and Economic Behavior* 95: 384-394.

[11] Charness, G., Frechette, G.R. and J.H. Kagel (2004) "How robust is laboratory gift exchange?" *Experimental Economics*, 7(2): 189-205.

[12] Crawford, V. and J. Sobel (1982) "Strategic information transmission" *Econometrica* 50(6): 1431-1451.

[13] De Marco, G. and G. Immordino (2014) "Reciprocity in the principal-multiple agent model" *B.E. Journal of Theoretical Economics* 14(1): 445-482.

[14] Dufwenberg, M. and M. Dufwenberg (2018) "Lies in disguise: A theoretical analysis of cheating" *Journal of Economic Theory* 175: 248-264.

[15] Dufwenberg, M. and G. Kirchsteiger (2004) "A theory of sequential reciprocity" *Games and Economic Behavior* 47(2): 268-298.

[16] Falk, A. (2007) "Gift exchange in the field" *Econometrica* 5: 1051-11.

[17] Falk, A. and U. Fischbacher (2006) "A theory of reciprocity" *Games and Economic behaviour* 54(2): 293-315.

[18] Farrell, J. and M. Rabin (1996) "Cheap talk" *Journal of Economic perspectives* 10(3): 103-118.

[19] Fehr, E., Kirchsteiger, G. and A. Riedl (1993) "Does fairness prevent market clearing? An experimental investigation" *The Quarterly Journal of Economics* 108(2): 437-459.

[20] Fischbacher, U. and F. Föllmi-Heusi (2013) "Lies in Disguise - An experimental study on cheating" *Journal of the European Economic Association* 11: 525-547.

[21] Geanakoplos, J., Pearce, D. and E. Stacchetti (1989) "Psychological games and sequential rationality" *Games and Economic Behavior* 1: 60-79.

[22] Gneezy, U. (2005) "Deception: The role of consequences," *American Economic Review* 95(1): 384-394.

[23] Gneezy, U., Kajackaite, A. and J. Sobel (2018) "Lying aversion and the size of the lie" *American Economic Review* 108(2): 419-453.

[24] Kartik, N. (2009) "Strategic communication with lying costs" *Review of Economic Studies* 76: 1359-1395.

[25] Khalmetski, K. and D. Sliwka (2017) "Disguising lies - Image concerns and partial lying in cheating games" mimeo.

[26] Krishna, V. and J. Morgan (2008) "Cheap talk" *The New Palgrave Dictionary of Economics* 1: 751-756.

[27] Netzer, N, and A. Schmutzler (2014) "Explaining gift-exchange – The limits of good intentions" *Journal of the European Economic Association* 12(6): 1586-1616.

[28] Rabin, M. (1993) "Incorporating fairness into game theory and economics" *American Economic Review* 83(5): 1281-302.

[29] Sanchez-Pages, S. and M. Vorsatz (2007) "An experimental study of truth-telling in a sender-receiver game" *Games and Economic Behavior* 61: 86-112.

[30] Sobel, J. (2013) "Giving and receiving advice" *Advances in economics and econometrics* 1: 305-341.

[31] von Siemens, F. A. (2013) "Intention-based reciprocity and the hidden costs of control" *Journal of Economic Behavior and Organization* 92: 55-65.