Clemens König:

Net-Loss Reciprocation and the Context Dependency of Economic Choices

Munich Discussion Paper No. 2013-18

Department of Economics
University of Munich

Volkswirtschaftliche Fakultät
Ludwig-Maximilians-Universität München

NET-LOSS RECIPROCATION AND THE CONTEXT DEPENDENCY OF ECONOMIC CHOICES[*]

Clemens König[°]

17 November 2013

Abstract This paper proposes a novel explanation for the context dependency of individual choices in two-player games. Context dependency refers to the well-established phenomenon that a player, when choosing from a given opportunity set created by the other player's strategy, chooses differently in different situations because of different alternatives to the other player's strategy. The utility model used to explain this kind of context dependency incorporates a preference for net-loss reciprocation. Net-loss reciprocation means that a player's willingness to impose a net loss (i.e., loss minus gain) on the other player increases in the net loss that he or she derives from the other player's strategy. I show that net-loss reciprocation together with the method for calculating net losses developed in this paper explains the context dependencies in individual behaviour that have been documented in a number of experimental studies, whereas existing models of intention-based reciprocity fail to explain all the evidence.

Keywords: Reciprocity, Fairness, Experimental economics, Game theory, Loss aversion.

JEL: C70, C91, D63, D64.

---

[*] I would like to thank Klaus Schmidt, Martin Kocher, Johannes Maier, Fabian Herweg, Piers Trepper and Linda Gratz for helpful comments. Financial support from the Deutsche Forschungsgemeinschaft through SFB-TR 15 is gratefully acknowledged.

[°] Clemens König, Department of Economics, University of Munich, Ludwigstrasse 28, D-80539 Munich, Germany, Email: clemens.koenig@lrz.uni-muenchen.de

1 Introduction

This paper studies the *context dependency* of individual choices in two-player games. Consider the following basic decision problem faced by each player: Choosing an outcome of the game from the opportunity set created by the other player's strategy. Context dependency holds if a player's choice from a fixed opportunity set of this kind is not constant across situations, but varies with the alternative strategies at the other's disposal. Examples are provided below. They show that context dependency may occur both in the context of negative reciprocity (punishment for unkind behaviour) and positive reciprocity (reward for kind behaviour). While experimental evidence suggests an important role for context dependency in negative reciprocity, experimental studies of positive reciprocity have not established strong context dependency.[1] In this paper, I propose a new behavioural theory called *net-loss reciprocation* that can explain these apparently contradictory findings.

A new theory is called for because existing theories struggle to explain context dependency. According to outcome-based theories of behaviour, players are motivated by a single preference ordering on outcomes, which directly rules out any context dependency.[2] Existing theories that are not outcome-based also fail to explain all the evidence.[3] From an economic point of view, context dependency is important because it impacts the possibility of reaching (materially) efficient outcomes in strategic interactions. Context dependency strongly influences the occurrence of punishment, which in itself harms efficiency, but may improve the overall efficiency of the interaction.[4] Positive reciprocity, which promotes efficient behaviour like trust

[1] See Brandts and Solá (2001) and Falk et al (2003) for evidence on negative reciprocity and Dufwenberg and Gneezy (2000), Charness and Rabin (2002), McCabe et al (2003), Cox (2004), Servatka and Vadovic (2009) and Cox et al (2010) for evidence on positive reciprocity.

[2] Outcome-based theories leave open the possibility that several outcomes are most preferred in a given opportunity set. This can rationalise isolated instances of context dependency, but can hardly be regarded as a systematic explanation for their prevalence. Outcome-based theories include Fehr and Schmidt (1999) and Charness and Rabin (2002).

[3] Below, I focus on the intention-based models of reciprocity by Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006).

[4] The efficiency-promoting role of punishment has been extensively studied in the context of public good games (Fehr and Gächter, 2000). The evidence on whether punishment opportunities promote overall material efficiency in these games is mixed.

in trading relationships, appears less affected by context dependency. Yet, it is interesting to explore the mechanisms behind this, which also sheds light on robustness.

Net-loss reciprocation builds on two key concepts: Firstly, the loss and gain that players derive from the other player's strategy. Secondly, the loss and gain that players themselves impose on the other through their own choice from the feasible set given the other's strategy. Net-loss reciprocation means that players' willingness to pay for increasing the net loss imposed on the other increases in the net loss derived from the other's strategy.[5] Net losses are simply losses minus gains, where the two need not count for the same. Net-loss reciprocation is consistent with an intuitive notion of reciprocity, according to which people's kindness to others increases in these others' kindness to them. More importantly, net-loss reciprocation can explain context dependency because players may derive different net losses from two strategies creating the same opportunity set because of different alternatives at the other's disposal. As a result, their preferences on the same opportunity set may differ.

For an illustration of context dependency in the domain of *negative reciprocity*, consider the ultimatum mini games studied by Falk et al (2003). In all of them, the proposer can make a fixed offer of dividing the surplus, namely, "8 for the proposer, 2 for the responder", and one alternative offer that varies. The responder can accept or reject any offer. Consider two such games: One where the alternative is "5 for both" and one where it is "2 for the proposer, 8 for the responder". Since the opportunity sets after the fixed offer are the same, we have context dependency if responders are more likely to reject the fixed offer in one game than the other. Falk et al (2003) report significantly more rejection in the first game.

For an illustration in the context of *positive reciprocity*, consider the trust games studied by Dufwenberg and Gneezy (2000). In all of them, the second mover can share 20 between himself and the first mover in the event of trust, while the games differ regarding the outcome in case of no trust. Consider two games: One where the no-trust outcome is "4 for the first mover, 0 for the second mover" and one where it is "16 for the first mover, 0 for the second mover". Since sharing opportunities are identical, we have context dependency if second movers share the 20 differentially in the two games. Despite the fact that the no-trust outcomes differ considerably, Dufwenberg and Gneezy (2000) report no significant difference in sharing.

---

[5] This willingness to pay may be negative or positive.

Below, I put forward a utility model incorporating net-loss reciprocation to explain these and other experimental findings. I now sketch the basic features of the calculus of losses and gains underlying the model. I focus on the loss and gain that a player derives from the other player's strategy. Analogous procedures are used for determining the loss and gain imposed on the other.

A player's *loss* from some strategy of the other player (called henceforth the "status quo") is composed of his material loss and his fairness loss. The player derives a *material loss* whenever there is an alternative strategy of the other under which the player could have earned more than what he can maximally earn under the status quo. Moreover, the player derives a *fairness loss*, which he adds to his material loss, if his forgone earnings (that make up his material loss) derive from outcomes that are fairer than the fairness of the status quo. The intuition is that the player in this case feels an entitlement ("fairness claim") to his forgone earnings, which causes him to feel an additional loss. Since his fairness loss is added to his material loss, his total loss then exceeds his material loss. Fairness is measured by a function that ranks the outcomes of the game according to their fairness and incorporates considerations of material efficiency and a concern for the less well-off player.

For an illustration, consider the two ultimatum mini games discussed above. If the alternative to the status quo offer "8 for the proposer, 2 for the responder" is "5 for both", responders derive both a material loss (of 3) and a fairness loss, while their loss is limited to their material loss (of 6) if the alternative is "2 for the proposer, 8 for the responder". Intuitively, "5 for both" is fairer than the status quo for containing the same payoff sum, but a lower minimal payoff, whereas "2 for the proposer, 8 for the responder" is only as fair. Moreover, I show below that the sum of the material and fairness loss in the first game can exceed the higher material loss in the second game. This (together with the fact that there is no gain from the status quo) implies a higher net loss from the status quo in the first game, which explains responders' higher willingness to impose a net loss on proposers by rejecting the status quo more often.

Likewise, a player's *gain* from the status quo is composed of his material gain and his fairness gain. The player derives a *material gain* whenever he can earn more under the status quo than what he can maximally earn under some alternative. Moreover, the player derives a *fairness gain*, which is added to his material gain, if the earnings that make up his material gain derive from outcomes that are *less* fair than the fairness of the alternative. The intuition is that in the converse scenario, where the outcomes creating his material gain are fairer than the alternative,

the player feels an entitlement to his material gain ("well deserved"), which causes him to feel no gain based on fairness considerations.

For an illustration, consider the two trust games introduced before. Recall that the games differ regarding the no-trust outcome: No trust entails "4 for the first, 0 for the second mover" in the first game and "16 for the first, 0 for the second mover" in the second game. In each case, second movers suffer no loss from being trusted, and their material gain is the same (namely, 20) because they can earn up to 20 after trust, but earn zero after no trust. Moreover, their fairness gain is zero in both cases because there is no outcome after trust that is less fair than either no trust outcome, which means that second movers feel entitled to their additional earning possibilities after trust. Even the outcome "0 for the first, 20 for the second mover", which second movers can implement after trust, is not less fair than either no-trust outcome because it contains a larger total payoff and the same minimal payoff. As a result, second movers derive the same (negative) net loss from trust in each case, which explains why Dufwenberg and Gneezy (2000) detect no context dependency across the two situations.[6]

In contrast, existing models of social preferences struggle to account for context dependencies. Outcome-based models like the inequality aversion model of Fehr and Schmidt (1999) or the model of Charness and Rabin (2002), which combines a taste for material efficiency with generosity towards those who are least well-off, build on the idea that strategic behaviour derives from a single ranking of the outcomes of the interaction. For this reason, these models cannot provide a systematic explanation of context dependency.[7]

Such an explanation can in principle be provided by intention-based models of reciprocity making use of psychological game theory.[8] Most widely used are Dufwenberg and Kirchsteiger (2004), who build on Rabin (1993), and Falk and Fischbacher (2006). One key feature of these models is their reliance on players' second-order beliefs, i.e., their beliefs about the other player's belief about their own choice of strategy. Thus, when faced with some strategy of the other player, players consult their second-order belief, which together with the strategy of the

---

[6] Of course, an insignificant treatment effect can have other reasons such as too few observations.

[7] Context dependency can only arise from several most preferred outcomes in the fixed opportunity set.

[8] Psychological games were first defined and analysed by Geanakoplos, Pearce and Stacchetti (1989). A framework for dynamic psychological games is provided by Battigalli and Dufwenberg (2009).

other pins down a unique outcome of the game, which serves to represent the other's strategy. In contrast, my approach does not rely on second-order beliefs, but represents the other's strategy and its alternatives by the entire sets of feasible outcomes that these strategies create. Consequently, my approach is more amenable to empirical testing using standard experimental data as it does not require measurement of higher-order beliefs.[9]

Regarding predictions, both models assert that players' willingness to be kind to the other increases in the kindness of the other's status quo strategy, which is similar in spirit to net-loss reciprocation. Problems arise in the conceptualisation of kindness. For instance, in Dufwenberg and Kirchsteiger (2004), if the outcome representing the status quo gives the player more (less) than half of what he maximally and minimally stands to earn under the alternatives, the status quo is perceived as kind (unkind). Thus, this approach limits itself to comparing earnings without allowing for players' sense of entitlement to these earnings. The evidence from the ultimatum mini games makes clear that such a sense of entitlement may override material considerations.

Also, reliance on second-order beliefs may lead to unintuitive predictions because it uses players' (likely) reaction to the strategy of the other as a means to assess the strategy's kindness. Yet, if players react to some strategy that they in fact perceive as unkind in a self-serving manner and more generously to some alternative they perceive as kind, they may end up with more own payoff when faced with the former, from which Dufwenberg and Kirchsteiger (2004) would conclude that the first strategy is kinder. This problem is also shared by Falk and Fischbacher (2006). When discussing applications, I explain in more detail this and other difficulties encountered by the two models, whose basic building blocks are laid out in more detail in Appendix B. All in all, net-loss reciprocation can account for larger parts of the evidence than either model of intention-based reciprocity.

At first blush, the model of net-loss reciprocation introduced in this paper could be thought of as a model of loss aversion (Kahnemann and Tversky, 1979; Shalev 2000; Köszegi and Rabin, 2006). While I allow for the possibility that "losses loom larger than gains", which is the cornerstone of this literature, there are important differences. Loss aversion builds on the idea that the consequences of decisions are evaluated relative to some (deterministic or stochastic)

[9] Dhaene and Bouckaert (2010) investigate the performance of the model of Dufwenberg and Kirchsteiger (2004) using measured second-order beliefs in a setting unrelated to context dependency.

reference point. If the utility of a consequence exceeds (falls short of) the reference point, individuals perceive a gain (loss). Loss aversion therefore presupposes some baseline utility attached to consequences from which losses and gains can be calculated. This is where net-loss reciprocation steps in, which is best described as a theory about how sensations of loss and gain derived from the other's behaviour act as a source of (social) preferences and hence as a source of baseline utility attached to the different outcomes of the game.[10]

That said, the utility model proposed below is qualitative in the sense that the details of the utility function up to the net-loss reciprocation property are left open. There is also no equilibrium analysis, which is refrained from because it is not needed to explain the phenomena this paper sets out to explain.[11] Indeed, the context dependencies addressed below relate to the behaviour of players who have certainty about the other player's strategy because they are second movers in sequential games where each player has one move. As a result, these players' beliefs are pinned down by the game's information structure, and a notion of best response is enough to explain their choices.[12] Regarding best responses, I take no stance on which specification of utility consistent with net-loss reciprocation is most appropriate. There are several plausible ways of incorporating net-loss reciprocation into a full-fledged utility model.[13] Comparing the relative performance of these modelling options is left for future work. The substantive question addressed in this paper is: Can the net-loss reciprocation property of preferences together with my method for calculating losses and gains explain the context dependencies (or absences thereof) observed in experiments? The answer is largely affirmative.

---

[10] In this sense, loss aversion is orthogonal to net-loss reciprocation. Shalev (2000) studies loss aversion in games.

[11] Cox et al (2008) propose a non-equilibrium model of reciprocity in sequential games. However, it is not suited to studying context dependency.

[12] Of course, net-loss reciprocation could also be used to explain the behaviour of first movers with the added complication that their beliefs about the other's strategy are unobservable. These beliefs could be measured experimentally. However, there are to the best of my knowledge no economic experiments documenting context dependencies in the behaviour of players who must form beliefs about others.

[13] E.g., players could be willing to sacrifice own payoff to match the net loss they impose on the other to the net loss imposed on them. Players could also be endowed with some baseline preferences on outcomes whose degree of altruism decreases in the net loss they derive from the other (although this specification does not perfectly fit the definition of net-loss reciprocation given below).

The remainder of this paper is structured as follows: I first show how to calculate the net loss that a player derives from the other player's strategy, which is followed by the method for calculating the net loss that the player himself imposes on the other as well as the utility model incorporating net-loss reciprocation. I then show how this qualitative model can account for context dependency in a number of well-known experimental studies and compare its performance to intention-based models of reciprocity. All proofs are in Appendix A. Appendix C contains a generalisation of my approach to $n \geq 3$ players including Nature.

2 Losses and Gains From the Other Player's Strategy

I limit attention to finite-horizon two-player (*i*, *j*) multi-stage games with observable past actions.[14] A player's inactivity at a stage is modelled by the respective action set being singleton. Let *H* be the set of non-terminal histories of the game. Player *i*'s *pure strategy* $s_i \in S_i$ assigns to each history $h \in H$ an action available to *i* at *h*.[15] I restrict attention to pure strategies. The set of pure strategy profiles is $S = S_i \times S_j$. Outcomes $\pi = (\pi_i, \pi_j)$ of the game are two-dimensional vectors of material payoffs. The function $\pi : S \to \mathbb{R}^2$ is the *outcome function* and $\Pi = \{\pi(s) : s \in S\}$ the set of attainable outcomes in the game. Moreover, the set of attainable outcomes or *opportunity set* for player *i* given that player *j* plays strategy $s_j \in S_j$ is given by $\Pi^{s_j} = \{\pi(s_i, s_j) : s_i \in S_i\}$ with $\Pi^{s_j} \subseteq \Pi$.

I first define player *i*'s *loss* from $s_j \in S_j$. Player *i*'s overall loss is derived from a more basic notion, namely, his loss from s_j relative to a particular alternative $\tilde{s}_j \in S_j$. I also refer to s_j as the *status quo* and to $\tilde{s}_j$ as the *alternative*. The basic idea is the following: Player *i* suffers a loss from s_j relative to $\tilde{s}_j$ only if he can earn more given $\tilde{s}_j$ than what he can maximally earn given s_j. Or, more formally and with a slight abuse of notation, let $\Pi^{\tilde{s}_j > s_j}$ be the set of outcomes in $\Pi^{\tilde{s}_j}$ that yield *i* a higher payoff than his highest attainable payoff in Π^{s_j}. A necessary condition for *i* suffering a loss is then that $\Pi^{\tilde{s}_j > s_j}$ is non-empty. The magnitude of his loss is determined by considering the different outcomes in $\Pi^{\tilde{s}_j > s_j}$. For each $\pi \in \Pi^{\tilde{s}_j > s_j}$, *i*

[14] I refer to *i* as "he" and *j* as "she".

[15] Action sets are assumed to be finite.

calculates both his *material loss*, which is the amount by which his payoff from π exceeds his maximal payoff in Π^{s_j}, and his *fairness loss*, which tracks the extent to which π is fairer than the fairest outcomes in Π^{s_j}.

While *i*'s material loss is guaranteed to be positive by the definition of $\Pi^{\tilde{s}_j \succ s_j}$, π may or may not be fairer than the fairest outcomes in Π^{s_j}. If π is not fairer, *i* perceives no fairness loss because he then has no fairness claim to π even though he could have earned more from it than what he can maximally earn given s_j. His loss is then limited to his material loss. In contrast, if π is fairer than the fairest outcomes in Π^{s_j}, *i* suffers a fairness loss because he now has a fairness claim to his additional earnings from π. All in all, *i*'s loss from s_j relative to a particular $\pi \in \Pi^{\tilde{s}_j \succ s_j}$ is the weighted sum of his material and fairness loss, while his loss from s_j relative to $\tilde{s}_j$ at large is his maximal loss from s_j relative to the outcomes in $\Pi^{\tilde{s}_j \succ s_j}$.

I now put more formal structure on these ideas. As stated above, Π^{s_j} is the set of attainable outcomes given s_j. The set $\Pi_i^{s_j}$ is the set of payoffs to player *i* contained in Π^{s_j}. Its maximal element is $\overline{\pi}_i^{s_j} = max\,\Pi_i^{s_j}$. Furthermore, $\Pi^{\tilde{s}_j \succ s_j} = \left\{\pi \in \Pi^{\tilde{s}_j} : \pi_i > \overline{\pi}_i^{s_j}\right\}$ is the set of attainable outcomes given $\tilde{s}_j$ that yield *i* more payoff than what he can maximally earn given s_j. Fairness is measured by a fairness function, isoquants of which are called *fairness curves*:

DEFINITION 1 The *fairness function* $f : \mathbb{R}^2 \to \mathbb{R}$ is given by

$$f(\pi) = \alpha\left(\pi_i + \pi_j\right)/2 + (1-\alpha)\underline{\pi}$$

where $\underline{\pi} = min\left\{\pi_i, \pi_j\right\}$ and $\alpha \in [0,1]$.

For an interpretation of the parameter α, consider the polar cases $\alpha = 1$ and $\alpha = 0$: If $\alpha = 1$, fairness boils down to material efficiency meaning that for any two outcomes π and $\tilde{\pi}$ we have $f(\pi) > f(\tilde{\pi})$ if and only if $\pi_i + \pi_j > \tilde{\pi}_i + \tilde{\pi}_j$. In contrast, if $\alpha = 0$, we have $f(\pi) > f(\tilde{\pi})$ if and only if $\underline{\pi} > \underline{\tilde{\pi}}$. For fairness to increase in this case, the player with less payoff must receive more.[16] In general, the lower α, the smaller (larger) the relative weight attached to efficiency (equality) considerations in fairness assessments. Yet, it is not equality *per se* that enters the

[16] The second case is reminiscent of a Rawlsian (or max-min) social welfare function. The first case has a utilitarian flavour.

fairness function, but the payoff of the less well-off player. It is this payoff that must increase for fairness to increase. If equality as such mattered, we could also reduce the payoff of the better-off player for fairness to increase.[17]

Player *i* derives a fairness loss from s_j relative to $\pi \in \Pi^{\tilde{s}_j > s_j}$ if and only if π lies on a higher fairness curve than the highest fairness curve reached in Π^{s_j}. Moreover, *i*'s fairness loss increases in the extent to which the fairness of π exceeds the maximal fairness in Π^{s_j}. To formalise this idea, let $\overline{f}^{s_j} = max_{\pi \in \Pi^{s_j}} f(\pi)$ be the highest fairness level attained in Π^{s_j}. The fairness gap between π and Π^{s_j} can then be expressed as $f(\pi) - \overline{f}^{s_j}$.[18] This lead to

DEFINITION 2 Player *i*'s *loss* from strategy $s_j \in S_j$ relative to strategy $\tilde{s}_j \in S_j$ is given by

$$l_i(s_j, \tilde{s}_j) = \begin{cases} max_{\pi \in \Pi^{\tilde{s}_j > s_j}} \left[\beta\left(\pi_i - \overline{\pi}_i^{s_j}\right) + (1-\beta) max\left\{ f(\pi) - \overline{f}^{s_j}, 0 \right\} \right] & \text{if } \Pi^{\tilde{s}_j > s_j} \neq \varnothing \\ 0 & \text{otherwise} \end{cases}$$

with $\beta \in [0,1]$. Moreover, *i*'s loss from s_j is given by $l_i(s_j) = max_{\tilde{s}_j \in S_j} l_i(s_j, \tilde{s}_j)$.

Thus, *i* assesses his loss from s_j relative to $\tilde{s}_j$ by considering the set $\Pi^{\tilde{s}_j > s_j}$. For each outcome $\pi \in \Pi^{\tilde{s}_j > s_j}$, he determines the weighted sum of his material loss $\pi_i - \overline{\pi}_i^{s_j} > 0$ and his fairness loss $max\left\{ f(\pi) - \overline{f}^{s_j}, 0 \right\}$. The weighting is provided by the parameter β: If $\beta = 1$, *i*'s loss coincides with his material loss. Conversely, if $\beta = 0$, *i* only pays heed to his fairness loss. Player *i*'s loss from s_j relative to $\tilde{s}_j$ is the maximal sum of this kind with respect to all outcomes in $\Pi^{\tilde{s}_j > s_j}$, while his loss from s_j at large is his maximal loss relative to all alternatives in S_j.

I next address player *i*'s *gain* from s_j. Relative to a particular $\tilde{s}_j$, *i* derives a gain only if $\Pi^{s_j > \tilde{s}_j}$, the set of feasible outcomes under s_j that give him more payoff than what he can

---

[17] That said, it would also be possible to use a notion of fairness where equality *per se* mattered. The present specification is mainly preferred for analytical convenience. A rigorous comparison of different fairness specifications is left for future work.

[18] The two terms can be thought of as the unique payoffs to *i* yielding the fairness levels $f(\pi)$ and $\overline{f}^{s_j}$, respectively, assuming that all other players earn the same. This reading of the fairness gap is invariant to rescalings of the fairness function.

maximally earn under $\tilde{s}_j$, is non-empty. Regarding the magnitude of his gain, fairness curves again play a central role. Player *i* derives a *fairness gain* from a given $\pi \in \Pi^{s_j > \tilde{s}_j}$ if and only if π lies on a lower fairness curve than the highest curve reached in $\Pi^{\tilde{s}_j}$. The intuition is that if π lay on the same or a higher curve, *i* would consider his material gain from π to be well-deserved for contributing to no decrease in fairness. This would lead him to feel no fairness gain, and his gain would be limited to his material gain. These considerations motivate

DEFINITION 3 Player *i*'s gain from strategy $s_j \in S_j$ relative to strategy $\tilde{s}_j \in S_j$ is given by

$$g_i\left(s_j, \tilde{s}_j\right) = \begin{cases} max_{\pi \in \Pi^{s_j > \tilde{s}_j}} \left[\beta\left(\pi_i - \overline{\pi}_i^{\tilde{s}_j}\right) + \left(1-\beta\right) max\left\{ \overline{f}^{\tilde{s}_j} - f\left(\pi\right), 0 \right\} \right] & \text{if } \Pi^{s_j > \tilde{s}_j} \neq \varnothing \\ 0 & \text{otherwise} \end{cases}$$

with $\beta \in [0,1]$. Moreover, *i*'s gain from s_j is given by $g_i\left(s_j\right) = max_{\tilde{s}_j \in S_j} g_i\left(s_j, \tilde{s}_j\right)$.

Thus, to assess his gain from s_j relative to $\tilde{s}_j$, *i* determines for each $\pi \in \Pi^{s_j > \tilde{s}_j}$ the weighted sum of his material gain $\pi_i - \overline{\pi}_i^{\tilde{s}_j} > 0$ and his fairness gain $max\left\{ \overline{f}^{\tilde{s}_j} - f\left(\pi\right), 0 \right\}$. Crucially, for the fairness gain to be positive, π must be *less* fair than what is maximally achievable given $\tilde{s}_j$. If π were more fair, *i* would feel entitled to his material gain and perceive no fairness gain.[19]

I assume that the same parameters α and β are used in the calculation of gains and losses. I allow for asymmetries between the two in defining *net losses*:

DEFINITION 4 Player *i*'s *net loss* from strategy $s_j \in S_j$ is given by

$$nl_i\left(s_j\right) = l_i\left(s_j\right) - \gamma g_i\left(s_j\right)$$

with $\gamma \in [0,1]$.

The case $\gamma < 1$ allows for the possibility that "losses loom larger than gains", which is a key

[19] Hence, the expression "fairness gain" does not refer to an increase in fairness, but to a sensation of gain based on fairness considerations.

assumption in the literature on loss aversion (Kahnemann and Tversky, 1979; Köszegi and Rabin, 2006). Yet, as discussed in the Introduction, this paper is not about loss aversion as understood by that literature.

The following is immediate:

LEMMA 1 If $S_j = \{s_j\}$, we have $nl_i(s_j) = 0$.

The lemma addresses the case where *j* is passive. If *j* has only one strategy, $\Pi^{\tilde{s}_j > s_j}$ and $\Pi^{s_j > \tilde{s}_j}$ are empty for all $\tilde{s}_j \in S_j$, which implies $l_i(s_j) = 0$ and $g_i(s_j) = 0$ and therefore $nl_i(s_j) = 0$.

3 Reciprocating the Other Player's Strategy

Given the strategy s_j of player *j*, player *i* must choose an outcome from the opportunity set Π^{s_j} created by s_j. In this section, I define a preference for *net-loss reciprocation* to explain this choice. Net-loss reciprocation means that *i*'s willingness to pay for increasing the net loss that he imposes on *j* increases in his own net loss from *j*'s strategy.

I first define the net loss imposed on *j*. Let $\pi^c \in \Pi^{s_j}$ be the outcome chosen by *i* and let $\Pi^{s_j > c} = \{\pi \in \Pi^{s_j} : \pi_j > \pi_j^c\}$ and $\Pi^{s_j < c} = \{\pi \in \Pi^{s_j} : \pi_j < \pi_j^c\}$ be the outcomes in Π^{s_j} yielding *j* more and less payoff than π^c, respectively. This leads to

DEFINITION 5 Player *j*'s *loss* from $\pi^c \in \Pi^{s_j}$ is

$$l_j(\pi^c, \Pi^{s_j}) = \begin{cases} max_{\pi \in \Pi^{s_j > c}} \left[\beta(\pi_j - \pi_j^c) + (1-\beta) max\{f(\pi) - f(\pi^c), 0\} \right] & \text{if } \Pi^{s_j > c} \neq \varnothing \\ 0 & \text{otherwise} \end{cases}$$

Moreover, *j*'s *gain* from π^c is

$$g_j(\pi^c, \Pi^{s_j}) = \begin{cases} max_{\pi \in \Pi^{s_j < c}} \left[\beta(\pi_j^c - \pi_j) + (1-\beta) max\{f(\pi) - f(\pi^c), 0\} \right] & \text{if } \Pi^{s_j < c} \neq \varnothing \\ 0 & \text{otherwise} \end{cases}$$

Finally, *j*'s *net loss* from π^c is

$$nl_j\left(\pi^c,\Pi^{s_j}\right)=l_j\left(\pi^c,\Pi^{s_j}\right)-\gamma g_j\left(\pi^c,\Pi^{s_j}\right).$$

The procedure for calculating *j*'s loss and gain from π^c is analogous to calculating *i*'s loss and gain from s_j. In particular, π^c takes the role of Π^{s_j} and the alternative outcomes $\pi\in\Pi^{s_j}$ the roles of the different $\Pi^{\tilde{s}_j}$. I also assume that the same parameters α, β and γ are used.

I now turn to player *i*'s preferences governing his choice from Π^{s_j}:

ASSUMPTION 1 Player *i*'s preferences on the outcomes in Π^{s_j} are represented by

$$u_i\left(\pi,s_j\right)=v\left(\pi_i\right)+r\left(nl_j\left(\pi,\Pi^{s_j}\right),nl_i\left(s_j\right)\right)$$

where $v:\mathbb{R}\to\mathbb{R}$ and $r:\mathbb{R}^2\to\mathbb{R}$ are continuous and $dv/d\pi_i>0$ as well as $\partial^2 r/\partial nl_j\partial nl_i>0$.

Thus, given s_j, *i*'s utility from $\pi\in\Pi^{s_j}$ is additively separable into the utility from his own payoff and a reciprocation term that depends on the net loss that π imposes on *j* as well as the net loss that *i* himself derives from s_j. The marginal utility of *i*'s own payoff is always positive. Moreover, $WTP=\partial r/\partial nl_j/dv/d\pi_i$, which is *i*'s willingness to pay for increasing the net loss imposed on *j*, increases in *i*'s net loss from s_j, where WTP itself may be negative or positive. Whenever $nl_j'>nl_j$, we therefore have $\partial\left[r\left(nl_j',nl_i\right)-r\left(nl_j,nl_i\right)\right]/\partial nl_i>0$, which means that the impact of an increase in *i*'s net loss from s_j is such that for any two outcomes that differ in terms of the net loss that they impose on *j* the utility advantage (disadvantage) of the outcome imposing the larger net loss becomes larger (smaller).[20]

Furthermore, I follow McFadden (1974) and McKelvey and Palfrey (1995, 1996) in making

ASSUMPTION 2 The probability of player *i* choosing outcome $\pi\in\Pi^{s_j}$ is given by

$$Pr\left(\pi,s_j\right)=exp\left[u_i\left(\pi,s_j\right)\right]/\sum_{\tilde{\pi}\in\Pi^{s_j}}exp\left[u_i\left(\tilde{\pi},s_j\right)\right].$$

---

[20] An example is $u_i=\pi_i-\left(nl_j-nl_i\right)^2$. The partial derivative of $-\left(nl_j'-nl_i\right)^2+\left(nl_j-nl_i\right)^2$ with respect to nl_i is $2\left(nl_j'-nl_i\right)-2\left(nl_j-nl_i\right)$, which is positive if and only if $nl_j'>nl_j$.

As explained by Goeree et al (2008), this quantal response structure can be justified by disturbances on individual decision making reflecting the effects of unobservables such as mood or perceptual variations. According to this interpretation, $u_i(\pi, s_j)$ corresponds to the average utility attached to $\pi \in \Pi^{s_j}$ with each player realising a mean-zero perturbation of $u_i(\pi, s_j)$.[21] Perturbations are assumed i.i.d. across outcomes and players.[22] Assumption 2 implies that every player chooses every available outcome with positive probability. This helps in interpreting experimental data, where typically not all subjects facing a given opportunity set $\hat{\Pi}$ choose the same outcome. In this context, context dependency refers to a statistically significant shift in the empirical choice distribution on $\hat{\Pi}$ for two strategies s_j and s_j' satisfying $\Pi^{s_j} = \Pi^{s_j'} = \hat{\Pi}$. Using Assumption 2, we can explain this shift if we can show a corresponding shift in the theoretical choice distributions. I draw extensively on this method in the following section.

I conclude this section by the following lemma, which is useful in what follows:

LEMMA 2 Consider any pair of strategies $s_j, s_j' \in S_j$ with $\Pi^{s_j} = \Pi^{s_j'} = \hat{\Pi}$ where $\hat{\Pi}$ is a fixed set of outcomes and let $Pr(\pi)$ and $Pr'(\pi)$ denote the probabilities of player *i* choosing a given $\pi \in \hat{\Pi}$ when faced with s_j and s_j', respectively. Suppose that $nl_i(s_j) = nl_i(s_j')$. We then have

$$Pr(\pi) = Pr'(\pi)$$

for all $\pi \in \hat{\Pi}$. Suppose next that $nl_i(s_j') > nl_i(s_j)$. We then have

$$Pr'(\pi')/Pr(\pi') > Pr'(\pi)/Pr(\pi)$$

for all $\pi, \pi' \in \hat{\Pi}$ such that $nl_j(\pi', \hat{\Pi}) > nl_j(\pi, \hat{\Pi})$.

---

[21] In this setup, the choice of scale for utilities is not without loss of generality. In particular, multiplication of all utilities by some constant $c > 1$ makes players likelier to choose the options yielding them the highest utility. For this reason, quantal response models contain a scaling parameter λ intended to capture the degree of players' "rationality", i.e., their likelihood of choosing their most preferred options. Since the level of λ does not affect the conclusions drawn below, I set λ to 1.

[22] Moreover, in order to generate the assumed logit structure, the perturbations must take a particular stochastic form. See Goeree et al (2008) for details.

Thus, if *i* is faced with two strategies creating the same opportunity set $\hat{\Pi}$ for him and from which he derives the same net loss, his probability of choosing any given outcome in $\hat{\Pi}$ is the same across the two situations. If he does not derive the same net loss, his choice probability possesses the *monotone likelihood ratio property* for outcomes that can be ordered according to the net loss imposed on *j*. For any such pair, there is a relative shift in probability mass towards the outcome imposing the higher net loss in the situation where *i* derives the higher net loss.

4 Applications

In this section, I show that net loss reciprocation can explain the context dependency that has been documented in a number of experimental studies. Each piece of evidence considered lends additional structure to the model: Firstly, the evidence from *The Hidden Cost of Control* suggests that material factors are not irrelevant in the calculation of losses and gains ($\beta > 0$). Secondly, the evidence from *Trust* implies that gains are not fully discounted when calculating net losses ($\gamma > 0$). Thirdly, the evidence from *Ultimatum Bargaining* rules out a purely efficiency-oriented notion of fairness and provides an upper bound for the importance of material factors ($\alpha < (1-2\beta)/(1-\beta)$ and $\beta < 0.5$). Finally, the evidence from *Lost Wallets* pinpoints the fairness parameter α to equal $(1-2\beta)/(2-2\beta)$.

Discussing each piece of evidence in turn, I also address the problems faced by the reciprocity models of Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) in accounting for these experimental results. Recall that both models make use of players' second-order beliefs, i.e., their beliefs about the other player's belief about their own choice of strategy. As the studies considered in this section do not measure these beliefs, I use the actual behaviour of experimental subjects as a "stand-in" for second-order beliefs. This is in keeping with the equilibrium spirit of these models, which requires beliefs to coincide with actual behaviour.

Although both models rely on second-order beliefs, they differ in how they define the kindness of the other player's strategy. Suppose that we want to evaluate the kindness of s_j to player *i*. To do so, Dufwenberg and Kirchsteiger (2004) compare *i*'s expected payoff from s_j given his second-order belief to what he could have minimally and maximally earned under the alternatives given again his second-order belief. The strategy is viewed as kind (unkind) if it

gives i more (less) than half of what he could have minimally and maximally earned. In contrast, Falk and Fischbacher (2006) focus on the expected outcome implemented by s_j given i's second-order belief. If i earns more (less) than j under this outcome, s_j is viewed as kind (unkind). The model also takes into account the alternatives available to j. For instance, if the outcome implemented by s_j puts i ahead, this is only viewed as fully kind if there are alternatives under which i would have earned less. The intuition is that the kindness embodied by s_j cannot be regarded as fully intentional otherwise. Appendix B contains a more detailed exposition of the two models.

4.1 The Hidden Cost of Control

Falk and Kosfeld (2006) study the reaction of agents (A) to a principal's (P) decision to control their choice of how much productive effort to exert. Two of their treatments are directly relevant for us: In the control game (CG), the principal first decides whether to control the agent or not. If not controlled by the principal, the agent can exert any effort level $e \in \{0,1,2,...,120\}$. Payoffs are given by $2e$ for the principal and $120-e$ for the agent. In contrast, if the principal has controlled the agent, effort is restricted to be at least ten, i.e., $e \in \{10,11,...,120\}$, the mapping from effort to payoffs being the same as after no control. The second treatment is a dictator game (DG) that is identical to the subgame of CG after control. As a result, the agent (who takes the role of player i) chooses from $\{(120-e,2e): e \in \{10,11,...,120\}\}$ both in DG and after control in CG.[23] We have

PROPOSITION 1 Suppose that $\beta > 0$. The net loss that the agent derives from the principal is given by $nl_A^{CG} = 10\beta$ after control in CG and by $nl_A^{DG} = 0 < nl_A^{CG}$ in DG. The net loss that the agent imposes on the principal decreases in effort. Hence, average effort is higher in DG.

The agent derives a net loss of zero in DG because the principal is passive in this treatment. After control in CG, the agent only derives a material loss (of 10) because control only rules out outcomes that are less fair than the outcomes attainable under control for being less efficient and

[23] Recall that outcomes have the format $\pi = (\pi_i, \pi_j)$.

containing a smaller minimal payoff. Hence, for the agent to derive a loss from control, the weight on his material loss must not be zero ($\beta > 0$). Regarding the net loss that the agent imposes on the principal, a one-unit increase in effort decreases the principal's material loss and increases her material gain. Fairness losses and gains do not counteract this: The principal's fairness loss is non-increasing in effort. In fact, it decreases except for high effort, where the principal may not feel entitled to additional effort if her fairness notion leans toward a concern for the less well-off. Likewise, the principal's fairness gain is non-decreasing in effort. It is zero for low effort levels, where the principal feels entitled to effort, but may increase for higher effort. As a result, as long as $\beta > 0$, the principal's net loss unambiguously decreases in effort.

Lemma 2 then implies that the effort distributions in the two situations possess the monotone likelihood ratio property, with probability mass shifting towards higher effort in the situation where agents derive the lower net loss. Consequently, average effort is predicted to be higher in *DG*. This matches the results of Falk and Kosfeld (2006), who report that average effort is significantly lower after control in *CG*. The agents in their experiment provide a mean effort of 17.5 after control in *CG*, but of 28.7 in *DG*.[24]

The reciprocity model of Dufwenberg and Kirchsteiger (2004), referred to as DK in what follows, struggles to account for these findings. Given that agents in *CG* exert more effort after no control than after control and exerting more effort means less payoff for agents themselves,[25] DK view control (somewhat counter-intuitively) as kind to agents. Consequently, they predict more effort, which is kinder to principals, after control in *CG* than in *DG*, where the principal is passive and hence neither kind nor unkind. Reliance on second-order beliefs has a perverse consequence here. From an intuitive viewpoint, it is clear agents are disgruntled at being controlled if the principal has had the choice of not controlling them. The effect of this disgruntlement, namely, that agents keep more money for themselves, is used by DK as "evidence" for the conclusion that they have no reason for being disgruntled. This confuses cause and effect of agents' emotional response to the principal's behaviour. My approach avoids this problem because it relies on agents' opportunity sets given the principal's strategy without taking into account their reaction to the latter.

---

[24] Falk and Kosfeld (2006) also study other control levels. However, they implement no corresponding dictator games, which means that no foregone-option effects can be studied.

[25] Falk and Kosfeld (2006) report a mean effort after no control of 23.0.

The reciprocity model of Falk and Fischbacher (2006), denoted FF in what follows, runs into similar problems. Given that responders exert less effort after control in *CG* than in *DG*, agents view themselves as being put further ahead of principals in *CG*. Moreover, principals could have put agents less ahead in *CG* by not controlling them. As a result, control in *CG* is kinder, which is at odds with the fact that agents are kinder to principals in *DG* by exerting more effort.

4.2 Trust

McCabe et al (2003) study a simple trust game (*TG*) in which the first mover (*FM*) can either implement the no-trust outcome $(20,20)$ or trust the second mover (*SM*). If trusted by the first mover, the second mover can choose between $(30,15)$ and $(25,25)$. The first entry in each payoff vector denotes the payoff to the second mover, who takes the role of player *i* in what follows. The authors compare second mover behaviour in this game to behaviour in a dictator game (*DG*) in which the first mover is passive and the second mover has a choice between the same two outcomes as after trust in *TG*. We have

PROPOSITION 2 Suppose that $\gamma > 0$ and either $\alpha < 2/3$ or $\beta > 0$. The net loss that the second mover derives from trust in *TG* is $nl_{SM}^{TG} = -\gamma\left(\beta 10 + (1-\beta)max\{5 - 7.5\alpha, 0\}\right)$, while his net loss in *DG* is $nl_{SM}^{DG} = 0 > nl_{SM}^{TG}$. The net loss that the second mover imposes on the first mover is $nl_{FM} = -10\gamma\beta$ if the second mover chooses $(25,25)$ and $nl'_{FM} = 10 - 7.5\alpha + 7.5\alpha\beta > nl_{FM}$ if he chooses $(30,15)$. Thus, second movers are more likely to choose $(25,25)$ after trust in *TG*.

We have $nl_{SM}^{DG} = 0$ because the first mover is passive in *DG*. As for *TG*, for the second mover's net loss from trust to be negative, we must have $\gamma > 0$ and either $\alpha < 2/3$ or $\beta > 0$. On the one hand, the gain from trust must not be fully discounted, which is guaranteed by $\gamma > 0$. Further, there are two ways for the gain from trust to be positive: Either we have $\beta > 0$, which ensures that the second mover puts positive weight on his material gain of 10. Or we have $\alpha < 2/3$, which means that fairness does not lean too much towards efficiency. In this case, the second mover perceives $(30,15)$, which he can implement after trust, as less fair than the no-trust outcome $(20,20)$ because of the smaller minimal payoff. He then derives a fairness gain from trust, which makes his overall gain positive irrespective of β.

Furthermore, the second mover clearly imposes a lower net loss on the first mover by choosing $(25,25)$. Net-loss reciprocation implies that the second mover is more likely to choose $(25,25)$ instead of $(30,15)$ in *TG* because of his lower derived net loss This context dependency is confirmed by McCabe et al (2003), who report that second movers choose $(25,25)$ with a frequency of 0.65 after trust in *TG* and of only 0.33 in *DG*, this difference being significant.[26]

4.3 Ultimatum Bargaining

Falk et al (2003) study four binary ultimatum games in each of which the offer "2 for the responder, 8 for the proposer" is available to the proposer. The games differ regarding the second offer. In three, there exists a true alternative, namely, "5 for both", "8 for the responder, 2 for the proposer" and "0 for the responder, 10 for the proposer", respectively. In the fourth, the proposer is effectively passive because the alternative offer is also "2 for the responder, 8 for the proposer". Thus, letting the responder take the role of player *i*, acceptance of the alternative implements $(5,5)$, $(8,2)$, $(0,10)$ and $(2,8)$, respectively. In what follows, I refer to the four treatments by these outcomes. We have

PROPOSITION 3 Suppose that $\alpha < (1-2\beta)/(1-\beta)$, $0 < \beta < 0.5$ and $\gamma > 0$. The net losses that the responder derives from "2 for the responder, 8 for the proposer" in the four treatments are given by $nl_R^{(5,5)} = 3 - 3\alpha + 3\alpha\beta$, $nl_R^{(8,2)} = 6\beta$, $nl_R^{(2,8)} = 0$ and $nl_R^{(0,10)} = -2\gamma\beta$, respectively, where we have $nl_R^{(5,5)} > nl_R^{(8,2)} > nl_R^{(2,8)} > nl_R^{(0,10)}$. The net loss that the responder imposes on the proposer by accepting "2 for the proposer, 8 for the responder" is $nl_P = -8\gamma\beta$, while rejection imposes $nl'_p = 2 + 3\alpha + 6\beta - 3\alpha\beta > nl_p$. We therefore have $Pr^{(5,5)} > Pr^{(8,2)} > Pr^{(2,8)} > Pr^{(0,10)}$, where Pr^x denotes the rejection probability in treatment $x \in \{(5,5),(8,2),(2,8),(0,10)\}$.

---

[26] The approaches of Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) can also account for this. Under the former, trust gives the second mover more in expected terms than no trust, which leads the second mover to view trust as kind, giving him an incentive to be kind to the first mover by repaying trust. Under the latter, given that trust is not always reciprocated, trust puts the second mover ahead of the first mover in expected terms. Moreover, the second mover would have earned less had he not been trusted. As a result, the kindness embodied by trust is fully intentional.

Crucially, Proposition 3 asserts that losses do not track material losses. The responder derives the highest overall loss in treatment $(5,5)$, whereas his material loss is largest in $(8,2)$. To see this, notice that "2 for the responder, 8 for the proposer" means a material loss of 6 for the responder if the alternative has been "8 for the responder, 2 for the proposer", but of only 3 if it has been "5 for both". The responder's fairness loss is $3-3\alpha$ in $(5,5)$, but is necessarily zero in $(8,2)$ because the outcome $(8,2)$ is always as fair as the outcome $(2,8)$ irrespective of α. The dissociation of losses from material losses is achieved by the fairness loss in $(5,5)$ being sufficiently large and fairness losses playing a sufficiently large role. Formally, $nl_R^{(5,5)} > nl_R^{(8,2)}$ is equivalent to $3-3\alpha+3\alpha\beta > 6\beta \Leftrightarrow \alpha < (1-2\beta)/(1-\beta)$. This can only be satisfied by $\alpha \geq 0$ if we have $1-2\beta > 0 \Leftrightarrow \beta < 0.5$.

Further, we have $nl_R^{(2,8)} = 0$ because the proposer is passive in $(2,8)$ and $nl_R^{(0,10)} = -2\gamma\beta$ because the responder derives no fairness gain in $(0,10)$. The reason is that he feels entitled to "2 for the responder, 8 for the proposer" if the alternative is "0 for the responder, 10 for the proposer". We have $6\beta > 0 > -2\gamma\beta$ because of $\beta > 0$ and $\gamma > 0$.

All in all, since rejection of the fixed offer "2 for the responder, 8 for the proposer" imposes a higher net loss on the proposer than acceptance, the responder is most likely to reject in $(5,5)$, second most likely in $(8,2)$ etc. This is largely consistent with the results of Falk et al (2003), who report the following rejection frequencies:

$$Pr^{(5,5)} = 0.44 > Pr^{(8,2)} = 0.27 > Pr^{(2,8)} = 0.18 > Pr^{(0,10)} = 0.09,$$

these differences being statistically significant except for the last one. Although the last difference has the right sign, my approach faces a difficulty here. The absence of a significant difference would be explained by net-loss reciprocation if $nl_R^{(2,8)} = nl_R^{(0,10)}$. Since we have $nl_R^{(2,8)} = 0$ because the proposer is passive in this treatment, whereas $nl_R^{(0,10)} = -2\gamma\beta$, we would have to assume either $\beta = 0$ or $\gamma = 0$ or both. Alternatively, if both γ and β are positive but small, their product can be viewed as approximately zero. I return to this issue below.

In contrast, DK face difficulties in rationalising the difference between the treatments $(5,5)$ and $(8,2)$. According to DK, the status quo offer "2 for the responder, 8 for the proposer" is *less* kind in $(8,2)$ than $(5,5)$ for the following reason: In $(5,5)$, the status quo is accepted with a probability of 0.66 and the alternative for sure, which makes for a kindness of the status quo

of $k = 0.66 \cdot 2 - 0.5[5 + 0.66 \cdot 2] = -1.84$.[27] In $(8,2)$, the acceptance probabilities are 0.73 for the status quo and 0.98 for the alternative. The fact that the status quo is accepted with a higher probability tends to make the status quo kinder. However, what works in the opposite direction is that responders could have earned more under the alternative. This second effect dominates since we have $k = 0.73 \cdot 2 - 0.5[0.98 \cdot 8 + 0.73 \cdot 2] = -3.72 < -1.84$. What DK do not take into account is responders' sense of entitlement, in particular, that they feel less entitled to their forgone earnings in $(8,2)$ because these forgone earnings derive from an outcome that is only as fair as status quo. Such considerations are at the heart of my approach.

The basic model in FF also fails to explain the difference in responder behaviour between $(5,5)$ and $(8,2)$. The reason is the binary nature of the intention factor (see Appendix B), which simply asks whether or not responders could have earned more in expected terms than under the status quo, which is the case in both treatments. Hence, the intention factor does not capture the fairness difference between the two alternatives and the differential sense of entitlement that this difference creates. The appendix in FF contains a richer version of their model designed to address this problem.

4.4 Lost Wallets

Dufwenberg and Gneezy (2000) fail to establish context dependency in second-mover behaviour in a series of trust games termed "Lost Wallet Games". Their common feature is that the second mover can split 20 units of payoff between himself and the first mover in the event of trust. The games differ regarding the no-trust outcome, which, letting the second mover take the role of player *i*, is given by $(0, f)$ with $f \in \{4,7,10,13,16\}$. That is, the no-trust payoff for the second mover is always zero, while the games differ with respect to the no-trust payoff for the first mover. For simplicity, I focus on the polar cases $f = 4$ and $f = 16$ because the absence of context dependency is most puzzling between them. I refer to the two treatments by the respective no-trust outcome, namely $(0,4)$ and $(0,16)$. Servatka and Vadovic (2009) draw on the basic setup of Dufwenberg and Gneezy (2000), while varying the inequality of the no-trust outcome. In their two treatments, the no-trust outcomes are given by $(0,10)$ and $(5,5)$. Like

[27] See Appendix B for a detailed exposition of the kindness function *k*.

Dufwenberg and Gneezy (2000), they fail to establish a significant difference in return transfers, i.e., there is again no context dependency.

Consistent with these empirical findings, we have

PROPOSITION 4 Suppose that we have either $\gamma > 0$, $0 < \beta \leq 0.5$ and $\alpha = (1-2\beta)/(2-2\beta)$ or $\gamma > 0$, $\beta = 0$ and $\alpha \geq 0.5$ or $\gamma = 0$ The net loss that the second mover derives from the first mover's trust is $nl_{SM}^{(0,4)} = -20\gamma\beta$ if the no-trust outcome is $(0,4)$ and $nl_{SM}^{(0,16)} = -20\gamma\beta = nl_{SM}^{(0,4)}$ if it is $(0,16)$. As a result, the repayment distributions do not differ between the two situations. If the no-trust outcome is $(0,10)$, the net loss derived from trust is $nl_{SM}^{(0,10)} = -20\gamma\beta$, while it is $nl_{SM}^{(5,5)} = -\gamma\left(15\beta + (1-\beta)\left(max\{5-10\alpha, 0\}\right)\right) = nl_{SM}^{(0,10)}$ if the no-trust outcome is $(5,5)$. Again, the repayment distributions do not differ between the two situations.

In the first two cases, the second mover's gain from trust is limited to his material gain of 20 because the no-trust outcomes $(0,4)$ and $(0,16)$ are not fairer than any outcome in the second mover's opportunity set after trust. Representing a payoff sum of less than 20, both no-trust outcomes are less efficient than the outcomes available after trust. Moreover, the minimal payoff is zero in each case, which is also the minimal payoff available after trust (if the second mover shares nothing). As a result, the responder feels entitled to his material gain causing his gain to be limited to the latter. As there is no loss from trust, the net loss is $-20\gamma\beta$ in each case.

For the same reasons, the second mover's net loss is $-20\gamma\beta$ if the no-trust outcome is $(0,10)$. If it is $(5,5)$, the second mover's material gain from trust is 15. For $nl_{SM}^{(5,5)} = -20\gamma\beta$ to hold, we can impose $\gamma = 0$ meaning that gains are fully discounted. Alternatively, if $\gamma > 0$, we can let $\beta = 0$ and $\alpha \geq 0.5$ meaning that the weight on material gains is zero, but the second mover derives no fairness gain from trust. Indeed, if $\alpha \geq 0.5$, the least fair outcome after trust, namely, $(20,0)$, is at least as fair as $(5,5)$ because efficiency receives sufficient weight in the fairness function. Finally, if $\gamma > 0$ and $\beta > 0$, the second mover must derive a positive fairness gain from trust in $(5,5)$ to offset his larger material gain in $(0,10)$. This is the case if $\alpha < 0.5$ because fairness then leans towards a concern for the less well-off. In these conditions, we have $nl_{SM}^{(5,5)} = -\gamma(5-10\alpha+10\beta+10\alpha\beta)$, which equals $-20\gamma\beta$ if and only if $\alpha = (1-2\beta)/(2-2\beta)$. This equality can only be satisfied by $\alpha \geq 0$ if $\beta \leq 0.5$. Also, given $\beta > 0$, $\alpha < 0.5$ as assumed.

Dufwenberg and Gneezy (2000) also implement a dictator treatment (*DG*) in which dictators

face the same opportunity set as second movers after trust. The authors report no significant difference in transfers between *DG* on the one hand and the treatments $(0,4)$ and $(0,16)$ on the other. For net-loss reciprocation to explain this, we must have $-20\gamma\beta = nl_{SM}^{DG} = 0$. We can make the equality hold by imposing either $\beta = 0$ or $\gamma = 0$. I return to this issue below.

Regarding the intention-based models, DK cannot account for there being no difference between $(0,10)$ and $(5,5)$. The problem is that DK pick up on second movers' differential earnings from the no-trust outcome. Given that second movers behave in the same way after trust, this yields the conclusion that trust is kinder in $(0,10)$ because second movers gain more from it in expected terms. My approach can navigate around this problem because second movers' higher material gain in $(0,10)$ can be offset by a higher fairness gain in $(5,5)$.

In contrast, FF can explain most absences of treatment differences. Given that average second-mover behaviour is the same, second movers view first movers as intending the same expected outcome (putting them ahead of first movers) in all treatments. Moreover, first movers could have treated second movers worse by not trusting them (except in *DG*). Hence, first movers are equally kind to second movers in all treatments rationalising the absence of context dependency. FF (like my approach) only struggle to explain behaviour in *DG* relative to the other treatments because first movers are passive there, which should cause second movers to share less.

5 Discussion

The examples considered in the preceding section are instructive with regard to the calibration of the model. The preferred specification is

$$0 < \gamma \leq 1,\ 0 < \beta < 0.5 \text{ and } \alpha = (1-2\beta)/(2-2\beta)$$

which is well-supported by the experimental data considered in this paper. The interpretation is that gains are not fully discounted in the calculation of net losses ($0 < \gamma \leq 1$) and that the weight on material losses and gains is neither zero nor too large ($0 < \beta < 0.5$). Given the restrictions on β, the condition on α implies $0 < \alpha < 0.5$ meaning that fairness leans towards a concern for

the less well-off.

Imposing $\beta > 0$ and $\gamma > 0$ fails to explain two pieces of evidence, namely, the treatment $(0,10)$ from *Ultimatum Bargaining* and *DG* from *Lost Wallets*. To account for them, the above specification could be modified by setting $\gamma = 0$. This parameterisation, which implies that gains are fully discounted, can account for all the evidence except that from *Trust*. In particular, the evidence from *Lost Wallets* is explained almost trivially by reducing net losses to zero in all treatments. Effectively, this specification negates the importance of positive reciprocity by asserting that people do not react to gains that they derive from others. Charness and Rabin (2002) provide further evidence that positive reciprocity is a less important motivational force than negative reciprocity.[28]

All in all, this section has demonstrated that net-loss reciprocation in conjunction with the method for calculating net losses developed in this paper can by and large account for the existence or absence of context dependency in a number of experimental studies. I have also shown that existing models of intention-based reciprocity face problems in explaining this evidence comprehensively. This is particularly true for the model of Dufwenberg and Kirchsteiger (2004), while at least the extended version of Falk and Fischbacher (2006) performs relatively well. Yet, even in its extended form, the latter only captures players' sense of entitlement in an approximate, qualitative fashion. My approach allows to precisely quantify this sense via fairness losses and gains.

6 Conclusion

This paper presents a qualitative preference model for two-player interactions building on the idea of net-loss reciprocation. Net-loss reciprocation asserts that a player's willingness to impose net losses on the other increases in the net loss that he derives from the other player's strategy. The paper shows that net-loss reciprocation can account for the context dependencies in individual behaviour (or absences thereof) that arise in a number of experimental studies.

The main difficulty faced by net-loss reciprocation relates to the status of positive reciprocity.

[28] In applications of loss aversion, it is often assumed for the sake of simplicity that only losses count

In the light of the evidence considered in this paper, it is not clear whether players fully discount any gain they derive from the other player's strategy or whether they take this gains into account leading them to become less willing to impose a net loss on the other. Apart from this, I find conclusive evidence that both material and fairness considerations matter to the determination of net losses, with fairness being somewhat more important. I also establish that a regard for the less well-off as opposed to a pure concern for material efficiency plays an important role in fairness assessments.

Given the relative success of my approach in explaining context dependency when compared to existing models of intention-based reciprocity, the development of full-fledged utility models incorporating net-loss reciprocation seems worthwhile. These models could be used to analyse more general classes of games.[29] An advantage of such models compared to intention-based models is their direct testability using standard experimental data as they do not rely on higher-order beliefs.

[29] In Appendix C, an extension of the model to more players (including Nature) is proposed.

Appendix A: Proofs

PROOF OF LEMMA 2

Let $nl_i\left(s'_j\right)=nl'_i$ and $nl_i\left(s_j\right)=nl_i$. Also, let $nl_j\left(\pi',\hat{\Pi}\right)=nl'_j$ and $nl_j\left(\pi,\hat{\Pi}\right)=nl_j$. The first part, where $nl_i=nl'_i$, is immediate since we have

$$Pr(\pi)=exp\left[v(\pi_i)+r\left(nl_j,nl_i\right)\right]\Big/\sum_{\breve{\pi}\in\hat{\Pi}}exp\left[v(\breve{\pi}_i)+r\left(n\breve{l}_j,nl_i\right)\right]$$

and

$$Pr'(\pi)=exp\left[v(\pi_i)+r\left(nl_j,nl'_i\right)\right]\Big/\sum_{\breve{\pi}\in\hat{\Pi}}exp\left[v(\breve{\pi}_i)+r\left(n\breve{l}_j,nl'_i\right)\right]$$

for all $\pi\in\hat{\Pi}$ by Assumptions 1 and 2.

Next, I show that $Pr'(\pi')/Pr(\pi')>Pr'(\pi)/Pr(\pi)\Leftrightarrow Pr'(\pi')/Pr'(\pi)>Pr(\pi')/Pr(\pi)$ for all $\pi,\pi'\in\hat{\Pi}$ if $nl'_i>nl_i$ and $nl'_j>nl_j$. By Assumption 2, we can express the second inequality as

$$\frac{exp\left[v(\pi')\right]\cdot exp\left[r\left(nl'_j,nl'_i\right)\right]}{exp\left[v(\pi)\right]\cdot exp\left[r\left(nl_j,nl'_i\right)\right]}>\frac{exp\left[v(\pi')\right]\cdot exp\left[r\left(nl'_j,nl_i\right)\right]}{exp\left[v(\pi)\right]\cdot exp\left[r\left(nl_j,nl_i\right)\right]}\Leftrightarrow$$

$$\frac{exp\left[r\left(nl'_j,nl'_i\right)\right]}{exp\left[r\left(nl_j,nl'_i\right)\right]}>\frac{exp\left[r\left(nl'_j,nl_i\right)\right]}{exp\left[r\left(nl_j,nl_i\right)\right]}.$$

Logarithmation of both sides yields

$$r\left(nl'_j,nl'_i\right)-r\left(nl_j,nl'_i\right)>r\left(nl'_j,nl_i\right)-r\left(nl_j,nl_i\right),$$

which holds by our assumptions on $r(\cdot)$.■

PROOF OF PROPOSITION 1

The principal is passive in *DG*. By Lemma 1, we have $nl_A^{DG}=0$. In *CG*, letting *NC* denote no

control, we have $\Pi^{NC>C} = \{(120-e,2e) : e \in \{0,1,...,9\}\}$. Thus, if not controlled, the agent can earn more than what he can maximally earn if controlled. As a result, $g_A^{CG} = 0$ and

$$l_A^{CG} = max_{\pi \in \Pi^{NC>C}} \left[\beta(\pi_A - 110) + (1-\beta) max\{f(\pi) - \overline{f}^C, 0\} \right],$$

where $\overline{f}^C$ is the highest fairness level attained in Π^C. However, the fairness of the outcomes in $\Pi^{NC>C}$ is below $\overline{f}^C$ because both efficiency and a concern for the less well-off mandate increasing effort to 40. Consequently, we have $l_A^{CG} = \beta 10 = n l_A^{CG} > 0$.

Next, I show that the net loss that the agent imposes on the principal decreases in *e*. Suppose that the agent chooses $e = x \in \{10,...,119\}$. We then have $\pi^c = (120-x, 2x)$,

$$l_P = max_{\pi \in \Pi^{C>c}} \left[\beta(\pi_P - 2x) + (1-\beta) max\{f(\pi) - f(\pi^c), 0\} \right] \text{ and}$$

$$g_P = max_{\pi \in \Pi^{C<c}} \left[\beta(2x - \pi_P) + (1-\beta) max\{f(\pi) - f(\pi^c), 0\} \right]$$

where $\Pi^{C>c} = \{(120-e,2e) : e \in \{x+1,...,120\}\}$ and $\Pi^{C<c} = \{(120-e,2e) : e \in \{10,...,x-1\}\}$. If effort increases by one unit, i.e., if $e = y = x+1$, we have $\pi^{c'} = (120-x-1, 2x+2)$ implying

$$l'_P = max_{\pi \in \Pi^{C>c'}} \left[\beta(\pi_P - 2x - 2) + (1-\beta) max\{f(\pi) - f(\pi^{c'}), 0\} \right] \text{ and}$$

$$g'_P = max_{\pi \in \Pi^{C<c'}} \left[\beta(2x + 2 - \pi_P) + (1-\beta) max\{f(\pi) - f(\pi^{c'}), 0\} \right]$$

where $\Pi^{C>c'} = \{(120-e,2e) : e \in \{x+2,...,120\}\}$ and $\Pi^{C<c'} = \{(120-e,2e) : e \in \{10,...,x\}\}$.

Suppose first that $x < 40$. We have $f(\pi^{c'}) > f(\pi^c)$ because both efficiency and a concern for the less well-off point towards increasing effort. As a result, $l_P > l'_P$ because the maximisation for determining l'_P takes place on the set $\Pi^{C>c'}$, which is a subset of the set $\Pi^{C>c}$ used for establishing l_P and $\beta > 0$. As for gains, the outcomes in $\Pi^{C<c}$ are less fair than π^c because they represent effort further away from 40. The same holds for $\Pi^{C<c'}$ and $\pi^{c'}$. As a result, gains are limited to material gains, and we have $g'_P > g_P$ because $\Pi^{C<c'}$ is a super-set of $\Pi^{C<c}$ and $\beta > 0$. All in all, $e = y$ imposes a smaller net loss on the principal than $e = x$.

Suppose next that $x \geq 40$. Regarding losses, if $f\left(\pi^{c'}\right) \geq f\left(\pi^{c}\right)$ because the fairness function leans towards efficiency, we have $l_P > l'_P$ for the same reasons as above. If $f\left(\pi^{c'}\right) < f\left(\pi^{c}\right)$ because the fairness function leans towards a concern for the less well off, the linear nature of the fairness function implies that π^{c} is fairer than all elements in $\Pi^{C>c}$ and likewise for $\pi^{c'}$ and $\Pi^{C>c'}$. As a result, losses are limited to material losses and we have $l_P > l'_P$ because $\Pi^{C>c}$ is a super-set of $\Pi^{C>c'}$ and $\beta > 0$. As for gains, if $f\left(\pi^{c'}\right) \geq f\left(\pi^{c}\right)$, all outcomes in $\Pi^{C<c}$ are not fairer than π^{c} and likewise for $\Pi^{C<c'}$ and $\pi^{c'}$, which implies that there are only material gains. We have $g'_P > g_P$ because $\Pi^{C<c'}$ is a super-set of $\Pi^{C<c}$ and $\beta > 0$. For the same reason, we have $g'_P > g_P$ if $f\left(\pi^{c'}\right) < f\left(\pi^{c}\right)$. Again, $e = y$ imposes a smaller net loss on the principal.

Since the principal's net loss decreases in effort, Lemma 2 together with $nl_A^{CG} > nl_A^{DG}$ implies that the effort distribution in *DG* first-order stochastically dominates the distribution after control in *CG*, which implies that average effort is higher in *DG*.■

PROOF OF PROPOSITION 2

Since first movers are passive in *DG*, we have $nl_{SM}^{DG} = 0$. In *TG*, second movers gain from trust in material terms, which implies their loss is zero. Their gain is given by

$$g_{SM}^{TG} = \beta 10 + (1-\beta)\left(max\left\{20 - \alpha 22.5 - (1-\alpha)15, 0\right\}\right) = \beta 10 + (1-\beta)max\left\{5 - 7.5\alpha, 0\right\}$$

because $(30,15)$ corresponds to a higher material and fairness gain than $(25,25)$. We have $nl_{SM}^{TG} = -\gamma\left(\beta 10 + (1-\beta)max\left\{5 - 7.5\alpha, 0\right\}\right) < 0$ because of our parameter assumptions.

I next show that the net loss imposed on first movers through $(30,15)$ exceeds that imposed through $(25,25)$. First movers derive no loss from $(25,25)$ and a material gain of 10. They derive no fairness gain because $(25,25)$ is superior from the viewpoint of both efficiency and a concern for the less well off. As a result, $nl_{FM} = -10\gamma\beta$. From $(30,15)$, first movers derive a loss of

$$\beta 10 + (1-\beta)\left(25 - \alpha 22.5 - (1-\alpha)15\right) = \beta 10 + (1-\beta)(10 - 7.5\alpha) = 10 - 7.5\alpha + 7.5\alpha\beta$$

and no gain, which implies that $nl'_{FM} = 10 - 7.5\alpha + 7.5\alpha\beta > nl_{FM}$. Lemma 2 together with

$nl_{SM}^{DG} > nl_{SM}^{TG}$ then implies that second movers are more likely to choose $(25,25)$ in *TG*.■

PROOF OF PROPOSITION 3

I refer to the offer "2 for the responder, 8 for the proposer" as *X* and the alternative offer in a given treatment as *Y*. We have $\Pi^X = \{(2,8),(0,0)\}$. In $(5,5)$, $\Pi^{Y>X} = \{(5,5)\}$. As a result, responders derive no gain from *X* in this case and a material loss of 3. The highest fairness level reached in Π^X is $f = \alpha 5 + (1-\alpha)2$, whereas $f((5,5)) = 5$. As a result,

$$nl_R^{(5,5)} = \beta 3 + (1-\beta)(5 - \alpha 5 - (1-\alpha)2) = 3 - 3\alpha + 3\alpha\beta .$$

In $(8,2)$, we have $\Pi^{Y>X} = \{(8,2)\}$, which implies a material loss from *X* of 6 and no gain. Responders derive no fairness loss because $(8,2)$ and $(2,8)$ lie on the same fairness curve irrespective of α. Consequently, $nl_R^{(8,2)} = 6\beta$. Since proposers are passive in $(2,8)$, $nl_R^{(2,8)} = 0$. Finally, we have $\Pi^{Y>X} = \varnothing$, but $\Pi^{X>Y} = \{(2,8)\}$ in $(0,10)$ meaning that responders derive no loss from *X* and a material gain of 2. The highest fairness level reached in Π^Y is $f = \alpha 5$. Since $\alpha 5 + (1-\alpha)2 \geq \alpha 5$, responders feel entitled to their material gain. We thus have $nl_R^{(0,10)} = -2\gamma\beta$. From our parameter assumptions, it follows that $nl_R^{(5,5)} > nl_R^{(8,2)} > nl_R^{(2,8)} > nl_R^{(0,10)}$.

I now turn to net losses imposed on proposers. Since $\Pi^X = \{(2,8),(0,0)\}$, proposers derive no loss from acceptance. Their gain is limited to $\beta 8$ because $(2,8)$ is fairer than $(0,0)$. All in all, $nl_P = -8\gamma\beta$. Conversely, next to a material loss of 8 from rejection, proposers suffer a fairness loss of $f((2,8)) - f((0,0)) = \alpha 5 + (1-\alpha)2 = 2 + 3\alpha$. As a result, their net loss is given by $nl'_p = \beta 8 + (1-\beta)(2 + 3\alpha) = 2 + 3\alpha + 6\beta - 3\alpha\beta > nl_p$. From $nl_R^{(5,5)} > nl_R^{(8,2)} > nl_R^{(2,8)} > nl_R^{(0,10)}$ and Lemma 2, it then follows that $Pr^{(5,5)} > Pr^{(8,2)} > Pr^{(2,8)} > Pr^{(0,10)}$.■

PROOF OF PROPOSITION 4

The opportunity set given trust is $\Pi^T = \{(20-r,r) : r \in \{0,1,...,20\}\}$ where *r* is the amount shared. Denoting no trust by *NT*, we have $\Pi^{T>NT} = \{(20-r,r) : r \in \{0,1,...,19\}\}$ in both treatments because all outcomes in Π^T except $(0,20)$ give the second mover more than the no-trust outcome. The fairness associated with no trust is $f = \alpha 2$ in $(0,4)$ and $f = \alpha 8$ in $(0,16)$,

while the fairness reached in $\Pi^{T>NT}$ as a function of $r \in \{0,1,...,19\}$ is given by

$$f = \alpha 10 + (1-\alpha) min\{r, 20-r\} > \alpha 8 > \alpha 2,$$

which implies that gains are limited to material gains. Hence, we have $g_{SM}^{(0,4)} = g_{SM}^{(0,16)} = \beta 20$ and $nl_{SM}^{(0,4)} = nl_{SM}^{(0,16)} = -\gamma\beta 20$. By Lemma 2, the repayment distributions are then the same.

Next, we show that $nl_{SM}^{(0,10)} = nl_{SM}^{(5,5)}$. In $(0,10)$, we have $nl_{SM}^{(0,10)} = -\gamma\beta 20$ for the same reasons as before. Since the second mover derives no loss from trust, $nl_{SM}^{(0,10)} = nl_{SM}^{(5,5)}$ is trivially satisfied if $\gamma = 0$ because all gains are then fully discounted.

Suppose instead that $\gamma > 0$, $0 < \beta \leq 0.5$ and $\alpha = (1-2\beta)/(2-2\beta)$ and notice that the last two conditions imply $0 \leq \alpha < 0.5$. In $(5,5)$, we have $\Pi^{T>NT} = \{(20-r, r) : r \in \{0,1,...,14\}\}$. Given that $\alpha < 0.5$, the least fair outcome in $\Pi^{T>NT}$, namely, $(20,0)$, is less fair than $(5,5)$, which implies a positive fairness gain from trust. As the maximisation of material and fairness gains points into the same direction, with $(20,0)$ maximising both, we obtain

$$g_{SM}^{(5,5)} = \beta 15 + (1-\beta)(5-\alpha 10) = 5 - 10\alpha + 10\beta + 10\alpha\beta.$$

We then have

$$nl_{SM}^{(5,5)} = nl_{SM}^{(0,10)} \Leftrightarrow -\gamma(5 - 10\alpha + 10\beta + 10\alpha\beta) = -\gamma\beta 20 \Leftrightarrow \alpha = (1-2\beta)/(2-2\beta),$$

as assumed.

Finally, if $\gamma > 0$, $\beta = 0$ and $\alpha \geq 0.5$, the second mover disregards his material gains, but his fairness gain from trust is zero in $(5,5)$ because fairness leans towards efficiency. This implies $nl_{SM}^{(5,5)} = 0 = nl_{SM}^{(0,10)}$.■

Appendix B: Intention-Based Models of Reciprocity

In this Appendix, I sketch the main features of the reciprocity models of Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006). I focus on how player *i* evaluates the *kindness* of player *j*'s pure strategy s_j. Conceptually, the kindness of s_j plays the same role as *i*'s net loss from s_j in my model. Since *i* observes s_j in the applications considered in this paper, *i* need not form a belief about s_j. At the same time, both approaches draw on what is called *i*'s *second-order belief*, i.e., *i*'s belief about *j*'s belief about *i*'s own strategy. Both allow this belief to refer to a behaviour strategy σ_i. I denote *i*'s second-order belief about σ_i by σ_{iji}. In this paper, I use for σ_{iji} the empirically observed choices of players *i*. The justification is that both models are equilibrium models and hence require beliefs to coincide with actual behaviour.

Dufwenberg and Kirchsteiger (2004) define

$$k\left(s_j,\sigma_{iji}\right)=\pi_i\left(s_j,\sigma_{iji}\right)-\pi_i^e\left(\sigma_{iji}\right)$$

where $\pi_i\left(s_j,\sigma_{iji}\right)$ is *i*'s expected payoff from *j*'s strategy s_j given his second-order belief σ_{iji}, i.e., the payoff to himself *i* thinks *j intends* him to receive, and $\pi_i^e\left(\sigma_{iji}\right)$ the payoff to himself *i* views as "equitable" given σ_{iji}. It is defined by

$$\pi_i^e\left(\sigma_{iji}\right)=0.5\cdot\left(\max_{s_j\in S_j}\pi_i\left(s_j,\sigma_{iji}\right)+\min_{s_j\in S_j}\pi_i\left(s_j,\sigma_{iji}\right)\right).$$

This formulation slightly simplifies the original model of Dufwenberg and Kirchsteiger (2004), which is inconsequential in the examples considered here. The interpretation is that *i* feels neutral about s_j $\left(k=0\right)$ if he believes *j* intends him to receive half of what he maximally and minimally stands to earn given *j*'s strategy set S_j and his second-order belief σ_{iji} and feels s_j is (un)kind whenever he receives more (less), to which correspond $k>0$ $\left(k<0\right)$. A direct implication is that the kindness of s_j equals zero if *j* is passive. Player *i* responds to the kindness of s_j as follows: If s_j is (un)kind, he is willing to increase the (un)kindness of his own behaviour to *j* at some material payoff cost to himself.

The reciprocity model of Falk and Fischbacher (2006) differs from Dufwenberg and Kirchsteiger (2004) in that distributional concerns directly influence kindness perceptions. The

kindness of strategy $s_j \in S_j$ as perceived by player *i* is given by

$$k\left(s_j,\sigma_{iji}\right)=\left[\pi_i\left(s_j,\sigma_{iji}\right)-\pi_j\left(s_j,\sigma_{iji}\right)\right]\cdot\Delta\left(s_j,\sigma_{iji}\right).$$

The first term is called the "outcome term". It consists of the inequality associated with the outcome $\pi\left(s_j,\sigma_{iji}\right)$ implemented by s_j given *i*'s second-order belief. If $\pi\left(s_j,\sigma_{iji}\right)$ puts player *i* ahead of (behind) *j*, *i* tends to view s_j as (un)kind. At the same time, the outcome term does not reflect the alternatives to s_j that *j* has at her disposal. This is where the second term (the "intention factor") comes into play. It takes on either the value 1 or $\varepsilon \in [0,1]$. For example, if $\pi\left(s_j,\sigma_{iji}\right)$ puts *i* ahead of *j*, the intention factor equals 1 if the feasible set of outcomes given σ_{iji} contains a payoff to *i* smaller than $\pi_i\left(s_j,\sigma_{iji}\right)$ and ε otherwise. The idea is that in the first case *j* could have treated *i* worse than giving him $\pi_i\left(s_j,\sigma_{iji}\right)$, whereas no such option was available in the second case. As a result, *i* discounts his advantage $\pi_i\left(s_j,\sigma_{iji}\right)-\pi_j\left(s_j,\sigma_{iji}\right)>0$ in the second case, but not in the first. The procedure for the case where $\pi\left(s_j,\sigma_{iji}\right)$ puts *i* behind is analogous.

Appendix C: A General Model of Net-Loss Reciprocation

As before, I limit attention to finite-horizon multi-stage games. The set of players is *I* where $|I| = n \geq 3$ and one of the players is Nature (denoted *N*). A player's inactivity at a stage is modelled by the respective action set being singleton. At every stage, players have certainty about what happened at the previous stages, i.e., the non-terminal history up to that stage. Let *H* be the set of non-terminal histories, which contains the empty history (or root of the game) $\varnothing$. The *pure strategy* $s_i \in S_i$ of player $i \in I$ assigns to each history $h \in H$ an action $a_i \in A_i(h)$ available to *i* at *h*, whereas *i*'s *behaviour strategy* $\sigma_i \in \Sigma_i$ assigns to each history a probability distribution on *i*'s available actions.[30] The set of pure and behaviour strategy profiles are given by $S = \prod_{i \in I} S_i$ and $\Sigma = \prod_{i \in I} \Sigma_i$, respectively. Outcomes $\hat{\pi}$ of the game are $(n-1)$-dimensional vectors of material payoffs.[31] The function $\hat{\pi} : S \to \mathbb{R}^{n-1}$ is the *outcome function*. It assigns to each pure strategy profile the payoff vector implemented by it. From $\hat{\pi}(s)$, we can derive $\pi : \Sigma \to \mathbb{R}^{n-1}$, which assigns to each profile of behaviour strategies the implemented vector of expected payoffs. $\Pi = \{\pi(\sigma) : \sigma \in \Sigma\}$ contains all feasible expected outcomes.[32]

Moreover, a few non-standard concepts are drawn on below. Firstly, $S_i(h) \subseteq S_i$ is the set of strategies of *i* that are *consistent* with history $h \in H$ in the following sense: If $h \neq \varnothing$, s_i is part of $S_i(h)$ if and only if it prescribes *i*'s actions contained in h. If $h = \varnothing$, we have $S_i(h) = S_i$. The set $\Sigma_i(h)$ is defined analogously: All actions in h must be prescribed with probability one for $\sigma_i \in \Sigma_i(h)$. Secondly, $s_i(s_i, h) \in S_i$ is the "update" of s_i that coincides with s_i except that it prescribes *i*'s actions contained in h and likewise for $\sigma_i(\sigma_i, h)$ where the actions in h are prescribed with probability one. Finally, $H(s_i) \subseteq H$ is the set of histories that are consistent with s_i in the sense that any $h \in H$ with $h \neq \varnothing$ is in $H(s_i)$ if and only if *i*'s actions contained in *h* are also actions prescribed by s_i. Moreover, $\varnothing \in H(s_i)$ for all $s_i \in S_i$.

To illustrate the model, I draw on a simple delegation game, which is implemented experimentally in Bartling and Fischbacher (2011). The game has four players:[33] One principal

---

[30] All action sets are assumed finite.

[31] Outcomes specify a payoff for each player except Nature.

[32] Note that $S \subseteq \Sigma$ since all pure strategy profiles are degenerate behaviour strategy profiles. As a result, all pure strategy profiles are in the domain of π.

[33] Strictly speaking, Nature is the fifth player, who is passive in this example.

(player A), one agent (player B) and two passive recipients, one of whom is called player C. The principal moves first. She can either implement a fair or unfair outcome directly or delegate this choice to the agent. The fair outcome yields 5 units of payoff to all parties, while the unfair outcome gives 9 units of payoff each to the principal and agent and 1 unit to each recipient. To simplify things, the punishment opportunities of the recipients after A or A and B have made their choice are left out of the picture.

C1 Derived Net Losses

Player *i*'s net loss from strategy s_j of player *j* consists of his loss minus gain from s_j where the two need not count for the same. Player *i* assesses his loss and gain by comparing the *opportunity set* of outcomes created by s_j to the opportunity sets created by *j*'s alternative strategies. In the example, suppose that C evaluates A's decision to delegate. He then compares the opportunity set created by A delegating to the opportunity sets created by A choosing the fair and unfair outcome directly.

Defining such opportunity sets raises several modelling issues. Firstly, the question arises which (if any) restrictions to place on the behaviour of third parties, i.e., on the other players besides *i* and *j*. A's decision to delegate is a case in point as its consequences depend on the behaviour of B, who is the third party in the relationship between A and C. In what follows, I assume that *i* considers the opportunity sets of *expected outcomes* created by s_j and its alternatives taking as given $\sigma_{-i,j} \in \prod_{k \in I \setminus \{i,j\}} \Sigma_k$, which is the profile of behaviour strategies of all other players including Nature and can be interpreted as *i*'s belief about these players' average behaviour. The idea is that *i*, when assessing the opportunity sets created for him by s_j and its alternatives, has some sense of how third parties are likely to act, which affects his sense of opportunity.[34] In the delegation example, B is expected to choose the unfair outcome after delegation with a probability of 0.34 according to the beliefs measured by Bartling and Fischbacher (2011). The opportunity set of expected outcomes created by delegation is therefore

[34] Alternatively, $\sigma_{-i,j}$ could be interpreted as *i*'s belief about *j*'s belief about the other players, i.e., *i*'s second-order belief.

$$\{(0.34\cdot 9+0.66\cdot 5,0.34\cdot 9+0.66\cdot 5,0.34\cdot 1+0.66\cdot 5,0.34\cdot 1+0.66\cdot 5)\}=\{(6.36,6.36,3.64,3.64)\},$$

while A choosing the fair and unfair outcome directly entail $\{(5,5,5,5)\}$ and $\{(9,9,1,1)\}$, respectively.[35] These opportunity sets are singleton because the punishment options are ignored.

A second issue is which perspective *i* adopts when assessing s_j and its alternatives given $\sigma_{-i,j}$. On the one hand, *i* could evaluate s_j from an *ex-ante* perspective meaning that he simply compares the opportunity set of expected outcomes created by s_j to the opportunity sets created by its alternatives taking as given $\sigma_{-i,j}$. On the other hand, he could adopt the perspective of some history $h\neq\varnothing$ of the game and evaluate *j*'s updated strategy $s_j(s_j,h)$ against its alternatives in $S_j(h)$ taking as given $\sigma_{-i,j}(\sigma_{-i,j},h)$. The importance of such conditioning on histories can be seen by considering a second example, namely, C's evaluation of B choosing the unfair outcome after delegation, the alternative being choosing the fair outcome. From an *ex-ante* perspective, the opportunity sets of expected outcomes created by these two strategies depend on the belief about A, i.e., about how likely A is to delegate in the first place. In the most extreme scenario, where A is not believed ever to delegate, the two opportunity sets would be the same implying C's net loss from the two strategies is the same, namely, zero. This dependence of the evaluation of B on the beliefs about A seems implausible. Intuitively, given the information structure, C knows that B knows that A has delegated when B chooses the unfair outcome and C wants to hold B to account for this knowledge. This consideration can be captured if we condition on the history "A has delegated", which means that the likelihood of delegation is set to one. For this reason, I posit that *i* when evaluating s_j adopts the perspective of all histories in the set $H(s_j)$, which is the set of histories consistent with s_j. The idea is that *i* restricts attention to histories not ruled out by s_j, which has intuitive appeal.[36] Consequently, I first define *i*'s loss and gain from s_j for a given $h\in H(s_j)$ and then define *i*'s overall loss and gain as his maximal history-contingent loss and gain with respect to $H(s_j)$ as a whole.

---

[35] The first entry in payoff vectors refers to the payoff of A, the second to the payoff of B and the last two to the payoffs of the recipients.

[36] Limiting the conditioning to $H(s_j)$ is also required for making the definition of *i*'s loss and gain consistent with the definition given above for a two-player setting, where there is no conditioning on histories ruled out by s_j. Consistency means that in any *n*-player game where the players in $I\setminus\{i,j\}$ are passive, *i*'s loss and gain from any given s_j is the same as his loss and gain from the corresponding s_j in the corresponding two-player game where the players in $I\setminus\{i,j\}$ are omitted.

Finally, when adopting the perspective of some $h \in H(s_j)$, the question arises if we should restrict *i*'s own behaviour to be in $S_i(h)$ in establishing the opportunity sets created by s_j and its alternatives. Limiting *i*'s behaviour to $S_i(h)$ appears not fully convincing because *i* aims to assess the "elbow room" left for him by s_j and its alternatives. From this angle, restricting attention to $S_i(h)$ seems misguided. Intuitively, *i* holds *j* responsible for choosing s_j rather than its alternatives in $S_j(h)$ given that third parties behave according to $\sigma_{-i,j}(\sigma_{-i,j}, h)$, but does not hold *j* responsible for ending up in h. Any part that *j* has played in bringing about h is dealt with by considering the rest of $H(s_j)$. Consequently, I define the opportunity set created by s_j from the perspective of h as $\{\pi(s_i, s_j, \sigma_{-i,j}(\sigma_{-i,j}, h)) : s_i \in S_i\}$ and likewise for the alternatives.

I now define player *i*'s *loss* and *gain* derived from $s_j \in S_j$ (the "status quo") relative to some alternative $\tilde{s}_j \in S_j$ without conditioning on histories. The sets Π^{s_j} and $\Pi^{\tilde{s}_j}$ are the opportunity sets of expected outcomes created by the two strategies. At this point, I only assume them to be non-empty without worrying about their precise definition, which is history-dependent and introduced later. I begin with *losses*. Analogous to before, let $\Pi_i^{s_j}$ be the expected payoffs to *i* in Π^{s_j}, $\overline{\pi}_i^{s_j} = max\,\Pi_i^{s_j}$ his maximal payoff given s_j and $\Pi^{\tilde{s}_j > s_j} = \{\pi \in \Pi^{\tilde{s}_j} : \pi_i > \overline{\pi}_i^{s_j}\}$ the set of feasible outcomes given $\tilde{s}_j$ yielding *i* more payoff than $\overline{\pi}_i^{s_j}$. Fairness is measured by a fairness function:

DEFINITION C1 The *fairness function* $f : \mathbb{R}^{n-1} \to \mathbb{R}$ is given by

$$f(\pi) = \underline{\pi} + \alpha \sum_{i \in I \setminus \{N\}} (\pi_i - \underline{\pi}) / (n-1)$$

with $\underline{\pi} = min\{\pi_i : i \in I \setminus \{N\}\}$ and $\alpha \in [0,1]$.

We can now define *i*'s loss from s_j relative to $\tilde{s}_j$, which is isomorphic to the two-player case.

DEFINITION C2 Player *i*'s *loss* from strategy $s_j \in S_j$ relative to strategy $\tilde{s}_j \in S_j$ is given by

$$l_i\left(s_j,\tilde{s}_j\right)=\begin{cases} max_{\pi\in\Pi^{\tilde{s}_j>s_j}}\left[\beta\left(\pi_i-\overline{\pi}_i^{s_j}\right)+\left(1-\beta\right)max\left\{f\left(\pi\right)-\overline{f}^{s_j},0\right\}\right] & \text{if } \Pi^{\tilde{s}_j>s_j}\neq\varnothing \\ 0 & \text{otherwise} \end{cases}$$

where $\beta\in[0,1]$.

I next turn to player *i*'s *gain* from s_j relative to $\tilde{s}_j$, which is again isomorphic to the two-player case.

DEFINITION C3 Player *i*'s gain from strategy $s_j\in S_j$ relative to strategy $\tilde{s}_j\in S_j$ is given by

$$g_i\left(s_j,\tilde{s}_j\right)=\begin{cases} max_{\pi\in\Pi^{s_j>\tilde{s}_j}}\left[\beta\left(\pi_i-\overline{\pi}_i^{\tilde{s}_j}\right)+\left(1-\beta\right)max\left\{\overline{f}^{\tilde{s}_j}-f\left(\pi\right),0\right\}\right] & \text{if } \Pi^{s_j>\tilde{s}_j}\neq\varnothing \\ 0 & \text{otherwise} \end{cases}$$

where $\beta\in[0,1]$.

I now address *i*'s loss and gain from s_j at large. The two are established by considering all histories consistent with s_j, which are collected in $H\left(s_j\right)$.

DEFINITION C4 From the perspective of history $h\in H\left(s_j\right)$, player *i*'s *loss* and *gain* from $s_j\in S_j$ given that the players in $I\setminus\{i,j\}$ follow $\sigma_{-i,j}\in\Sigma_{-i,j}$ are given by

$l_i\left(s_j,h\right)=max_{\tilde{s}_j\in S_j(h)}\,l_i\left(s_j,\tilde{s}_j\right)$ and $g_i\left(s_j,h\right)=max_{\tilde{s}_j\in S_j(h)}\,g_i\left(s_j,\tilde{s}_j\right)$, respectively,

where $\Pi^{s_j}=\left\{\pi\left(s_i,s_j,\sigma_{-i,j}\left(\sigma_{-i,j},h\right)\right):s_i\in S_i\right\}$ and likewise for each $\Pi^{\tilde{s}_j}$. Moreover, *i*'s overall loss and gain from s_j are given by

$l_i\left(s_j\right)=max_{h\in H\left(s_j\right)}\,l_i\left(s_j,h\right)$ and $g_i\left(s_j\right)=max_{h\in H\left(s_j\right)}\,g_i\left(s_j,h\right)$.

Thus, given $\sigma_{-i,j}$, *i* assesses his loss and gain from s_j history-wise by considering each element in $H\left(s_j\right)$. Adopting the perspective of some such history, *i* determines his maximal loss and

gain from s_j relative to its alternatives in $S_j(h)$ taking as given $\sigma_{-i,j}(\sigma_{-i,j},h)$.[37] Player *i*'s overall loss and gain from s_j are given by his maximal history-contingent loss and gain with respect to $H(s_j)$.

As before, players react to the *net loss* imposed on them by others:

DEFINITION C5 Player *i*'s *net loss* from strategy $s_j \in S_j$ is given by

$$nl_i(s_j) = l_i(s_j) - \gamma g_i(s_j)$$

with $\gamma \in [0,1]$.

C2 Imposed Net Losses and Preferences

Faced with some profile s_{-i} of the other players' strategies, player *i* must choose an outcome from his opportunity set $\Pi^{s_{-i}} = \{\pi(s_i, s_{-i}) : s_i \in S_i\}$. As before, let $\pi^c \in \Pi^{s_{-i}}$ be the outcome chosen by *i* and let $\Pi^{s_{-i}>c,j} = \{\pi \in \Pi^{s_{-i}} : \pi_j > \pi_j^c\}$ and $\Pi^{s_{-i}<c,j} = \{\pi \in \Pi^{s_{-i}} : \pi_j < \pi_j^c\}$ contain the outcomes in $\Pi^{s_{-i}}$ yielding player *j* more and less payoff than π^c, respectively. This leads to

DEFINITION C6 Player *j*'s *loss* from $\pi^c \in \Pi^{s_{-i}}$ is

$$l_j(\pi^c, \Pi^{s_{-i}}) = \begin{cases} max_{\pi \in \Pi^{s_{-i}>c,j}} \left[\beta(\pi_j - \pi_j^c) + (1-\beta) max\{f(\pi) - f(\pi^c), 0\} \right] & \text{if } \Pi^{s_{-i}>c,j} \neq \varnothing \\ 0 & \text{otherwise} \end{cases}$$

Moreover, *j*'s *gain* from $\pi^c \in \Pi^{s_{-i}}$ is

$$g_j(\pi^c, \Pi^{s_{-i}}) = \begin{cases} max_{\pi \in \Pi^{s_{-i}<c,j}} \left[\beta(\pi_j^c - \pi_j) + (1-\beta) max\{f(\pi) - f(\pi^c), 0\} \right] & \text{if } \Pi^{s_{-i}<c,j} \neq \varnothing \\ 0 & \text{otherwise} \end{cases}$$

Finally, *j*'s *net loss* from $\pi^c \in \Pi^{s_{-i}}$ is

[37] $S_j(h)$ is guaranteed to include s_j since *h* is taken from $H(s_j)$.

$nl_j\left(\pi^c,\Pi^{s_{-i}}\right)=l_j\left(\pi^c,\Pi^{s_{-i}}\right)-\gamma g_j\left(\pi^c,\Pi^{s_{-i}}\right)$.

As for *i*'s preferences governing his choice from $\Pi^{s_{-i}}$, they are given by

ASSUMPTION C1 Player *i*'s preferences on the outcomes in $\Pi^{s_{-i}}$ are represented by

$$u_i\left(\pi,s_{-i}\right)=v\left(\pi_i\right)+\sum\nolimits_{j\in I\setminus\{i,N\}} r\left(nl_j\left(\pi,\Pi^{s_{-i}}\right),nl_i\left(s_j\right)\right)$$

where the continuous $v:\mathbb{R}\to\mathbb{R}$ and $r:\mathbb{R}^2\to\mathbb{R}$ satisfy $dv/d\pi_i>0$ and $\partial^2 r/\partial nl_j\partial nl_i>0$ for every $j\in I\setminus\{i,N\}$.

As a result, $WTP=\partial r/\partial nl_j\,/\,dv/d\pi_i$, which is *i*'s willingness to pay for increasing any other player *j*'s net loss, increases in the net loss that *i* himself derives from s_j.

References

Battigalli, Pierpaolo and Martin Dufwenberg (2009). "Dynamic psychological games". *Journal of Economic Theory*, 144(1), 1-35.

Bartling, Björn and Urs Fischbacher (2011). "Shifting the Blame. On Delegation and Responsibility". *Review of Economic Studies*, 79(1), 67-87.

Brandts, Jordi and Carles Solá (2001). "Reference Points and Negative Reciprocity in Simple Sequential Games". *Games and Economic Behavior*, 36(2), 138-157.

Charness, Gary and Matthew Rabin (2002). "Understanding Social Preferences With Simple Tests". *The Quarterly Journal of Economics*, 117(3), 817-869.

Cox, James C. (2004). "How to identify trust and reciprocity". *Games and Economic Behavior*, 46(2), 260-281.

Cox, James C., Friedman, Daniel and Steven Gjerstad (2007). "A tractable model of reciprocity and fairness". *Games and Economic Behavior*, 59(1), 17-45.

Cox, James C., Daniel Friedman and Vjollca Sadiraj (2008). "Revealed Altruism". *Econometrica*, 76(1), 31-69.

Cox, James C., Maroš Servátka and Radovan Vadovič (2010). "Saliency of outside options in the lost wallet game". *Experimental Economics*, 13(1), 66-74.

Dhaene, Geert and Jan Bouckaert (2010). "Sequential reciprocity in two-player, two-stage games. An experimental analysis". *Games and Economic Behavior*, 70(2), 289-303.

Dufwenberg, Martin and Uri Gneezy (2000). "Measuring beliefs in an experimental lost wallet game". *Games and Economic Behavior*, 30(2), 163-182.

Dufwenberg, Martin and Georg Kirchsteiger (2004). “A theory of sequential reciprocity”. *Games and Economic Behavior*, 47(2), 268-298.

Falk, Armin, Ernst Fehr and Urs Fischbacher (2003). “On the nature of fair behavior”. *Economic Inquiry*, 41(1), 20-26.

Falk, Armin and Urs Fischbacher (2006). “A theory of reciprocity”. *Games and Economic Behavior*, 54(2), 293-315.

Falk, Armin and Michael Kosfeld (2006). “The hidden costs of control”. *The American Economic Review*, 1611-1630.

Fehr, Ernst and Klaus M. Schmidt (1999). “A theory of fairness, competition, and cooperation”. *The Quarterly Journal of Economics*, 114(3), 817-868.

Gächter, Simon and Ernst Fehr (2000). “Cooperation and punishment in public goods experiments”. *The American Economic Review*, 90(4), 980-994.

Geanakoplos, John, David Pearce and Ennio Stacchetti (1989). “Psychological games and sequential rationality”. *Games and Economic Behavior*, 1(1), 60-79.

Goeree, Jacob K., Charles A. Holt and Thomas R. Palfrey (2008). "Quantal response equilibria". In Steven N. Durlauf and Lawrence E. Blume ed.s., *The New Palgrave Dictionary of Economics (Second Edition)*, Palgrave Macmillan.

Kahneman, Daniel and Amos Tversky (1979). “Prospect theory. An analysis of decision under risk”. *Econometrica*, 47(2), 263-291.

Köszegi, Botond and Matthew Rabin (2006). “A model of reference-dependent preferences”. *The Quarterly Journal of Economics*, 121(4), 1133-1165.

McFadden, Daniel (1974). “Conditional logit analysis of qualitative choice behavior”. In P. Zarembka ed., *Frontiers in Econometrics*, Academic Press, New York, 105-142.

McCabe, Kevin A., Mary L. Rigdon and Vernon L. Smith (2003). “Positive reciprocity and intentions in trust games”. *Journal of Economic Behavior & Organization*, 52(2), 267-275.

McKelvey, Richard D. and Thomas R. Palfrey (1995). “Quantal response equilibria for normal form games”. *Games and Economic Behavior*, 10(1), 6-38.

McKelvey, Richard D. and Thomas R. Palfrey (1996). “A Statistical Theory of Equilibrium in Games”. *Japanese Economic Review*, 47(2), 186-209.

Rabin, Matthew (1993). “Incorporating fairness into game theory and economics”. *The American Economic Review*, 83(5), 1281-1302.

Servátka, Maroš and Radovan Vadovic (2009). “Unequal Outside Options in the Lost Wallet Game”. University of Canterbury Working Paper.

Shalev, Jonathan (2000). “Loss aversion equilibrium”. *International Journal of Game Theory*, 29(2), 269-287.

Von Siemens, Ferdinand (2011). “Intention-Based Reciprocity and the Hidden Costs of Control”. Tinbergen Institute Discussion Paper.