A Theory of Reciprocity with Trust

Marcus Roel*

m.roel@lse.ac.uk

London School of Economics

July 16, 2018

Abstract

People are reciprocal if they reward kind actions and punish unkind ones. I propose a new theory of intention-based reciprocity that addresses the question of when a mutually beneficial action is kind. When both benefit from the action, a player's motive is unclear: he may be perceived as kind for improving the other player's payoff, or as self-interested and not-kind for improving his own. I use trust as an intuitive mechanism to solve this ambiguity. Whenever a player puts himself in a vulnerable position by taking such an action, he can be perceived as kind. In contrast, if this action makes him better off than his alternative actions do, even if it is met by the most selfish response, he cannot be kind. My model explains why papers in the literature fail to find (much) positive reciprocity when players can reward and punish. In particular, I show how negative reciprocity crowds out positive reciprocity. By allowing for interactions between rewards and punishments, my model provides a theoretical framework to analyze institutional design and incentive structures when people are motivated by reciprocity.

Keywords: Reciprocity, Intentions, Trust, Social Preferences, Psychological Games

JEL Codes: A13, D02, C70, D63

*I first and foremost thank Erik Eyster who spent too many late Friday afternoons providing guidance, support, and constructive criticism. I would also like to thank Francesco Nava, Nava Ashraf, Björn Bartlin, Andrew Ellis, Rafael Hortala-Vallve, Gilat Levy, Matthew Levy, Nick Netzer, Ronny Razin, Armin Schmutzler, Balazs Szentes, Séverine Toussaert, Roberto Weber and seminar participants at the London School of Economics for helpful comments and suggestions.

1. Introduction

People are willing to sacrifice their own material wellbeing to reward those who are *kind* (positive reciprocity), and to punish those who are *unkind* (negative reciprocity). This deviation from pure selfishness has important economic consequences. After a pleasant dinner with great service, we leave a generous tip for the waiter even when we don't expect to return again. We are more likely to donate to charity when solicitation letters include gifts (Falk (2007)), i.e. we respond to gifts with counter-gifts. Akerlof (1982) argues that this idea of *gift exchange* may explain involuntary unemployment in the labor market: when workers respond to generous wage offers by working harder, firms are incentivized to raise wages above the market clearing wage. Fehr et al. (1993) and Fehr and Falk (1999) demonstrate this experimentally. Bewley (1995) provides field evidence for this view in the form of a large interview study. Employers cite worries about lower morale (and thus lower effort) as the reason for not cutting wages in recessions. Reciprocity may also give rise to acts of sabotage when workers punish unfair or unkind behavior. For example, Giacalone and Greenberg (1997) report a rise in employees' theft rates after wage cuts. Krueger and Mas (2004) find that tires produced at the Bridgestone-Firestone plant were ten times more likely to be defective as a result of a three-year labor dispute.

All these raise the fundamental question as to what constitutes kind and unkind behavior. Studies highlight that the perception of what is kind (or fair) is not only determined by distributional concerns, e.g. inequity-aversion (Fehr and Schmidt (1999)), but also by *how* this payoff distribution comes about. For example, a one-sided offer is perceived as less unkind if the only alternative is even more one-sided (Falk and Fischbacher (2006)) and is thus rejected less often (Falk et al. (2003)). This underlines that people consider the intentions and motives behind other people's actions and not just the respective outcomes that these actions induce.

In his seminal paper, Rabin (1993) (henceforth Rabin) formalizes intention-based reciprocity for normal form games and suggests a definition of kindness. Dufwenberg and Kirchsteiger (2004) (henceforth DK04) extend intention-based reciprocity to sequential games. In these models, players form beliefs about the intentions behind the other players' actions. For instance, upon receiving a gift, the receiver forms beliefs about whether the giver expects a gift in return or not. He then evaluates the giver's kindness based on these beliefs. An action is perceived as kind (unkind) if it yields an intended payoff that is larger (smaller) than a reference point. The reference point and thus kindness perceptions differ slightly between the two papers, however.[1] In Rabin, an action is only kind if it

[1] See also Netzer and Schmutzler (2014), who apply Rabin's notion of kindness to a sequential gift-exchange.

comes at a personal cost, whereas in DK04, a mutually beneficial action, i.e. an action that improves both players' payoffs, can be kind. As a result, a gift that also benefits the gift-giver, for example due to an expected counter-gift, is only kind in DK04.

In this paper, I revisit the central issue of when a mutually beneficial action is kind and provide a new definition of kindness. When an action is perceived to be mutually beneficial, a player's motive is unclear: He may be perceived as kind for improving the other player's payoff, or as selfish for improving his own. The concept of *trust* offers a psychologically intuitive mechanism to solve this ambiguity: Whenever the player puts himself in a vulnerable position by taking such an action, he is perceived as kind. In contrast, if his action makes him better off than the alternative, even if it is met by the most selfish response, he cannot be kind. Since players are only willing to reward kind actions, this distinction helps to explain why some papers in the literature do not find much positive reciprocity. It also offers new insights into the interaction of rewarding and punishing actions.

	Simultaneous choice		Sequential choice				
Study	cooperation rate	N	player 1	player 2 after C	player 2 after D	N	Strategy Method
Khadjavi and Lange, 2013 (students)	37.0%	36	63.0%	62.1%	0.0%	46	no
Khadjavi and Lange, 2013 (prisoners)	55.6%	46	46.3%	60.0%	3.4%	54	no
Ahn et al., 2007	32.5%	80	30.0%	35.0%	5.0%	80	yes
Bolle and Ockenfels, 1990	18.6%	59	17.3%	19.7%	4.9%	122	yes
Hayashi et al., 1999	36.0%	50	56.3%	61.1%	0.0%	63	no
Watabe et al, 1996	55.6%	27	82.6%	75.0%	12.0%	68	no
Cho and Choi, 2000	47.5%	59	52.4%	72.7%	0.0%	42	no
Average	40.4%		49.7%	55.1%	3.6%		

Figure 1: Choice data from prisoner's dilemmas

Figure 1 lists all studies that cover both the simultaneous and sequential prisoner's dilemma.[2] The data highlights that cooperation in a prisoner's dilemma is not unconditional and hence cannot be explained by a simple model of altruism: in the sequential prisoner's dilemma hardly any player 2 cooperates after defection. The sequential prisoner's dilemma, see also game 1 on page 4, is an example of a social dilemma, in which the first player's (he) action may improve both his own and the second player's payoff if the second player (she) positively reciprocates. In four out of six studies, it is empirically payoff-maximizing for player 1 to cooperate. The data, hence, suggests that player 2 views player 1's cooperative choice as kind even if it is in player 1's best interest. Malmendier and Schmidt (2017) observe a similar behavior in response to gifts. In their experiment, most participants are aware that the gift was intended to influence their behavior; yet they still positively reciprocate. Malmendier

[2] All studies are one-shot interactions, incentivized, and feature a participant for each role. I did not include studies that use deception, i.e. do not have an actual player 1, or that are not incentivized.

and Schmidt argue that players feel obligated to reciprocate. Finally, figure 1 also indicates that there is more cooperation in the sequential prisoner's dilemma than in the simultaneous one.[3]

In my model, player 2 perceives cooperation as kind even if she believes that player 1 expects her to cooperate in response (second order belief). Tempted to defect, player 2 wonders 'what if I take advantage of him?' If she defects, player 1 would be better off had he defected himself, and therefore exposed vulnerability by cooperating. Player 2 perceives his choice as trusting, concludes that player 1's action is kind, which in turn motivates her to cooperate. To determine whether a mutually beneficial action is kind, player 2 asks the simple question 'is it trusting?'

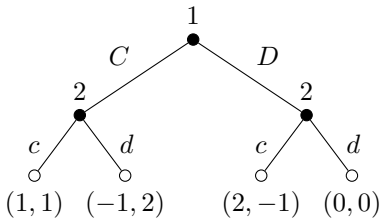
DK04, instead, suggest that a mutually beneficial action is kind as long as there exists *a strategy for player 2* for which player 1 is better off by taking the alternative choice.[4] It follows that only a mutually beneficial action that is also player 1's (payoff) dominant action cannot be considered kind. While DK04 predict the same behavior in the prisoner's dilemma as I do, this is not true in general: In my model, actions tend to be perceived as less kind than in DK04, giving rise to less positive reciprocity. Not only does this explain why some papers in the literature do not find much positive reciprocity, it also provides new insights into the interaction of punishing and rewarding actions.

This is best illustrated by Orhun (2018). She is interested in how player 2 responds to cooperation in a prisoner's dilemma when player 2's available choices after defection vary. In particular, she compares behavior in the usual prisoner's dilemma (game 1) to behavior in a prisoner's dilemma with punishment (game 2). In the latter, player 2 has the option to punish player 1 after he defects. Orhun finds that the availability of the option to punish significantly alters the players' perception of the game. On average, player 1 believes that in 41% of the time player 2 punishes after D, and player 2 holds a second order belief that he thinks she punishes in 54% of all cases. For these beliefs, cooperation maximizes player 1's payoff even if player 2 defects after C.[5] Relatively to the sequential prisoner's dilemma, player 2's cooperation rate after C drops by 22 percentage points in the one with punishment.

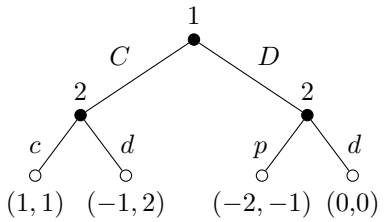
[3] While this can be a result of player 2's knowledge of 1's action (in contrast to expecting cooperation from player 1, player 2 observes his action), it provides further evidence that player 2 is willing to reward the mutually beneficial actions: In the simultaneous game, cooperation is, if anything, kinder. By cooperating in the simultaneous game, player 1 improves player 2's payoff more than in the sequential game, and does so at his own expense. Despite these two forces, the sequential games features more cooperation. Note that in a simultaneous version of game 1, player 1 increases 2's payoff by 2 units when he cooperates, regardless of 2's choice. Given that player 2 defects after defection, he only increases 2's payoff by 2 when she defects after C. When she cooperates after C, he only increases her payoff by 1.

[4] In the prisoner's dilemma, this is satisfied for the strategies 'always cooperate', 'always defect', and 'defect after cooperation and cooperate after defection'.

[5] The actual payoffs in Orhun's experiment differ slightly from those in the figures.



Game 1: Sequential prisoner's dilemma



Game 2: Prisoner's dilemma with punishment

My model, as far as I am aware, is the only model that predicts full conditional cooperation in the prisoner's dilemma, and defection (after C) and punishment (after D) in the prisoner's dilemma with punishment. Similar to the ultimatum game, player 2 punishes in response to the unkind action D, which leads player 1 to cooperate. While cooperation improves 2's payoff, it is not trusting: player 1 is better off even if player 2 defects in response. As a result, she perceives C as not kind and defects. In DK04, C is always seen as kind, since there is a strategy (always defect) for which D is optimal for player 1. If anything, their model predicts more, not less positive reciprocity in the prisoner's dilemma with punishment as players tend to punish, lowering their own alternative payoff.

This example highlights how kindness perceptions are not simply affected by the set of available choices for player 1, but also by how player 2 responds to these alternative. My model explains why some papers fail to find much positive reciprocity, i.e. Offerman (2002), Al-Ubaydli and Lee (2009) and Orhun (2018). Since the standard intention-based reciprocity model of Dufwenberg and Kirchsteiger (2004) predicts positive reciprocity in such games, Offerman's (2002) paper was often used (in combination with other papers) to argue that negative reciprocity is stronger than positive reciprocity. My model highlights that this need not be true. It shows how negative reciprocity can instead crowd out positive reciprocity. An action can be perceived very differently when the alternative action is followed by punishment rather than by a selfish response.

By allowing for complex interactions between punishment and rewards, my model provides a theoretical framework to analyze institutional design and incentive structures when people are motivated by reciprocity. It explains, for example, the lower demand for rewards in Andreoni et al. (2003) when players can punish, and how employees reduce their efforts when employers impose fines for shirking, Fehr and Gächter (2001).

In a companion paper, Roel (2018b), I extend my model to incomplete information and thereby provide a general framework to analyze sequential information sharing, sequential bilateral exchange, etc. With reciprocity preferences, I show that sequential mechanisms can be more efficient than simultaneous ones, Bierbrauer and Netzer (2016).

The rest of this paper is organized as follows: In the next section, I discuss the related literature. In section 3, I present my model and show that an equilibrium exists. Sections 4 and 5 characterize the differences of my model to competing theories of Rabin and DK04. The key difference to Rabin is the addition of trust. When an action is perceived as kinder in my model than in Rabin, the action is trusting. This allows players to reciprocate mutually beneficial actions. In contrast to fundamental preferences for trust, a trusting action can be unkind. In such cases, trust is predicted to be betrayed. The comparison with DK04 highlights how negative reciprocity can crowd out positive reciprocity. It also suggests yet-to-be-explored games, in which DK04's prediction of positive reciprocity appears implausible. In section 6, I revisit a variety of experimental papers and show how my model describes behavior in games, where players can reward and punish, better. Equilibrium predictions across all models are summarized in section 7. The paper ends with concluding remarks, section 8. Proofs, as well as the mathematical detail for most examples can be found in Appendix A.

2. Literature Review

Intention-based reciprocity models are built on the general framework of psychological games, Geanakoplos et al. (1989). Psychological games allow for beliefs to directly affect utility, and not just indirectly through expectation formation. In Rabin (1993), a player uses her belief about the other's action, as well her second order beliefs about her own, to asses whether that person intends to help or hurt her, whether he is kind or not. This directly affects her preferences. Dufwenberg and Kirchsteiger (2004) extend intention based-reciprocity to sequential games. Their key observation is that a player needs to update her beliefs about how kind the other player is as the game progresses. As discussed in the introduction, a second, possibly more crucial, difference to Rabin is their definition of the reference point. In comparison to Rabin's original model, or more recent versions that apply his reference point to sequential games (allowing for updating of beliefs, Netzer and Schmutzler (2014) and Le Quement and Patel (2017)), actions are perceived as kinder in DK04. The kinder an action, the more a player is willing to sacrifice own material gains to help him. As a consequence, DK04 predicts more positive reciprocity than models in the spirit of Rabin.[6]

Reciprocity models have been very successful in explaining non-selfish behavior in the laboratory. Participants generally reward trust, Berg et al. (1995), choose high levels of costly effort in response to above market wage offers, Fehr et al. (1993), and reject low offers in the ultimatum game, Güth et al. (1982). Studies also highlight the importance of intentions in motivating non-selfish actions.

[6] A formal description of each reference point can be found in section 4 and 5.

For example, Blount (1995) compares rejection rates in a normal ultimatum game, to one where the offer is made by a random number generator. In stark contrast to an offer made by a human subject, almost all zero-offers are accepted when they are chosen at random. Similarly, Falk et al. (2003) show that in an ultimatum game rejection rates for the same offer differ systematically with the availability of alternative offers. For instance, a split of $(8, 2)$ is rejected more frequently when player 1 could have chosen $(5, 5)$, than when his only other alternative is $(10, 0)$. Falk and Fischbacher (2006) report subjective kindness perceptions of such divisions. Player 2 perceives $(8, 2)$ as very unkind if $(5, 5)$ and $(2, 8)$ are alternatives, but much less unkind if the only other alternatives are $(9, 1)$ and $(10, 0)$. McCabe et al. (2003) vary player 1's choice set in a binary trust game. 65% of responders repay trust if player 1 has a choice between trusting and not trusting, while only 33% return money if the alternative 'not to trust' is eliminated. These studies highlight that people consider the intentions and motives behind other people's actions; they are not motivated by preferences over relative payoffs alone (Fehr and Schmidt (1999)). A zero-offer is not unkind if it is chosen at random; perceptions of a seemingly selfish or generous actions depend the set of alternative actions.

Theorists have developed a variety of other reciprocity models. Instead of modelling kindness in terms of absolute payoffs, it can also be modelled through relative payoffs between agents. This is done by Falk and Fischbacher (2006). A recent paper in the literature is Celen et al. (2017), who propose a novel definition of kindness based on the notion of blame. Here a player puts himself in the other's position and wonders if she would take an action that is worse or nicer, and blames him if the latter is true. Examples of models that do not rely on psychological games are Cox et al. (2007), Cox et al. (2008), Charness and Rabin (2002), Levine (1998), and Gul and Pesendorfer (2016), among others. For a good summary see Sobel (2005). These models, like mine, focus on the internal preferences for reciprocity. As such, they fail to account for social pressure or social image concerns, and, for example, cannot explain why people avoid, at a cost, situations in which they are asked to share (Dana et al. (2006), Dana et al. (2007)). Malmendier et al. (2014) discuss how these issues apply to reciprocity.

By introducing the idea of trust to reciprocity, my model is related to the trust literature, Berg et al. (1995), Cox (2004), etc. I will show how it relates to Cox et al. (2016), who define trust for general two-stage games. While trust and kindness often coincide in games, they are different concepts since they primarily focus on different peoples' payoffs. As a result, an action can be trusting but also unkind.

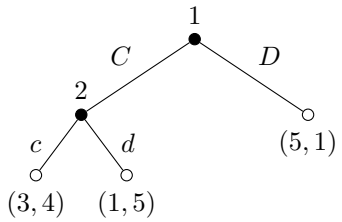
3. The Model

Basic idea of reciprocity. Reciprocity models allow for utility to depend on one's own as well on another player's payoff. In a two player game, player 2's utility takes the simple form of

$$U_2(\cdot) = \underbrace{\pi_2(\cdot)}_{\text{own payoff}} + \underbrace{\kappa_1(\cdot) \times \pi_1(\cdot)}_{\text{utility from reciprocity}}$$

In contrast to models of altruism, $\kappa_1 > 0$, or spite, $\kappa_1 < 0$, κ_1 varies with player 2's perception of how kind 1 is towards her. Player 1's kindness, as perceived by player 2, is defined as an expression that compares 2's perceived payoff against a reference point π_2^r. When the payoff is larger than the reference point, we say player 1 is kind ($\kappa_1 > 0$), when it is lower, we say he is unkind ($\kappa_1 < 0$).

Game 3 captures a simple scenario in which player 1 can improve player 2's payoff at his own cost. Suppose that 2's reference point is her lowest payoff in the game, $\pi_2^r = 1$. In this case, player 2 will perceive action C as kind since it makes her strictly better off than the reference point. Apart from answering the simple question 'is 1 kind?' when 2 observes C, she may also wonder 'how kind is he'. Since her payoff is either 4 or 5, she needs some criterion to decide between the two. In intention-based reciprocity models, she uses her beliefs about what player 1 think she would do after C. These beliefs are called second order beliefs and capture the intended consequences after player 1's action. For example if she believes 1 thinks she cooperates, she perceives the kindness of action C as $\kappa_1 = 4-1 = 3$; kindness is simply the difference between the payoff and the reference point. If instead she believes that 1 believes she defects, she would perceive him to be more kind, $\kappa_1 = 5-1 = 4$. Since she prefers to cooperate for the lowest kindness perception, $U_2(c) = 4 + (3) \cdot 3 \geq 5 + (3) \cdot 1 = U_2(d)$, she will positively reciprocate either way.



Game 3: A simple example of kindness

I now proceed to introducing the formal notation and kindness definitions.[7] In contrast to the

[7] I opt for a notation that is closer to DK04 than the more recent, general framework of Battigalli and Dufwenberg (2009), who extend psychological games to allow for updated higher-order beliefs, beliefs of others, and plans of actions to directly affect utility. This has the advantage that differences between models are more explicit, and that a more familiar equilibrium notion is used.

previous example, the reference point represents a statistic on a subset of payoffs, and not necessarily all payoffs in the game.

Game. Let the game be a 2-player, finite, multi-stage game, with perfect information and finite actions. Hence, choices occur sequentially and are fully observed.[8]

Players, actions, and strategies. Let $N = \{1, 2\}$ be the set of players, and H be the set of all non-terminal histories. Terminal histories are denoted by Z. $A_{i,h}$ describes the (possibly empty) set of actions for player $i \in N$ at node $h \in H$. A history of length l is a sequence $h = (a^1, a^2, \ldots, a^l)$, where $a^t = (a_1^t, a_2^t)$ is a profile of actions chosen at time t $(1 \leq t \leq l)$. Player i's behavior strategy is denoted by $\sigma_i \in \times_{h \in H} \Delta(A_{i,h}) =: \Delta_i^H$. It assigns at each node $h \in H$ a probability distribution $\sigma_i(\cdot|h)$ over the set of pure actions. Define $\Delta^H = \prod_{i \in N} \Delta_i^H \ni \sigma$.

Player i's *material payoff* is defined as $\pi_i : Z \to \mathbb{R}$. It represents the 'selfish' payoff, which is independent of any feelings of reciprocity, obligation, or behavioral concerns. Since behavior strategies induce a probability distribution over terminal notes, material payoffs can be redefined as $\pi_i : \Delta^H \to \mathbb{R}$.

In this paper, I employ the notational convention that i and j always refer to different people. In all examples, player 1 is male and player 2 is female.

Beliefs and updating. Players form beliefs about their opponent's strategies (first order belief) and what they think their opponent thinks of their own strategies (second order belief). Denote player i's first order belief about j's behavior strategy σ_j by $\alpha_j \in \Delta_j^H$, and her second order belief by $\beta_i \in \Delta_i^H$. A key observation in Dufwenberg and Kirchsteiger (2004) is that psychological games require updating of beliefs for each history:

Definition 1: *For any $\alpha_j \in \Delta_j^H$ and $h \in H$, let $\alpha_j|h \in \Delta_j^H$ be the updated first-order belief (about strategies) which assigns probability 1 to all actions of player j in (the sequence) h. Beliefs for any other history $h' \neq h$ are left unchanged. β_i is updated in the same fashion.*

After observing some action a_j (or more generally history h), each player updates their first and second order belief to match this past actions. For instance, suppose player 2 holds the initial belief that player 1 cooperates in a prisoner's dilemma, $\alpha_1(C) = 1$. After observing defection, she updates her belief to $\alpha_1(C)|D = 0$.[9] As a result, actions are always seen as intentional, not as mistakes. Notice

[8] While the theory can be applied one-to-one to games with simultaneous choice in stage games, the usual updating process assumed in the literature leads to some unappealing behavior. Footnote 16 discusses this point further after all concepts are introduced.

[9] While these beliefs are non-strategic, they affect kindness perceptions and hence require some form of updating if

that such updating behavior implies that players give up on non-degenerate probabilistic beliefs (about the past) after observing the other player's action. Randomized choice is interpreted not as conscious randomization, but rather as choice frequencies at the population level.[10] True randomization can be introduced by public randomization-devices; for detail see Sebald (2010).

Define the set $\Delta_j^H|h$ as the set of j's strategies that lead to history h with probability 1 (assuming i also plays the respective actions in h with certainty). It follows that $\alpha_j|h \in \Delta_j^H|h$. Whenever a term features multiple updated beliefs, or also conditions on history h, I simply condition once at the end, i.e. $\pi_i(\beta_i, \alpha_j|h) := \pi_i(\beta_i|h, \alpha_j|h \mid h)$.

Perceived kindness. Player i forms beliefs about j's kindness by comparing the payoff she thinks she obtains, $\pi_i(\alpha_j, \beta_i)$, against a reference point $\pi_i^r(\beta_i)$. The reference point is a combination of the highest and lowest payoff that player j can induce by using a subset of strategies $E_j \subseteq \Delta_j^H$. Besides second order beliefs, the subset E_j is the essential ingredient in reciprocity models with intentions. Its definition critically affects all kindness perceptions, preferences and behavior.

Given that the definition of E_1 is central to this paper, let's look at a simple choice problem in order to understand why the literature defines the reference point on a subset of payoffs. Suppose player 1 can either play a_1, which leads to payoffs $(\pi_1, \pi_2) = (10, 10)$, or a suboptimal alternative a_1', which results in $(-100, 0)$; $A_1 = \{a_1, a_1'\}$. Assume the reference point is the simple average of player 2's highest and lowest payoff resulting from actions in E_1. If $E_1 = A_1$ then $\pi_2^r = (10 + 0)/2$ which would imply that player 2 perceives action a_1 as kind.[11] Although a_1 makes player 2 better off, it is also player 1's dominant choice. As a result, it is unclear whether player 2 perceives such selfish action as kind. 'The fact that my neighbor doesn't throw stones at my window doesn't make him kind.'[12] Indeed, if such actions were to have an effect, any level of kindness could be generated by simply adding dominated payoffs to the game. If instead E_1 is defined over the set of Pareto efficient actions, that is $E_1 = \{a_1\}$, then a_1 is neither kind nor unkind as it yields the same payoff as the reference point, $\pi_2^r = 10$. Here, the notion of Pareto efficiency reduces the set A_1 to what player 2 considers the 'sensible set' E_1.

In this paper, I introduce the idea of trust-efficiency to define E_1. Trust-efficiency uses a notion

unexpected events occurs. The simplest example of this is a sequential prisoners dilemma. Without updating (C, cc) is an equilibrium. Player 2 cooperates no matter what given her (correct) belief that player 1 is kind due to C. However, if the first player were to defect, this belief would not be sustainable. By updating the initial beliefs when defection occurs, the (equilibrium) strategy for player 2 becomes conditional cooperation.

[10] Battigalli and Dufwenberg (2009) interpret randomized choices of player i as a common first-order belief of i's opponents about i's strategy. This results in an "equilibrium in beliefs" as in Aumann and Brandenburger (1995).

[11] While E_j is technically defined as a subset of behavioral strategies, for all examples, I will indicate which (pure) actions are part of it.

[12] The payoffs implicitly assume that the neighbor is not a rebellious teenager, who enjoys breaking windows.

of Pareto efficiency that is generally based on i's second order belief regarding her own response, but which is adjusted for her hypothetical thought process of 'what if I act selfishly?'

Definition 2: *A behavior strategy $\sigma_j \in \Delta_j^H$ is **Pareto efficient given** $\sigma_i \in \Delta_i^H$ if there is no other strategy $\sigma_j' \in \Delta_j^H$ that gives at least one player strictly more, without making the other worse-off, that is $\pi_k(\sigma_j', \sigma_i|h) \geq \pi_k(\sigma_j, \sigma_i|h)$ for all $h \in H$, $k \in \{1,2\}$ with strict inequality for at least one player.*

Define player i's material best-response as the behavior strategy $\sigma_i^{mBR}(\alpha_j)$ that maximizes i's payoff for all possible histories taking i's first order belief about j's strategy α_j as given; that is for all $h \in H$

$$\sigma_i^{mBR}(\alpha_j) \in \arg\max_{\sigma_i \in \Delta_i^H} \pi_i(\sigma_i, \alpha_j|h).$$

In case $\sigma_i^{mBR}(\alpha_j)$ is not unique, abusing notation, let it refer to a pure strategy that also maximizes j's payoff at every $h \in H$ (among $\sigma_i^{mBR}(\alpha_j)$). Denote the optimal choice at each h that make up this pure strategy by $a_{i,h}^{mBR}(\alpha_j)$, that is $\sigma_i^{mBR}(a_{i,h}^{mBR}(\alpha_j)|h) = 1$. Finally let $\sigma_i \backslash x_h$ refer to the behavior strategy that replaces the local choice at h in σ_i by $x_h \in \Delta(A_{i,h})$.[13] With these terms, I can define deviations from the material best response. An action is called generous (punishing) if it gives the other player more (less) than what he would get as a result of the material best-response.

Definition 3: *Player i's action $a_i \in A_{i,h}$ at h is **generous** if $\pi_j(\sigma_i^{mBR}(\alpha_j)\backslash a_i, \alpha_j|h) > \pi_j(\sigma_i^{mBR}(\alpha_j), \alpha_j|h)$. Action $a_i \in A_{i,h}$ is **punishing** if $\pi_j(\sigma_i^{mBR}(\alpha_j)\backslash a_i, \alpha_j|h) < \pi_j(\sigma_i^{mBR}(\alpha_j), \alpha_j|h)$. Denote player i's set of generous actions at h by $A_{i,h}^G(\alpha_j)$ and the respective set of punishing actions by $A_{i,h}^P(\alpha_j)$.*

This brings us to the central definition of E_j, which is called $TE_j(\beta_i)$ in my model:

Definition 4 (Trust Efficiency)**:** *A behavior strategy $\sigma_j \in \Delta_j^H$ is trust-efficient if it is Pareto efficient given $\beta_i^{TE} \in \Delta_i^H$, with β_i^{TE} defined as*

$$\beta_i(a_i|h)^{TE} := \begin{cases} 0 & \text{if } a_i \in A_{i,h}^G(\alpha_j) \\ \sum_{x \in A_{i,h}^G(\alpha_j) \cup a_{i,h}^{mBR}(\alpha_j)} \beta_i(x|h) & \text{if } a_i = a_{i,h}^{mBR}(\alpha_j) \\ \beta_i(a_i|h) & \text{if } a_i \in A_{i,h}^P(\alpha_j) \end{cases}$$

for all $h \in H$, $a_i \in A_{i,h}$. The set of trust-efficient strategies is denoted by $TE_j(\beta_i)$.

To illustrate this definition, take a simple game where player 1 moves first and player 2 responds.

[13] When x_h is (pure) action, $x_h \in A_{i,h}$ it is implicitly understood that it refers to $\sigma_i(x_h|h) = 1$ and $\sigma_i(a_h|h) = 0$ for all other actions.

Player 1's trust-efficient actions are his Pareto efficient actions given player 2's adjusted second order-belief, which uses her material best-response instead of any generous action that she thinks player 1 thinks she takes. For instance, if player 2's second order belief in the Prisoner's dilemma with punishment, game 2, is $\beta_2(c|C) = 1$ and $\beta_2(d|D) = 1$, then she evaluates the Pareto efficiency of C and D using $\beta_2(c|C)^{TE} = 0$ and $\beta_2(d|D)^{TE} = 1$ instead. While D is not Pareto efficient given β_2, it is given β_2^{TE}. If player 2 defected after C, player 1 would be better off by defecting himself. Action C makes player 1 vulnerable to being exploited. As a result, player 2 considers both actions to be trust-efficient. This will enable C to be perceived as kind as the reference point is determined by trust-efficient actions. The trust-efficient set for player 2, in contrast, is rather trivial, as she faces a simple decision problem at h, which doesn't depend on any future player.

β_i^{TE} treats generous and punishing actions rather differently; it adjusts beliefs in the first, but leaves beliefs in the latter unchanged. Suppose, for instance, that player 2 holds the belief $\beta_2(c|C) = 1$ and $\beta_2(p|D) = 1$ in game 2. In this case $\beta_2(c|C)^{TE} = 0$ while $\beta_2(p|D)^{TE} = 1$; only C is trust-efficient. The asymmetric treatment of generous and punishing actions captures the idea player 2's perception of player 1's cooperative action depends on whether she thinks he avoids punishment or exposes vulnerability.

In a more general environment, i.e. when players move more than once, the material-best response doesn't just focus on realized play but also takes the opponent's overall strategy into account, i.e. is forward looking. The idea remains the same, in the sense that β_i^{TE} transfers any belief in generous actions to material best-replies for any given node in the game.[14]

The reference point is a simple convex combination of the highest and lowest material payoff, with payoffs restricted to the trust-efficient actions.

Definition 5: *Let player i's reference point be*

$$\pi_i^r(\beta_i|h) := \lambda \cdot \max_{\sigma_j \in TE_j(\beta_i|h)} \pi_i(\beta_i|h, \sigma_j) + (1 - \lambda) \cdot \min_{\sigma_j \in TE_j(\beta_i|h)} \pi_i(\beta_i|h, \sigma_j)$$

for some $\lambda \in [0, 1]$.

As a punishing action of player i can make a strategy of player j inefficient, the reference point may be discontinuous in β_i. If this is the case, let π_i^r refer to the smoothed out, continuous version of the reference point in all subsequent expressions.[15]

[14] For simplicity, I opted to not explicitly indicate that β_i^{TE} is a function of α_j. I hope that this helps making expressions more easily understood instead of having the opposite effect.

[15] In contrast to generous actions, I am unaware of a game that actually requires mixed strategies in punishing actions. In general, when a player prefers to take a punishing action a_i and holds beliefs that $\beta_i(a_i|h) \in [0, 1)$ then she will also want to punish for $\beta_i(a_i|h) = 1$; the simple, non-continuous reference point is usually enough. For details on

Player i forms beliefs over the kindness of j's strategy; She compares her perceived payoff $\pi_i(\alpha_j, \beta_i)$ against the reference point $\pi_i^r(\beta_i)$.

Definition 6: *Player i perceives j's kindness from strategy α_j at h according to the function $\kappa_j : \Delta_j^H \times \Delta_i^H \to \mathbb{R}$ with*

$$\kappa_j\left(\alpha_j, \beta_i \mid h\right) := k\left(\pi_i(\alpha_j, \beta_i), \pi_i^r(\beta_i) \mid h\right)$$

with $\frac{\partial k(\cdot)}{\partial \pi_i} \geq 0$, $\frac{\partial k(\cdot)}{\partial \pi_i^r} \leq 0$, $k(\pi_i = \pi_i^r, \cdot) = 0$, and a continuous $k(\cdot)$.

Example: If $k(\cdot)$ is linear, the function reduces to the usual $\kappa_j(\alpha_j, \beta_i|h) = \pi_i(\alpha_j, \beta_i|h) - \pi_i^r(\beta_i|h)$. This function will be used in all examples.

In general, $k(\cdot)$ can describe more general functional forms such as bounding kindness (to 1), or allowing for dimishing effects as payoffs scale up.

Utility and Equilibrium.

Definition 7: *The utility of player i at $h \in H$ is a function $U_i : \Delta_i^H \times \Delta_j^H \times \Delta_i^H \to \mathbb{R}$ defined by*

$$U_i\left(\sigma_i, \alpha_j, \beta_i|h\right) = \pi_i(\sigma_i, \alpha_j|h) + \gamma_i \cdot \kappa_j\left(\alpha_j, \beta_i|h\right) \cdot \pi_j(\sigma_i, \alpha_j|h) \tag{1}$$

where γ_i is a non-negative parameter capturing i's concern for reciprocity.

The equilibrium is defined using the multi-selves approach. An agent (i, h) maximizes i's conditional utility at h by choosing the local action, taking 'her' strategy at all other nodes as given. This approach is necessary as the agent's preferences may change over time.

Definition 8: $(\sigma^*, \alpha^*, \beta^*)$ *is a reciprocity with trust equilibrium (RTE) if for all $i \in N$, for each $h \in H$, and for any $a_i^* \in A_{i,h}$ it holds that*

- *if* $\sigma_i^*(a_i^*|h) > 0$ *then* $a_i^* \in \arg\max_{a_i \in A_{i,h}} U_i\left(\sigma_i^* \backslash a_i, \alpha_j^*, \beta_i^*|h\right)$
- $\alpha_i^* = \sigma_i^*$
- $\beta_i^* = \sigma_i^*$

The equilibrium has the usual feature that players make optimal decisions at every h taking behavior and beliefs in other unreached histories as given. Moreover, first and second order beliefs are correct

how to smooth out the reference point, see Appendix A in Rabin (1993); in this regard, see also the discussion of conditional-efficiency in section 4.

and are updated as the game progresses. The updating process views unexpected actions as intentional, not as mistakes.[16]

Proposition 1: *An equilibrium exists if $\kappa_i(\cdot)$ is continuous for all $i \in N$.*

The proof follows the strategy of DK04. The key observation is that behavior at unreached nodes (or rather second order beliefs about it) has direct effects on preferences due to kindness perceptions. As a result, the usual backward induction argument fails. Instead, the existence proof requires that all histories are analysed simultaneously.

Example. Sequential prisoner's dilemma, game 1. Suppose for simplicity that player 1 is selfish,$\gamma_1 = 0$, and that the reference point is the minimum efficient payoff, $\lambda = 0$.[17] For any β_2, player 1's efficient set of actions is $TE_1(\beta_2) = \{C, D\}$. To understand this, start with player 2's second order belief that she conditionally cooperates. For such belief C Pareto-dominates D so that C could not be perceived as kind. However, the efficiency notion, instead, uses 2's material best-response after C. Given such response, both actions are Pareto efficient. Player 1 makes himself vulnerable by playing C as subsequent defection would lower his payoff below what he could have obtained by playing D. This leads player 2 to perceived the mutually beneficial action C as kind.

At C her choices yield utilities of

$$U_2(c, \beta_2|C) = 1 + \gamma_2 \left[\beta_2(c|C) + 2(1 - \beta_2(c|C)) - (-\beta_2(c|D))\right] \cdot 1, \text{ and}$$

$$U_2(d, \beta_2|C) = 2 + \gamma_2 \left[\beta_2(c|C) + 2(1 - \beta_2(c|C)) - (-\beta_2(c|D))\right] \cdot (-1).$$

These two expressions clarify how the second order belief about behavior at unreached nodes affects 2's perception of 1's kindness. The more 2 believes 1 believes she cooperates at D, the kinder she perceives him to be.

Since D minimizes 2's payoff it can never be kind, however. Player 2 defects after D, $\sigma_2(c|D) = \beta_2(c|D) = 0$. She cooperates at C if and only if $2\gamma_2 \left(\beta_2(c|C) + 2(1 - \beta_2(c|C))\right) \geq 1$. Thus for $\gamma_2 \geq \frac{1}{2}$ she cooperates. For $\gamma_2 < \frac{1}{4}$, she defects, and for intermediate values she randomizes with

[16] At this point, I should comment on why the model is only defined over games with strictly sequential choices. The game matching pennies illustrates an interesting issue that occurs when there are simultaneous choices in sequential games.

	L	R
T	1,0	0,1
B	0,1	1,0

Suppose both people were to belief that the other player is perfectly randomizing. Ex-ante, this leads to zero-kindness. Ex-post one player wins, the other loses. The updating process in my model - and Dufwenberg and Kirchsteiger (2004) or Battigalli and Dufwenberg (2009) - places probability 1 on the observed actions. Hence, ex-post, the winner considers the loser as kind, and the loser views the winner as unkind. If they had the opportunity to reward or punish in a subsequent period - they would choose to do so. While they may want to do so for status concerns, it seems counterintuitive that this is a result of reciprocity when they agreed ex-ante that kindness is zero.

[17] Clearly, using the more familiar $\lambda = 1/2$ doesn't add any additional insight to this particular example, but gives rise to a more complicated looking reference point. Whenever there is no punishment, using $\lambda = 0$ is often better.

$\sigma_2(c|C) = \beta_2(c|C) = 2 - \frac{1}{2\gamma_2}$. Player 1 cooperates if and only if $\gamma_2 \geq \frac{3}{8}$. The intermediate case highlights that an equilibrium in pure strategies may not exist when players are not purely motivated by material-payoffs, unlike in Zermelo (1913). Player 2 only views 1 as sufficiently kind (to motivate her to cooperate) when she thinks he thinks she defects, but not when she thinks he thinks she cooperates.

4. Trust and Conditional Efficiency

In this section, I compare my model to the reciprocity models of Rabin (1993) and Netzer and Schmutzler (2014). By also comparing it to a model of trust, i.e. Cox et al. (2016), I will explain why it is called reciprocity *with trust*, and how it differs from pure trust.

For the remainder of this paper, the games of interest are two-stage games where player 1 moves first and 2 responds. The focus will be on player 2's equilibrium response as player 1's preference is identical across all models.

In this setting, I will use $TE_1(\beta_2)$ to refer to player 1's trust-efficient *actions* (instead of behavior strategies). Moreover, for $a_1, a_1' \in A_1$, a_1 Pareto-dominates a_1' given $\beta_2 \in \Delta_2^H$, in short a_1 $\mathbf{PD}(\beta_2)$ a_1', if $\pi_k(a_1, \beta_2) \geq \pi_k(a_1', \beta_2)$ for all $k \in \{1, 2\}$, with strict inequality for at least one player. Similarly if a_1 dominates a_1' given $\beta_2^{TE} \in \Delta_2^H$ I use a_1 $\mathbf{PD}(\beta_2^{TE})$ a_1'.

Conditional-efficiency. Rabin (1993) models efficiency *conditional* on player 2's response (or rather the second-order belief thereof) when defining reciprocity for normal-form games. For a two-stage game, it translates to:

Definition 9 (Conditional Efficiency, Rabin '93)**:** *An action $a_1 \in A_1$ is **conditionally efficient** if it is Pareto efficient given $\beta_2 \in \Delta_2^H$. Denote the set of conditional efficient actions by $CE_1(\beta_2)$.*

The fundamental difference between Rabin's original model and mine is his definition of efficiency. He simply uses second order beliefs to determine whether an action is efficient, which will be consistent with actual choices in equilibrium. In my model, I start with second order beliefs to determine efficiency, but use material-best replies instead of any beliefs in generous actions.

Since Rabin focused on normal form games, his model featured no belief updating, however. Netzer and Schmutzler (2014) and Le Quement and Patel (2017) apply the notion of conditional-efficiency to sequential games, allowing for such updating. For this paper, I define a *conditional Reciprocity Equilibrium (conRE)* as an equilibrium that takes all ingredients from section 3, but replaces trust-

efficiency with conditional efficiency.

The difference between the two efficiency notions is best illustrated by revisiting the prisoner's dilemma. From earlier, we know that player 2 defects after defection, $\beta_2(c|D) = 0$. The second order belief about how player 2 responds after cooperation is crucial, however. If 2 thinks that 1 believes she defects, $\beta_2(c|C) = 0$, then both C and D are efficient: C is better for 2, while D is better for 1. If she thinks he believes she cooperates, $\beta_2(c|C) = 1$, C is the only efficient action as it is mutually beneficial. In general, C is the only efficient action if it also maximizes player 1's payoff, i.e. $\beta_2(c|C) \geq 1/2$. As the reference point is based on the efficient actions only, it follows that when both actions are Pareto efficient, $\beta_2(c|C) < 1/2$, she perceives action C as kind, $\kappa_1(C, \beta_2(c|C)) = 2 - \beta_2(c|C) - 0$. For $\beta_2(c|C) \geq 1/2$ only action C is efficient and so the reference point is identical to her (perceived) payoff, $\kappa_1(C, \beta_2) = 2 - \beta_2(c|C) - (2 - \beta_2(c|C)) = 0$. While player 2 benefits from C, she attaches zero kindness to 1's action. A reciprocity model based on the conditional-efficiency notion adopts the cynical perspective that an action can only be kind when it occurs at 1's expense. This places a strict limit of how much player 2 can cooperate in equilibrium. Even when she is sufficiently motivated by reciprocity, $\gamma_2 \geq 1/3$, it must be that she cooperates with slightly less than $1/2$ probability. For $\gamma_2 < 1/3$, both RTE and conRE coincide.[18]

The prisoner's dilemma example suggests that (a) trust-efficiency features a more generous reference point, and that (b) the difference between trust-efficiency and conditional-efficiency is trust. I will now explore each idea and show how they relate to each other.

Proposition 2: *Let β_2 be an equilibrium belief for either RTE or conRE (or both). Then*

$$\min_{a_1 \in TE_1(\beta_2)} \pi_2(a_1, \beta_2) \leq \min_{a_1 \in CE_1(\beta_2)} \pi_2(a_1, \beta_2).$$

This proposition confirms the notion that actions are perceived as kinder in my model. When positive reciprocal responses make actions inefficient, they remain efficient under trust-efficiency. Since a lower minimum efficient payoff translates into a lower reference point, actions are perceived as kinder.[19] I will now show under which conditions the reference points differ. To do so, I introduce a notion of trust and illustrate how it relates to the notion of trust- and conditional-efficiency.

[18] Netzer and Schmutzler (2014) make a similar observation in a gift-exchange game, where a firm is known to be selfish. They highlight that a high wage offered by a firm (that moves first) isn't kind if the firm expects the worker to reciprocate by exerting high effort. If a 'low wage' leads to 'low effort' and 'high wage' to 'high effort', and the payoff set from the second dominates the first, the efficient set collapses to a singleton. This makes any high-wage offer not kind. Their goal is to highlight the limits of reciprocity when one player is known to be selfish. In contrast, this paper is motivated by trying to understand when cooperation is possible if both are (known to be) reciprocal - whenever selfishness of player 1 is assumed in this paper, it is mainly done to simply derivations.

[19] All results in this paper are based on equilibrium beliefs. For an example that highlights what can happen when only rationalizability is required, see Appendix B.1, game 11. The example underlines that RTE is best used as an equilibrium model.

Trust. Cox et al. (2016) define trust for two-stage games as follows: For any $a_1, a_1' \in A_1$, a_1 is more trusting than a_1' if and only if

$$\pi_1(a_1, \sigma_2^{mBR}) < \pi_1(a_1', \sigma_2^{mBR}) \text{ and } \max_{a_2 \in A_{2,a_1}} \pi_1(a_1, a_2) > \pi_1(a_1', \sigma_2^{mBR}).$$

The first condition captures vulnerability. Given 2's selfish responses, player 1 is worse off by playing a_1 than a_1'. The second condition requires that there is a response to a_1 that makes player 1 better off than the selfish outcome after a_1'. For the purpose of this paper, I use a slight variation of their definitions, namely:

Definition 10: *Let $a_1, a_1' \in A_1$. a_1 is **more trusting than** a_1' if and only if*

$$\pi_1(a_1, \sigma_2^{mBR}) < \pi_1(a_1', \sigma_2^{mBR}) \text{ and } \max_{a_2 \in A_{2,a_1}} \pi_1(a_1, a_2) > \pi_1(a_1', \beta_2).$$

This definition keeps the critical first condition that focuses on the relative material best-response payoffs, while relaxing the second condition to only require a_1 to potentially do better than a_1' given β_2.[20] Cox et al. (2016) remark that some definitions of trust allow for the possibility that the second player can be better off, but chose not to include it as it would reflect gifts or generosity, not trust. I will argue later, that player 2's payoff is rather relevant to predict when trust is reciprocated or betrayed.[21]

Proposition 3: *Let $(\sigma^*, \alpha^*, \beta^*)$ be an RTE and let player 1 only have two actions, $A_1 = \{a_1, a_1'\}$. If $TE_1(\beta_2^*) = \{a_1, a_1'\}$ and $CE_1(\beta_2^*) = \{a_1\}$ then a_1 is more trusting than a_1'.*

The proposition can be read as 'RTE is conditional efficiency plus trust'. While a_1 Pareto dominates a_1' given β_2^* player 2 is tempted by her selfish option (after a_1) and understands that if she took it, player 1 would have been better-off under the alternative. This makes a_1 *trusting*. From a mechanical standpoint, note that it is not the trusting action that becomes efficient, but the alternative action that was less trusting. Consequently the trusting action appears kind, which allows for the possibility of it being rewarded.

When generalizing this result to $|A_1| \geq 2$, it is useful to introduce notation for the efficient action that minimizes 2's payoff. Denote this action for the respective efficiency notions, TE and CE, by

[20] This definition is not meant to capture the best-definition of trust (which may want to hold constant expected behavior off-path), but rather, to be a useful language for describing selfish payoffs. For most games in the experimental literature, this definition *implies* Cox et al. (2016)'s definition.

[21] They also define 2's action as trustworthy after a_1 if it gives 1 (weakly) more than the payoff he would get if 2 acted selfishly when 1 chooses the least-trusting action relative to a_1.

$$M_1^{TE_1(\beta_2)} = \underset{a_1 \in TE_1(\beta_2)}{\arg\min} \ \pi_2(a_1, \beta_2) \qquad \text{and} \qquad M_1^{CE_1(\beta_2)} = \underset{a_1 \in CE_1(\beta_2)}{\arg\min} \ \pi_2(a_1, \beta_2).$$

Since the maximum payoff is always efficient and thus is identical across models, the reference point differs if and only if these two actions are different.[22]

Proposition 4: *Let $(\sigma^*, \alpha^*, \beta^*)$ be an RTE. If $M_1^{TE_1(\beta_2^*)} \neq M_1^{CE_1(\beta_2^*)}$ then any action $a_1 \in A_1$ that Pareto-dominates $M_1^{TE_1(\beta_2^*)}$ given β_2^* is more trusting than $M_1^{TE_1(\beta_2^*)}$.*

By proposition 2, we know that $M_1^{TE_1}$ induces a weakly lower payoff than $M_1^{CE_1}$. Hence, when the minimum efficient payoff is strictly lower under trust-efficiency, the proposition states that it is due to trust. This reiterates that RTE can be interpreted as conditional efficiency plus trust. Since the proposition only describes when preferences are different, it is not clear whether the equilibrium itself must be different. This is tackled next.

Proposition 5: *Let $(\sigma^*, \alpha^*, \beta^*)$ be an RTE. If $M_1^{TE_1(\beta_2^*)} \neq M_1^{CE_1(\beta_2^*)}$ then $(\sigma^*, \alpha^*, \beta^*)$ cannot be a conditional Reciprocity Equilibrium.*

Whenever trust plays a role, an RTE cannot be an equilibrium in the Rabin model. To understand this proposition, recall that when player 2 conditionally cooperates in the prisoner's dilemma, cooperation is the only efficient choice for player 1 given the notion of conditional-efficiency. In this case, it cannot be perceived as kind and so player 2 cannot cooperate in response. The proposition generalizes this observation. Whenever there is an equilibrium where trust-efficiency yields a different minimum payoff than conditional-efficiency (for the same second order belief), the respective action that induces 2's minimum payoff for trust-efficiency is followed by a positive reciprocal response. But such response isn't feasible in a conditional Reciprocity equilibrium as this action is not perceived as kind.

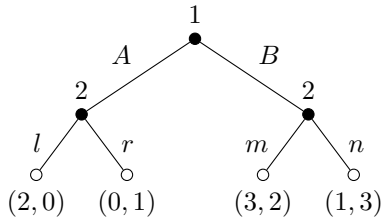
If the minimum payoff is the same, on the other hand, preferences are identical across both models so that the equilibrium is also a conditional RE.

4.1. Trust Does Not Imply Efficiency

Cox et al. (2016) intentionally define trust using only player 1's payoffs. As a result, it is a very different notion than efficiency and kindness. It turns out that a choice can be trusting, but also unkind and Pareto dominated. When players are motivated by reciprocity, such trust is likely to be betrayed. Game 4 illustrates this idea.

[22] In general it does not need to be true that $CE_1(\beta_2) \subseteq TE_1(\beta_2)$. Game 12 in Appendix B.2 illustrates a case where action $a_1 \in CE_1(\beta_2)$ but is not in $TE_1(\beta_2)$. It also makes clear why such cases aren't problematic.

In this example action A makes player 2 strictly worse off. Since $\pi_1(A, \sigma_2^{mBR}) = 0 < 1 = \pi_1(B, \sigma_2^{mBR})$ and $max_{a_2 \in A_{2,A}} \pi_1(A, a_2) = 2 > 1 = \pi_1(B, \sigma_2^{mBR})$ action A is more trusting than B. As 2's second order belief must assign probability one to r, B Pareto dominates A for any second order belief of how she responds after B. It follows that the only efficient action is B despite the fact that A is more trusting than B. As a result, player 2 always takes her selfish action rn. She betrays player 1's trust after A and does not reward the mutually beneficial action B.



Game 4: Trust doesn't imply efficiency

5. Dufwenberg and Kirchsteiger '04

In this section, I compare the reciprocity with trust model to Dufwenberg and Kirchsteiger (2004) - which is the standard intention-based reciprocity model for sequential games. I will argue that their model classifies 'too many' actions as efficient, giving rise to a reference point that is often too low, and thus predicting too much positive reciprocity. Once again, the games of interest will be two-stage games.

Definition 11 (Unconditional-efficiency, Dufwenberg and Kirchsteiger '04)**:** *An action $a_1 \in A_1$ is* ***unconditionally-efficient****, if it is Pareto-efficient for at least one strategy of player 2, $\sigma_2 \in \Delta_2^H$. Denote the set of unconditional efficient actions by UE_1.*

Efficiency no longer takes (the second order belief about) 2's strategy as given, but instead requires that it isn't Pareto-dominated by some $a_1' \in A_1$ for *all* possible strategies of player 2. We can think about this from an ex-ante perspective: Without knowing how player 2 might respond, any action that is dominated for all possible responses is eliminated. Define a unconditional Reciprocity Equilibrium (unRE) as an equilibrium that takes all ingredients from section 3, but replaces trust-efficiency with unconditional efficiency.

Returning once again to the prisoner's dilemma, it should be clear that both actions are unconditionally efficient, $UE_1 = \{C, D\}$. If player 2 always defects, C is better for 2, while D is better for 1. Consequently, the equilibrium predictions of RTE and unRE are identical. In general,

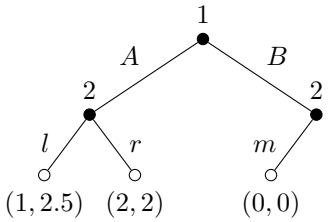
unconditional efficiency is less restrictive than trust-efficiency. While trust-efficiency is based on a particular strategy, β_2^{TE}, unconditional efficiency only requires the existence of any strategy for which player 1's action is not Pareto-dominated.

Proposition 6: *For any β_2 it holds that $TE_1(\beta_2) \subseteq UE_1$ and thus*

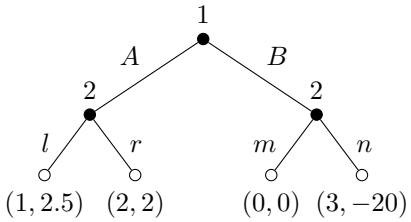
$$\min_{a_1 \in UE_1} \pi_2(a_1, \beta_2) \leq \min_{a_1 \in TE_1(\beta_2)} \pi_2(a_1, \beta_2).$$

Compared to trust-efficiency, actions tend to be perceived as kinder when unconditional efficiency is used. As a result unRE predicts more (less) positive (negative) reciprocity than RTE.

We now turn to game 5 and 6. In game 5, A Pareto-dominates B regardless of player 2's action; A is not kind. In game 6, A is no longer the only unconditionally-efficient action. B is efficient because it yields the largest payoff for player 1 if, for example, player 2 were to play strategy ln. Consequently A is kind, which makes player 2 want to reciprocate by using strategy rm in equilibrium. Notice, however, that B gives player 2 strictly less than A. As a result, it cannot be kind and so player 2 will always choose m after B.



Game 5



Game 6

This example highlights the fundamental problem of the unconditional-efficiency notion: Since unconditional-efficiency is independent of what players actually do (or want to do), it opens up for the possibility of adding various unused choices to create kindness. Note that this example doesn't require equilibrium beliefs. Since player 2 never plays n, player 1 cannot believe that she does, and so player 2 must hold the second order belief that she plays m (after B): n is not rationalizable, yet affects kindness perceptions.[23] While she may view action B as greedy, it is unlikely that n's existence motivates player 2 to play l. This suggests that DK04's approach of modelling efficiency without any link to second order beliefs leads to a model that predicts too much positive reciprocity. In the next subsection, I will explore how this example generalizes.

[23] An action $a_2 \in A_2(h)$ cannot be rationalizable if there exists no second order belief β_2 under which she wants to take such action. For a full definition of rationalizability in psychological games, see Battigalli and Dufwenberg (2009).

5.1. General Comparison

I now proceed to describe in which cases the reference point in the RTE differs from DK04. There are two main classes of payoffs, one that generalizes the previous example and one that is linked to games where player 2 punishes in some nodes of the game. The latter case sheds light on why some experimental papers don't observe much positive reciprocity.

The easiest way to compare RTE to DK04 is by looking at games where player 1 has only two actions, $|A_1| = 2$. Moreover, I make the following equilibrium selection assumption:

Assumption 1: *Player 2 doesn't punish after $a_1 \in A_1$ if there is an alternative action $a_1' \in A_1$ with $\pi_2(a_1, \sigma_2^{mBR}) > \pi_1(a_1', \sigma_2^{mBR})$.*

The assumption eliminates very pessimistic beliefs that could theoretically lead to punishing behavior after an action a_1 that makes player 2 better off than under alternative a_1' if she takes her material best-response after each.[24] [25] Note that since player 1 only has two actions, only one of the two can be seen an unkind and punished. The assumption does not eliminate punishment in general. It can be shown that an equilibrium still exists even when the above is assumed.[26]

Proposition 7: *Let player 1 have two actions, $A_1 = \{a_1, a_1'\}$, and let $(\sigma^*, \alpha^*, \beta^*)$ be an RTE. If assumption 1 holds, $UE_1 = \{a_1, a_1'\}$ and $TE_1(\beta_2^*) = \{a_1\}$ then one of the following holds:*

1. *$\pi_1(a_1, \sigma_2^{mBR}) > \pi_1(a_1', \sigma_2^{mBR})$ and $\pi_2(a_1, \sigma_2^{mBR}) > \pi_2(a_1', \sigma_2^{mBR})$, or*
2. *$\pi_1(a_1, \sigma_2^{mBR}) < \pi_1(a_1', \sigma_2^{mBR})$ and $\pi_2(a_1, \sigma_2^{mBR}) > \pi_2(a_1', \sigma_2^{mBR})$.*

The first case is the most extreme. When player 2 uses her material best-response after both actions, a_1 Pareto dominates a_1'. Both players are better off. Here, it appears difficult to rationalise why a_1 is perceived as kind. One example of this was game 6. There are various other ways in which additional actions lead to the same result.[27]

The second case is slightly more interesting. Under material best-replies, both action are indeed efficient: Player 1 would prefer a_1' over a_1. As player 2 considers a_1' unkind, however, she punishes him in response to a_1', making it Pareto-dominated. Moreover, she understands that by choosing

[24] Take a game with $A_1 = \{a_1, a_1'\}$. a_1 induces payoffs for player 2 of 1 or -1 (depending on 2's response), while a_1' leads to payoff 0. Let $\lambda = 1/2$. Suppose 1's payoffs are such that both actions are (always) efficient. Let $\tilde{\beta}$ be the second order belief that she takes the action that leads to a payoff of 1. In this case $\kappa_1(a_1, \tilde{\beta}) = (\tilde{\beta} - (1-\tilde{\beta}) - (\tilde{\beta} - (1-\tilde{\beta}) + 0)/2) = \tilde{\beta} - 1/2$. If 2 believes 1 thinks she punishes after a_1, then she may want to punish if it lowers 1's payoff sufficiently. While $\tilde{\beta} = 1$ represents 'normal' beliefs, $\tilde{\beta} = 0$ takes an extremely negative view player 1 aims to hurts player 2. For most normal interactions such belief appears unlikely.

[25] While the assumption is written in terms of behavior, it could have also been written in terms of the equivalent beliefs, which implies such behavior.

[26] See Appendix A, Lemma 9 for detail.

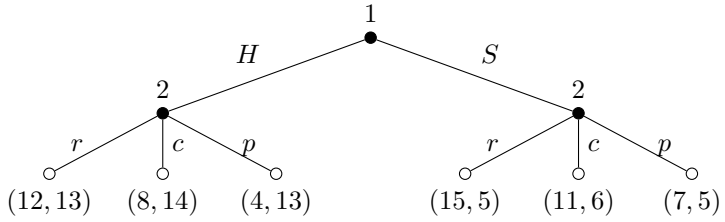
[27] For instance a_1' may be become efficient if there exists generous action after a_1 that results in a lower payoff for player 2 than her selfish payoff after a_1'. Alternatively, it may be due to a punishment action after a_1, which once again leads to a lower payoff for 2 than her selfish payoff after a_1'.

a_1, player 1 avoids punishment. While he improves her payoffs, she views player 1's action as selfish because it is not trusting. This example highlights the interaction between rewards and punishment and is explored in more detail in the next section.

The proposition generalizes to $|A_1| > 2$, yet requires a more involved assumptions given the larger set of actions. It is discussed in Appendix B.

6. Applications

This section revisits games in the literature where player 2 can reward and punish. Game 7 is taken from Offerman (2002) and represents a perfect example for part 2 of proposition 7. The second player has the option, at a cost of 1 unit, to reward (r) or to punish (p) player 1 by 4 units. Player 1 can be helpful (H) or selfish (S). In a SPNE with selfish players, $\gamma_1 = \gamma_2 = 0$, player 1 plays S and player 2 always acts cool, cc.



Game 7: Offerman (2002)

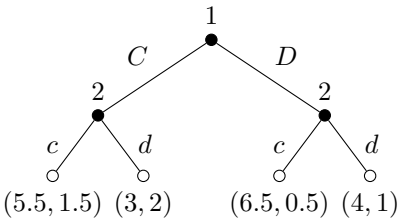
The key observation for this game is that when player 2 punishes after S, H Pareto dominates S for both $\beta_2 = rp$ and $\beta_2 = cp$. As a result, only H is trust-efficient and therefore not kind. Player 2's optimal response to H is c.

In one treatment, Offerman (2002) allows player 1 to make a choice himself, while in the other, player 1's choice is made for him by a computer He finds clear evidence of negative reciprocity but limited, not statistically significant, evidence of positive reciprocity: 83.3% of the second movers punish the selfish choice (vs. 16.7% in the 'random treatment'), whereas 75% of second players reciprocate helpful choices (vs 50% in the 'random treatment'). Offerman concludes that negative intentionality matters more than positive intentionality and explains this with self-serving attribution.[28] Al-Ubaydli and Lee (2009) repeat Offerman's experiment employing a structural approach, in which the reciprocity model by Falk and Fischbacher (2006) is used to account for asymmetries due to inequity aversion.

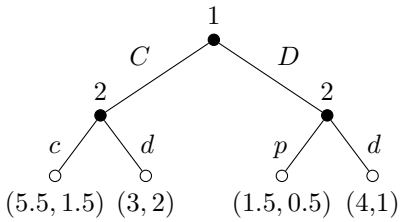
[28] Intentional harm hurts player 2's self-esteem and therefore induces punishment. However, she attributes being treated well by players or nature to being 'a good person deserves help'. As a result, there is no need to reciprocate.

They also find that negative intentions are more likely to be followed by punishment than positive ones, and subscribe to Offerman's conclusion of self-serving attribution. My analysis emphasizes that this conclusion does not need to be correct, since it neglects to account for the interaction between punishment and rewards. It is because player 2 punishes in response to the selfish choice, that the helpful action is no longer perceived as kind.[29]

The Offerman result was surprising because both actions are efficient in DK04, and thus C is perceived as kind. Moreover, DK04's model predicts that either player 2 rewards H and punishes S, or always acts neutral. This is due to the fact that the game is fully symmetric, and that their reference point put equal weights on the maximum and minimum efficient payoff. Simply relaxing the parametric specification of equal weights, i.e. for example to $\lambda > 1/2$, is not the solution as we will see next. In a recent experimental paper, Orhun (2018) observes similar behavior as in Offerman. Instead of replacing player 1's choice with a computer, she varies the set of choices in the game. In particular she varies player 2's choice in a sequential prisoner's dilemma after defection, see game 8 and 9. In the game 8, player 2 has the usual option to cooperate, whereas in game 9 she can punish player 1 instead.



Game 8: Sequential prisoner's dilemma



Game 9: Prisoner's dilemma with punishment

The option to punish significantly alters the players' perception of the game. On average, player 1 believes that in 41% of the times player 2 punishes after D, and player 2 holds a second order belief that he thinks she punishes in 54% of all cases. Under these beliefs, cooperation is player 1's payoff maximizing choice. Orhun finds that cooperation rates (after C) fall significantly from 57% in game 8 to 35% in game 9.[30]

Orhun remarks that DK04 cannot predict the drop in cooperative behavior. Indeed, it cannot be explained for any possible weighting assigned to the minimum and maximum payoff in the reference point. If anything, the option to punish increases the kindness perception of C by lowering the minimum efficient payoff. In contrast, RTE predicts this exact change in behavior. Given that player 2 punishes C $\mathbf{PD}(\beta^{TE})$ D, so that C is not perceived as kind. With reciprocal players RTE predicts

[29] In the random treatment, neither outcome is kind or unkind, and player 2 always plays c.
[30] Unfortunately, she doesn't compare this to what player 2 would have done in a dictator game.

(C, cd) in the usual prisoner's dilemma and (C, dp) in the prisoner's dilemma with punishment. I am unware of any other model that makes the same prediction.[31]

For example, it is unclear how type-based models, i.e. Levine (1998) or Gul and Pesendorfer (2016), could explain player 2's response. Whenever it is optimal for player 1 to cooperate in both games, player 2's belief about 1's type must be the same. As a result, she must also cooperate in the prisoner's dilemma with punishment.[32]

I conclude this section with a game in the spirit of Andreoni et al. (2003). We will see that in game 10, each reciprocity equilibrium, RTE, conRE, unRE, makes a different equilibrium prediction.

Andreoni et al. are interested in the incentive effects of voluntary rewards and punishments, and how this can be used to shape economic institutions. In their experiment, player 1 decides how much of his endowment to give to player 2. The choice set of player 2 varies by treatment. She either has no choice (dictator game), can reward, punish, or reward or punish the sender. They find that offers are lowest in the dictator game, second lowest in treatment that only allows for punishment, second highest in the treatment that only allows for rewards, and highest in the treatment that allows for both rewards and punishment. In general, punishment eliminates very selfish offers, while rewards incentivise high offers. Like all papers in this section, they observe that the option to punish lowers the demand for rewards. This pattern remains significant even for the most generous offers. The authors find this behavior puzzling and conjecture that an explanation may require a definition of kindness that changes with the treatment.

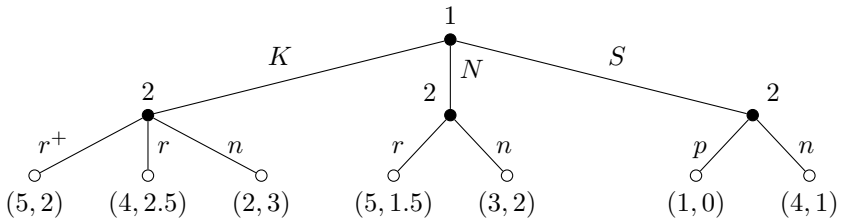
From our discussion in the previous paragraphs, it should be clear by now that this is exactly the behavior RTE predicts. When the most selfish offers are met with punishment, they become inefficient, giving rise to an increase in the reference point. As a result, more generous offers appear less kind and the demand for rewards is lower relative to a game in which player 2 cannot punish.

Instead of varying the choice sets as in the previous section, I will solely use game Game 10, which is a simplified, nonlinear version of the continuous game in Andreoni et al. (2003), to highlight the different predictions of RTE, conRE, and unRE. For simplicity, player 2's choices are very limited, favoring relevance of actions over symmetry.[33]

[31] I should point out that a conditional Reciprocity model can explain the drop in cooperation rates, but cannot explain that cooperating is optimal for player 1 in the basic prisoner's dilemma.

[32] This argument takes as given that the result is not driven by very spiteful types who prefer to (D,dp) over (C,dp).

[33] The equilibrium predictions remain the same in the fully symmetric game that features mediocre (r) and strong rewards (r^+) as well as punishment (p) after each action. The equilibrium prediction for a dictator treatment would be the most selfish offer S. In the reward-only treatment both RTE and unRE predict the highest offer K with is fully rewarded (r^+). In the punish-only treatment, all reciprocity models predict that player 1 offers N to avoid punishment



Game 10

For a selfish player 1 and a reciprocal player 2, with $\gamma_2 = 1$, the RTE is (K, rnp), the unRE is (K, r^+rp) and the conRE is $\sigma_1(N) = 1$ and $\sigma_2(r|K) = 1/2 - \epsilon$, $\sigma_2(n|K) = 1/2 + \epsilon$, $\sigma_2(n|N) = 1$, $\sigma_2(p|S) = 1$, with a small, positive ϵ.

All predictions feature punishment in response to the selfish offer S. In both the conditional Reciprocity Equilibrium and RTE player 2 responds to action N with a neutral response due the fact that N Pareto dominates S, while unRE allows for some positive reciprocity.[34] After action K unRe predicts the largest reward. In RTE K is perceived as relatively less kind since the efficient minimum is $\pi_2(N, n) = 2$ and not 0. As a result, player 2 rewards less. Even at the top, punishment crowds out rewards. Player 2 is the most cynical in conRE. She rewards with less than $1/2$ probability after K, as otherwise K would be in player 1's material interest - in which case it wouldn't be kind. As a result, player 1 plays N, compared to K in RTE and unRE.

My theory can also be used to explain how incentive structures affects the responder's behavior. Fehr and Gächter (2001) show that sanctions set by employers undermine voluntary cooperation in gift-exchange games. The mechanism is similar: sanctions affect the minimum efficient wage offer as they enforce higher levels of effort, altering the reference point. This lowers kindness perceptions and positive reciprocity.[35]

after S.

[34] At closer inspection, this example highlights one negative feature of the conRE and RTE model. Punishment after S not only lowers K's kindness but also makes the neutral action N unkind. The same is true, for example in an ultimatum game, where equal splits will be perceived as unkind when low offers are punished. This feature isn't very appealing. More generally, punishing very unkind actions can make other unkind action even more unkind, resulting in an increased demand for punishment. This feature can be avoided by using an efficient set that is based on the material best-replies for unkind actions and trust-efficiency to determine how kind a seemingly kind action really is. For all games of interest, kindness for the later notion is below that of the first. Lastly, set kindness to 0 when trust-efficient kindness is negative, while it is positive under material-efficiency.

[35] Fehr and Gäechter have two treatments. In the first, they run a simple gift-exchange where firms set wages (and suggest a desired work level) and workers respond by choosing effort levels $\in [1, 10]$. They find that effort is increasing in the generosity of the wage. In a second treatment, they allow firms to set a costly sanction that has to be paid when workers shirk. It is exogenously verified with probability $1/3$ if the employee shirks. This essentially allows firms to (rationally) enforce an effort level of 4, larger than the minimum effort level of 1 without sanctions. Firms make use of such sanctions. However, it reduces voluntary cooperation - even below the rational level. To understand how my model works in this setting, start with a second order belief that she responds with the minimum rational effort. In this case, the minimum efficient wage offer in treatment 1 is $w_{T1}^{min} = 1$, while in the incentive treatment it is $w_{T2}^{min} = 4$. For these second order beliefs, an actual offer of $w = 4$ is potentially kind in treatment 1, while it is unkind in the second treatment. Since the reference point is higher in the second treatment, wage offers are perceived as relatively less kind. This can induce the worker to actually lower his effort below the rational effort for low enough wage offers.

7. Summary of Equilibria

After comparing the Reciprocity Equilibrium with Trust to the conditional and unconditional Reciprocity Equilibrium, I now summarize the general equilibrium prediction across all models.

Proposition 8: *If* $(\sigma^*, \alpha^*, \beta^*)$ *is a conditional and unconditional Reciprocity Equilibrium, then it is also an RTE.*

I have argued that kindness perceptions in RTE are less cynical than in conditional RE, but lower than in unconditional RE. When the equilibrium coincides for the two extreme kindness perceptions, it must hence also be an equilibrium in my model.

Figure 2 provides a graphical summary of how equilibrium predictions differ across all thee models. Intersection 1 is the visual equivalent of proposition 8. The equilibrium can coincide for three reasons. First, player 1's action is unambiguously kind. His action improves 2's payoff at his own expense. In this case, kindness perceptions are identical for each model. Second, player 2 is simply not motivated (enough) by concerns for reciprocity, $\gamma_2 \approx 0$, in which case different kindness perceptions become irrelevant. The selfish SPNE is nested in each model. Third, perceptions may differ but player 2 may simply not have relevant choices to respond differentially; her action set could be rather limited, or positive and negative reciprocal actions could be too costly.

Equilibria in intersection 2 feature actions that are mutually beneficial, yet are perceived as kind due to trust. RTE coincides with unconditional RE, whereas it cannot be an equilibrium for conditional RE, area 3. A simple example of this is the trust game or the sequential prisoner's dilemma.

Intersection 4 captures equilibria where actions are perceived as less kind in a RTE than in a unRE, area 5. This can be the result of either unused actions (at any history), or punishing actions.

While it is useful to have a single model that can explain and predict strictly positive reciprocal responses, as well as purely selfish payoff-maximizing choices (intersections 2, 4), area 6 highlights that RTE also makes unique predictions in more complex games. Game 10 is an example of this.

8. Discussion

The starting point behind this paper was the idea that if two reciprocal player achieve full cooperation in the simultaneous version of social dilemma, then they should also be able to achieve this in the sequential version. This is consistent with experiments on the prisoner's dilemma, see figure 1. To

Quantitatively, it is unclear, however, why effort levels remain so extremely flat for all wage offers, see Figure 3 in Fehr and Gächter (2000).

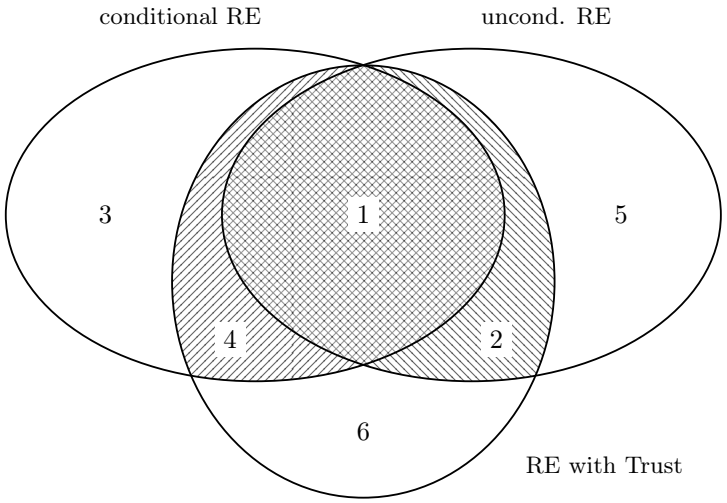


Figure 2: Equilibrium predictions overview

that end, I have proposed a new way of modelling reciprocity with intentions. By adding the idea of trust to reciprocity models, kindness perceptions become less cynical than in Rabin (1993). This allows player 2 to fully reciprocate actions that improved both her own and player 1's payoff.

Netzer and Schmutzler (2014) argued, using the conditional Reciprocity Equilibrium, that when a firm is known to be selfish, a worker does not respond to a high wage with high effort as in this case a high wage would be in the firm's best interest.[36] When the second player knows that player 1 is *surely* self-interested, it becomes easy for her to decide whether player 1 took an action for his own benefit, or also with her in mind. To capture this, the model can be extended in a way that trust-concerns are no longer relevant when a player is sufficiently confident that her opponent is selfish. It would represent a similar mechanism to the one put forward by Rotemberg (2008) for altruism. In this regard, the extension would adopt ideas from the literature of type-based reciprocity, Levine (1998), Ellingsen and Johannesson (2008) and Gul and Pesendorfer (2016). While these models can often be simpler to solve, it is unclear how they could explain the behavior in Orhun (2018). As a result, I view type-based and intention-based reciprocity models as complements.

We have also observed that actions tend to be perceived as less kind in my model than in Dufwenberg and Kirchsteiger (2004).[37] Linking efficiency (more closely) to actual behavior has the advantage that the reference point is affected less by unchosen actions. In terms of directions for future research,

[36] While all examples in this paper featured a 'selfish player' player 1, this was simply done for analytical convenience.

[37] Game 6 actually originated from my work on incomplete information. It turns out that due to the unconditional efficient set, kindness perceptions in DK04 can become independent of the prior belief over types. This gives rise to implausible behavior. For more detail, see Roel (2018b).

it would be interesting to test games like 5 and 6 in the lab.

My model also provides new insights into the interaction of rewards and punishment, and how the latter can crowd out the former. It helps to explain why some papers fail to find much positive reciprocity, i.e. Offerman (2002) and Al-Ubaydli and Lee (2009), and provides a potential solution to the positive reciprocity puzzle, Orhun (2018). Intention-based reciprocity models are often criticised for being complex and having little predictive power due to multiple equilibria. This may not necessarily be a drawback, however, since reciprocity is complex by nature. Rather, its complexity makes it ideal for analyzing institutional design and incentive structures.

A. Appendix A: Proofs

A.1. Model

Proof of proposition 1. Define the local best response correspondence $r_{i,h} : \Delta^H \to \Delta(A_{i,h})$ by

$$r_{i,h}(\sigma) = \underset{x_{i,h} \in \Delta(A_{i,h})}{\arg\max} \; U_i\left(\sigma_i \backslash x_{i,h}, \sigma_j, \sigma_i | h\right)$$

and best response correspondence $r(\sigma) : \Delta^H \to \prod_{(i,h)\in N\times H} \Delta(A_{i,h})$ by

$$r(\sigma) = \prod_{(i,h)\in N\times H} r_{i,h}(\sigma)$$

As $\prod_{(i,h)\in N\times H} \Delta(A_{i,h})$ and Δ^H are topologically equivalent, we can define an equivalent function $\tilde{r} : \Delta^H \to \Delta^H$ and look for a fixed point. A fixed points under $\tilde{r}$ satisfy the RTE conditions since player (i,h) maximizes her utility, and first and second order beliefs are correct (and are updated along the path given h).

Kakutani's fixed point theorem applies in this setup. To see this, notice the local choice set $\Delta(A_{i,h})$ is compact, convex and non-empty. Next, $r_{i,h}$ is non-empty as U_i is continuous in (i,h)'s own choice $(x_{i,h})$, the set is compact and hence attains a maximum. $r_{i,h}$ is convex as U_i is indeed linear in (i,h)'s own choice. Upper hemi-continuity of $r_{i,h}$ follows from the fact that U_i is continuous (π_i, π_j, and κ_i are continuous).

Since these properties extend from $r_{i,h}$ to $\tilde{r}_{i,h}$ and $\tilde{r}$, all conditions of Kakutani's fixed point theorem are satisfied. It follows that an RTE exists. □

A.2. Trust

conRE - prisoner's dilemma.

For $\beta_2(c|C) = 0$ player 2 wants to cooperate if $U_2(c, \beta_2|C) = 1 + \gamma_2(2-0)\cdot 1 > 2 + \gamma_2(2-0)\cdot(-1) = U_2(d, \beta_2|C)$ - which is exactly the same inequality as in the RTE model. In contrast, she wants to defects when $\beta_2(c|C) \geq 1/2$, $U_2(c, \beta_2|C) = 1 < U_2(d, \beta_2|C) = 2$. It follows that for $\gamma_2 < 1/3$ the equilibrium is identical to my model as $\sigma_2(c|C) < 1/2$. Yet for a player 2 with $\gamma_2 \geq 1/3$ it must be that she cooperates with slightly less $1/2$ probability. To find the exact probability, a small technical adjustment needs to be introduced to ensure continuity at $\beta_2(c|C) = 1/2$. In particular, we take a

very small $\epsilon > 0$, such the kindness of C is

$$\kappa_1(C, \beta_2) = \begin{cases} 2 - \beta_2(c|C) & \text{if} \quad \beta_2(c|C) \leq 1/2 - \epsilon \\ \left(\frac{3}{2} + \epsilon\right) \frac{1/2 - \beta_2(c|C)}{\epsilon} & \text{if} \quad 1/2 - \epsilon < \beta_2(c|C) < 1/2 \\ 0 & \text{if} \quad \beta_2(c|C) \geq 1/2 \end{cases}$$

Derivation for interior part: Let $f(b)$ at b, $f(1/2) = 0$. To connect the two for $x \in [b, 1/2]$ take $f^c(x) = f(b) - (x - b)f(b)/(1/2 - b)$. And plug in.

The exact equilibrium probability depends on how we close the discontinuity in the kindness function. In equilibrium $\left(\frac{3}{2} + \epsilon\right) \frac{1/2 - \beta_2(c|C)}{\epsilon} 2 = 1$ or $\beta_2(c|C) = \frac{1}{2} - \frac{\epsilon}{3+2\epsilon}$ which goes to $1/2$ as $\epsilon \to 0$.

A.2.1. ESSENTIAL LEMMAS AND PROPERTIES OF $\mathbf{PD}(\cdot)$

Before proceeding to the proofs of this section, it is useful to establish some properties of the $\mathbf{PD}(\cdot)$ operator and the respective efficient sets CE_1 and TE_1.

Lemmas for conditional efficiency, $CE_1(\beta_2)$.

Lemma 1: $\mathbf{PD}(\beta_2)$ *is transitive.*

Proof of lemma 1. If a_1 $\mathbf{PD}(\beta_2)$ a_1' and a_1' $\mathbf{PD}(\beta_2)$ a_1'' then $\pi_k(a_1, \beta_2) \geq \pi_k(a_1', \beta_2)$ and $\pi_k(a_1', \beta_2) \geq \pi_k(a_1'', \beta_2)$ for all k with strict inequalities for some. Consequently $\pi_k(a_1, \beta_2) \geq \pi_k(a_1'', \beta_2)$ for all k with strict inequalities for some. □

Lemma 2: *The conditionally efficient action* $M_1^{CE_1(\beta_2)}$ *that minimizes 2's payoffs,* $M_1^{CE_1(\beta_2)} \in \arg\min_{a_1 \in CE_1(\beta_2)} \pi_2(a_1, \beta_2)$*, also maximizes 1's payoffs,* $M_1^{CE_1(\beta_2)} \in \arg\max_{a_1 \in CE_1(\beta_2)} \pi_2(a_1, \beta_2)$.

Proof of lemma 2. Suppose it doesn't, that is there is some $a_1 \in A_1$ that is better for player 1, while not being worse for player 2. This would imply that a_1 $\mathbf{PD}(\beta_2)$ $M_1^{CE_1(\beta_2)}$, leading to the contradiction that $M_1^{CE_1(\beta_2)}$ is not conditionally efficient, $M_1^{CE_1(\beta_2)} \notin CE_1(\beta_2)$. □

Lemma 3: *If* $a_1 \notin CE_1(\beta_2)$*, then there* $\exists a_1' \in CE_1(\beta_2)$ *that* a_1' $\mathbf{PD}(\beta_2)$ a_1.

Proof of lemma 3. By definition of being dominated, there must exist $a_1' \in A_1$ that a_1' $\mathbf{PD}(\beta_2)$ a_1. If a_1' itself is not efficient, $a_1' \notin CE_1(\beta_2)$, then there must be an action $a_1'' \in A_1$ that a_1'' $\mathbf{PD}(\beta_2)$ a_1'. Since $\mathbf{PD}(\beta_2)$ is a transitive operator a_1'' $\mathbf{PD}(\beta_2)$ a_1. If a_1'' is also not efficient, repeat the argument.

As there are a finite amount of actions, and thus only a finite amount of in-efficient actions, it must be that there exist some $a_1''' \in CE_1(\beta_2)$ that a_1''' $\mathbf{PD}(\beta_2)$ a_1. □

Lemmas for trust-efficiency, $TE_1(\beta_2)$.

Lemma 4: $\mathbf{PD}(\beta_2^{TE})$ *is transitive.*

Proof of lemma 4. Suppose $a_1', a_1'', a_1''' \in A_1$, a_1' $\mathbf{PD}(\beta_2^{TE})$ a_1'' and a_1'' $\mathbf{PD}(\beta_2^{TE})$ a_1'''. If no action is followed by generous responses, the operator is identical to $\mathbf{PD}(\beta_2)$, which is transitive. If any of the actions is followed by generous responses, the material best response is used. Denote the payoffs vector by $\pi = (\pi_1, \pi_2)$ and let I_{a_1} be the indicator function that takes value of 1 if $a_1 \in A_1$ is followed by a generous response. Write a_1' $\mathbf{PD}(\beta_2^{TE})$ a_1'' as $I_{a_1'}\pi(a_1', \sigma_2^{mBR}) + (1 - I_{a_1'})\pi(a_1', \beta_2) \geq I_{a_1''}\pi(a_1'', \sigma_2^{mBR}) + (1 - I_{a_1''})\pi(a_1'', \beta_2)$ and a_1'' $\mathbf{PD}(\beta_2^{TE})$ a_1''' as $I_{a_1''}\pi(a_1'', \sigma_2^{mBR}) + (1 - I_{a_1''})\pi(a_1'', \beta_2) \geq I_{a_1'''}\pi(a_1''', \sigma_2^{mBR}) + (1 - I_{a_1'''})\pi(a_1''', \beta_2)$ which shows that a_1' $\mathbf{PD}(\beta_2^{TE})$ a_1''' (clearly, any respective strict inequality remains strict). □

Lemma 5: *If* $a_1 \notin TE_1(\beta_2)$*, then there* $\exists a_1' \in TE_1(\beta_2)$ *that* a_1' $\mathbf{PD}(\beta_2^{TE})$ a_1.

Proof of lemma 5. Repeat proof of lemma 3 together fact that $\mathbf{PD}(\beta_2^{TE})$ is transitive by lemma 4. □

A.2.2. Proofs for trust-section

Lemma 6: *Let* $M_1 := \arg\min_{a_1 \in E_1 \subseteq A_1} \pi_2(a_1, \beta_2^*)$*. In any reciprocity equilibrium based on* E_1*,* β_2^* *cannot attach a positive probability to any generous action after* $a_1 \in A_1$ *if* $\pi_2(a_1, \beta_2^*) \leq \pi_2(M_1, \beta_2^*)$.

The lemma applies for RTE, conRE, and DK04. In equilibrium, any action that induces a payoff that is (weakly) lower than the lowest efficient payoff cannot be kind, and hence player 2 must respond either by a material best-response or a punishing action.

Proof of Lemma 6. Suppose β_2^* attaches positive probability to the generous action $\tilde{a}_2$ after a_1, $\beta_2^*(\tilde{a}_2|a_1) > 0$. Since a_1 yields less than the minimum efficient payoff, $\pi_2(a_1, \beta_2^*) \leq \pi_2(M_1, \beta_2^*)$, it must be that $\kappa_1(a_1, \beta_2^*) \leq \kappa_1(M_1, \beta_2^*) \leq 0$. As a result player 2 prefers the material best-response over $\tilde{a}_2$: $U_2(\tilde{a}_2, \beta_2^*|h = a_1) = \pi_2(a_1, \tilde{a}_2) + \gamma_2\kappa_1(a_1, \beta_2^*)\pi_1(a_1, \tilde{a}_2) < \pi_2(a_1, a_{2,a_1}^{mBR}) + \gamma_2\kappa_1(a_1, \beta_2^*)\pi_1(a_1, \tilde{a}_2) \leq \pi_2(a_1, a_{2,a_1}^{mBR}) + \gamma_2\kappa_1(a_1, \beta_2^*)\pi_1(a_1, a_{2,a_1}^{mBR}) = U_2(a_{2,a_1}^{mBR}, \beta_2^*|h = a_1)$. □

Lemma 7: *Let β_2 be an RTE-belief. If $a_1 \in TE_1(\beta_2)$ and a_1 $\mathbf{PD}(\beta_2^{TE})$ a_1' with $\pi_2(a_1', \beta_2) \leq \min_{x_1 \in TE_1(\beta_2)} \pi_2(x_1, \beta_2)$ then a_1 $\mathbf{PD}(\beta_2)$ a_1'.*

Proof of lemma 7. By lemma 6, 2 cannot respond with a generous action after a_1'. If 2 responds to a_1 either selfishly or with a punishing action, the statement is vacuously true - the payoff for each action is the same given β_2 and β_2^{TE}. If 2 responds with a generous action, a_1 $\mathbf{PD}(\beta_2^{TE})$ a_1' implies that $\pi_1(a_1, \beta_2) > \pi_1(a_1, \sigma_2^{mBR}) \geq \pi_1(a_1', \beta_2)$. Combine lemma 6, $\pi_2(a_1, \beta_2) > \min_{x_1 \in TE_1(\beta_2)} \pi_2(x_1, \beta_2)$, as otherwise player 2 wouldn't want to take a generous action, together with the assumption $\min_{x_1 \in TE_1(\beta_2)} \pi_2(x_1, \beta_2) \geq \pi_2(a_1', \beta_2)$ to get the result. □

Corollary 1: *Let β_2 be an RTE belief. Then $\min_{a_1 \in TE_1(\beta_2)} \pi_2(a_1, \beta_2) \leq \min_{a_1 \in CE_1(\beta_2)} \pi_2(a_1, \beta_2)$.*

Proof of corollary 1. By lemma 7 any action $a_1' \notin TE_1(\beta_2)$ that gives player 2 less than her minimum efficient payoff $\min_{a_1 \in TE_1(\beta_2)} \pi_2(a_1, \beta_2)$ cannot be in $CE_1(\beta_2)$. But then $\min_{a_1 \in CE_1(\beta_2)} \pi_2(a_1, \beta_2)$ must be weakly larger. □

Lemma 8: *Let β_2 be an conditional Reciprocity equilibrium belief. Then $\min_{a_1 \in TE_1(\beta_2)} \pi_2(a_1, \beta_2) \leq \min_{a_1 \in CE_1(\beta_2)} \pi_2(a_1, \beta_2)$.*

Proof of lemma 8. Suppose $M_1^{CE_1(\beta_2)}$ induces a lower payoff than $M_1^{TE_1(\beta_2)}$. By lemma 5, there exists an action $a_1 \in TE_1(\beta_2)$ that a_1 $\mathbf{PD}(\beta_2^{TE})$ $M_1^{CE_1(\beta_2)}$. Moreover a_1 must be followed by a generous response as otherwise a_1 $\mathbf{PD}(\beta_2)$ $M_1^{CE_1(\beta_2)}$, which would imply that $M_1^{CE_1(\beta_2)} \notin CE_1(\beta_2)$ (note that lemma 6 requires that 2 cannot be generous after $M_1^{CE_1(\beta_2)}$). Using both observations together, a_1 must satisfy $\pi_1(a_1, \beta_2) > \pi_1(a_1, \sigma_2^{mBR}) > \pi_1(M_1^{CE_1(\beta_2)}, \beta_2)$ and $\pi_2(a_1, \sigma_2^{mBR}) > \pi_2(M_1^{CE_1(\beta_2)}, \beta_2) > \pi_2(a_1, \beta_2)$. But since $\pi_2(M_1^{TE_1(\beta_2)}, \beta_2) > \pi_2(M_1^{CE_1(\beta_2)}, \beta_2)$ it cannot be that $M_1^{TE_1(\beta_2)}$ induces the minimum payoff given $TE_1(\beta_2)$. □

Proof of proposition 2. Follows directly from corollary 1 and lemma 8. □

Proof of proposition 3. Since a_1 $\mathbf{PD}(\beta_2^*)$ a_1' a_1' must induce the minimum payoff. By lemma 6, it must be that player 2 (i) takes the material best-response after a_1', in which case $\pi_1(a_1, \sigma_2^{mBR}) < \pi_1(a_1', \beta_2) = \pi_1(a_1', \sigma_2^{mBR})$, or (ii) punishes (possibly mixing over punishment and material best-responses), which yields $\pi_1(a_1, \sigma_2^{mBR}) < \pi_1(a_1', \beta_2) < \pi_1(a_1', \sigma_2^{mBR})$. Lastly, by dominance of a_1, it is clearly true that there exist some payoff $\pi_1(a_1, a_2) > \pi_1(a_1', \beta_2)$. □

Proof of proposition 4. Identical to the proof of $|A_1| = 2$. By lemma 6, it must be that player 2 either (i) takes the material best-response after $M_1^{TE_1(\beta_2)}$, in which case $\pi_1(a_1, \sigma_2^{mBR}) < \pi_1(M_1^{TE_1(\beta_2)}, \beta_2) = \pi_1(M_1^{TE_1(\beta_2)}, \sigma_2^{mBR})$, or (ii) punishes (possibly mixing over punishment and material best-responses), which yields $\pi_1(a_1, \sigma_2^{mBR}) < \pi_1(M_1^{TE_1(\beta_2)}, \beta_2) < \pi_1(M_1^{TE_1(\beta_2)}, \sigma_2^{mBR})$. Since it a_1 Pareto dominates $M_1^{TE_1(\beta_2)}$ given β the second condition is also satisfied. □

Proof of proposition 5. First, I show that $M_1^{CE_1(\beta_2^*)}$ **PD**(β_2^*) $M_1^{TE_1(\beta_2^*)}$.

If $|A_1| = 2$ and thus $|CE_1(\beta_2^*)| = 1$, clearly $M_1^{CE_1(\beta_2^*)}$ **PD**(β_2^*) $M_1^{TE_1(\beta_2^*)}$ as it is the only conditionally efficient action, and thus, by definition, must Pareto-dominate all other actions.

If $|CE_1(\beta_2^*)| \geq 2$, suppose it is not true that $M_1^{CE_1(\beta_2^*)}$ **PD**(β_2^*) $M_1^{TE_1(\beta_2^*)}$. In this case there must exist (at least) another action $a_1 \in CE_1(\beta_2^*)$ (lemma 3) that satisfies a_1 **PD**(β_2^*) $M_1^{TE_1(\beta_2^*)}$. Since $M_1^{CE_1(\beta_2^*)}$ induces player 2's minimum payoff in $CE_1(\beta_2^*)$, it must be that $\pi_2(M_1^{CE_1(\beta_2^*)}, \beta_2^*) < \pi_2(a_1, \beta_2^*)$ as well as $\pi_1(M_1^{CE_1(\beta_2^*)}, \beta_2^*) > \pi_1(a_1, \beta_2^*)$ (lemma 2). But since $\pi_1(a_1, \beta_2^*) > \pi_1(M_1^{TE_1(\beta_2^*)}, \beta_2^*)$ and $\pi_2(M_1^{CE_1(\beta_2^*)}, \beta_2^*) > \pi_2(M_1^{TE_1(\beta_2^*)}, \beta_2^*)$ (proposition 2), it follows that $M_1^{CE_1(\beta_2^*)}$ **PD**(β_2^*) $M_1^{TE_1(\beta_2^*)}$.

Finally for $M_1^{TE_1(\beta_2^*)} \in TE_1(\beta_2^*)$, β_2^* must assign positive probability to a generous action after $M_1^{CE_1(\beta_2^*)}$ as otherwise $M_1^{CE_1(\beta_2^*)}$ **PD**(β_2^{TE}) $M_1^{TE_1(\beta_2^*)}$. By lemma 6, this cannot occur in a conditional Reciprocity Equilibrium. □

A.3. Dufwenberg and Kirchsteiger '04

Proof of proposition 6. Since unconditional efficiency doesn't just require an action to be Pareto-dominated for some response (β_2), but for all responses, it must be that $TE_1(\beta) \subseteq UE_1$. Next observe that $\min_{a_1 \in X} \pi_2(a_1, \beta_2)$ is (weakly) lower the larger the set X. □

Lemma 9: *When player 1 has only 2 actions, $A_1 = \{a_1, a_1'\}$, and assumption 1 holds, an RTE exists.*

Proof of lemma 9. Suppose $\pi_2(a_1, \sigma_2^{mBR}) > \pi_2(a_1', \sigma_2^{mBR})$. Eliminate all punishing actions from player 2's set of actions after a_1, that is $A_{2,a_1}^{new} = A_{2,a_1} \setminus A_{2,a_1}^P$. By proposition 1, an equilibrium still exists for this restricted set of actions. Moreover, in any RTE, player 2 cannot take a generous action after a_1' (or be considered kind). To see this, suppose she does. Since only one of the two actions can be kind, this would imply that player 2 takes her material best-response action after a_1 for sure. But since any generous action $a_2 \in A_{2,a_1'}^G$ leads to a payoff of $\pi_2(a_1', a_2) < \pi_2(a_1', \sigma_2^{mBR}) < \pi_2(a_1, \sigma_2^{mBR})$, it must be that a_1' is perceived as unkind. As a result, she must either punish or take her material best-response after a_1'. Note further that, in equilibrium, a_1 is either kind or is at least not unkind. This

equilibrium is also an RTE for the unrestricted set of actions. Since a_1 is not unkind, player 2 prefers to take a generous or material best-response over a punishing action. The existence of punishing has no impact on the equilibrium responses. □

Proof of proposition 7. By lemma 6, player 2 cannot respond generously after a_1', hence $\pi_k(a_1', \beta_2^*) \leq \pi_k(a_1', \sigma_2^{mBR})$ for all k. Since a_1 is the only efficient action, it must be that $\pi_k(a_1, \beta_2^*) = \pi_k(a_1, \sigma_2^{mBR})$ for all k. Case (1) thus follows immediately if player 2 doesn't punish after a_1', as in this case a_1 **PD**(β_2^*) a_1'. If player 2 punishes after a_1' then by assumption 1 it must be that $\pi_2(a_1', \sigma_2^{mBR}) < \pi_2(a_1, \sigma_2^{mBR})$. If $\pi_1(a_1', \sigma_2^{mBR}) < \pi_1(a_1, \sigma_2^{mBR})$ then we are still in case (1); if instead $\pi_2(a_1', \sigma_2^{mBR}) > \pi_2(a_1, \sigma_2^{mBR})$ we are in case (2). □

A.4. Applications

For this section assume $\lambda = 1/2$.

Game 7.

RTE: If $\beta_2 = cp$ then 2 punishes after S if $U_2(c, \beta_2 = cp|S) = 6 + \gamma_2(5-14)11 \leq 5 + \gamma_2(5-14)5 = U_2(p, \beta_2 = cp|S)$ or $\gamma_2 \geq 1/54$. It is cheap to punish and S is rather unkind. Notice that believing that 2 punishes, makes S appear less kind, and punishment easier to sustain. If we start with selfish beliefs, $\beta_2 = cc$, punishment requires $U_2(c, \beta_2 = cc|S) = 6 + \gamma_2(6 - (14+6)/2)11 \leq 5 + \gamma_2 \cdot (-4) \cdot 5 = U_2(p, \beta_2 = cp|S)$ or $\gamma_2 \geq 1/24$. Thus for any $\gamma_2 > 1/24$, punishment is the unique equilibrium belief, and there are multiple equilibria for $\gamma_2 \in [1/54, 1/24]$. If player 2 punishes her optimal choice after H is c.

DK04: player 2 rewards H and punishes S, or always acts neutral.

Suppose $\beta_2 = cc$, then 2 reciprocates after H in DK04 if $U_2(r, \beta_2 = cc|H) = 13 + \gamma_2(14-(14+6)/2)12 \leq 14 + \gamma_2(4)8 = U_2(c, \beta_2 = cc|H)$ or $\gamma_2(4)4 \geq 1$, and punishes after S if $U_2(p, \beta_2 = cc|S) = 5 + \gamma_2(6 - (14+6)/2)7 \leq 6 + \gamma_2(-4)11 = U_2(c, \beta_2 = cc|S)$ and hence again $\gamma_2(4)4 \geq 1$. Clearly, the same holds true if $\beta_2 = rp$, where $\kappa_1(H, \beta_2 = rp) = 13-(13+5)/2 = 4 = -\kappa_1(S, \beta_2 = rp) = -(5-(13+5)/2) = 4$. The symmetry in responses clearly only holds when $\lambda = 1/2$ - which is assumed in all papers I am aware off.

Game 10.

Suppose $\gamma_2 = 1$. Notice that even for the kindest beliefs (lowest max, highest min), $\beta_2 = r^+rn$, 2 wants to punish after S, $U_2(p, \beta_2|S) = \gamma_2(1-(2+1)/2)\cdot 1 = -\gamma_2/2 > U_2(n, \beta_2|S) = 1+\gamma_2(1-(2+1)/2)4 = 1-2\gamma_2$ (This true for $\gamma > 0.4$).

When 2 punishes after S, N **PD**$(\cdot\cdot p)$ S and so conRE and RTE opt for the selfish response after N. For unRE, 2 reciprocates after N as $U_2(r, r^+rp|N) = 1.5+\gamma_2(1.5-(2+0)/2)5 > U_2(n, r^+rp|N) = 2+\gamma_2(1.5-(2+0)/2)3$ or $\gamma_2 \geq 1/2$.

Finally after K, 2 reciprocates strongly for unREL: $U_2(r^+, r^+rp|K) = 2+\gamma_2(2-(2+0)/2)5 = 2+5\gamma_2$, $U_2(r, r^+rp|K) = 2.5+4\gamma_2$, and $U_2(n, r^+rp|K) = 3+2\gamma_2$, that is she prefers r^+ for $\gamma_2 \geq 1/2$, r for $1/2 > \gamma_2 \geq 1/4$.

For RTE $U_2(r^+, rnp|K) = 2+\gamma_2(2.5-(2.5+2)/2)5 = 2+\gamma_2 5/4$, $U_2(r, rnp|K) = 2.5+\gamma_2$, and $U_2(n, rnp|K) = 3+\gamma_2/2$ and so she prefers r^+ only for $\gamma_2 \geq 2$ (recall $\gamma_2 = 1$ was assumed), and prefers for r for $2 > \gamma_2 \geq 1$. Note that indifference is always broken in favor of the more extreme as any randomization increases perceived kindness of K.

The Equilibrium behavior for conRE follows the solution from the sequential prisoners's dilemma. The key ingredient, again, is to smooth out the discontinuity.

A.5. Summary of Equilibria

Proof of proposition 8. By proposition 6 we know that that $\min_{a_1 \in UE_1} \pi_2(a_1, \beta_2^*) \leq \min_{a_1 \in TE_1(\beta_2^*)} \pi_2(a_1, \beta_2^*)$. Moreover by lemma 8, it must also be that $\min_{a_1 \in TE_1(\beta_2^*)} \pi_2(a_1, \beta_2^*) \leq \min_{a_1 \in CE_1(\beta_2^*)} \pi_2(a_1, \beta_2^*)$.

If $\min_{a_1 \in CE_1(\beta_2^*)} \pi_2(a_1, \beta_2^*) = \min_{a_1 \in UE_1} \pi_2(a_1, \beta_2^*)$, together with the observation that $\min_{a_1 \in TE_1(\beta_2^*)} \pi_2(a_1, \beta_2^*)$ is sandwiched in between the two, the minimizing action must be identical in all three. In that case, preferences in all three models are identical, and $(\sigma^*, \alpha^*, \beta^*)$ is an RTE. Note that there is no need to look at player 1. Player 2's efficient set $E_2(h)$ at $h \in H$ represent a simple decision problems and thus all three efficiency notions coincide, leading to the same preferences for player 1 given 2's identical response across models.

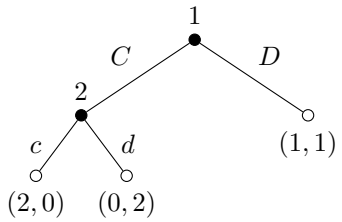
If instead, $\min_{a_1 \in CE_1(\beta_2^*)} \pi_2(a_1, \beta_2^*) < \min_{a_1 \in UE_1} \pi_2(a_1, \beta_2^*)$, yet player 2 prefers the same actions, then she must also prefer the same action given a reference point in between. If this isn't immediately obvious, simply take the utility difference of any $a_2, a_2' \in A_2(h)$, which can be written as $\kappa_1(a_1, \beta_2^*)(\pi_1(a_2) - \pi_1(a_2')) \geq \pi_2(a_2') - \pi_2(a_2)$. If this inequality holds for two different kindness levels, it must also hold for some convex combination of the two. □

B. Appendix B: Further Detail

B.1. RTE is an Equilibrium Concept

Game 11 highlights why imposing equilibrium is often necessary for models with second order beliefs. If player 2 is sufficiently reciprocal, she wants to play c if $\beta_2 = d$. In contrast, when $\beta_2 = c$, then she clearly wants to play d. But this indicates that both combinations of action and belief are rationalizable. Player 1 can think she plays c for sure since he thinks she thinks $\beta_2 = d$, and vice versa. Most importantly, this example shows that player 1 can hold a belief that player 2 reciprocates, minimizing her payoff in the process.

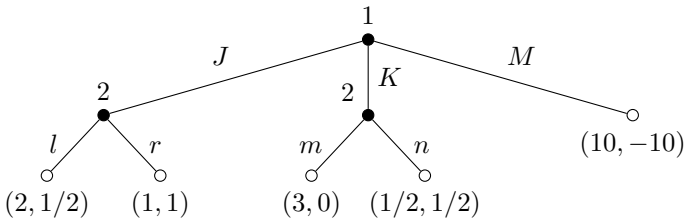
Clearly, this can never be an equilibrium belief and isn't very sensible. But without imposing a more restrictive utility function, such beliefs are indeed rationalizable. This suggests that imposing equilibrium is often needed - which I have done throughout the paper. It should be clear that when there are more than two choices, even more behavior can be rationalizable.



Game 11

B.2. Relationship Between Efficient Sets

Game 12 highlights that it does not need to be true that $CE_1(\beta_2) \subseteq TE_1(\beta_2)$, even if β_2 is a RTE-belief. Clearly $\beta_2 = lm$ is an equilibrium (M is always efficient, the minimum payoff -10). $CE_1(lm) = \{J, K, M\}$. However, given the selfish responses, $\beta_2^{TE} = rn$, J **PD**(rn) K, so that $TE_1(lm) = \{J, M\}$.



Game 12

B.3. Dufwenberg and Kirchsteiger '04, $|A_1| \geq 2$

If player 1's action set is finite, $|A_1| \geq 2$, it is helpful to split up the proposition into two ideas.

Proposition 9: *Let* $(\sigma^*, \alpha^*, \beta^*)$ *be an RTE. If* $M_1^{TE_1(\beta_2^*)} \neq M_1^{UE_1}$ *then* $M_1^{TE_1(\beta_2^*)}$ **PD**(β_2^*) $M_1^{UE_1}$.

This result highlights that whenever the reference point in DK04 differs from my model, the action that induces the minimum payoff under trust-efficiency Pareto-dominates the respective action that induces the minimum payoff in theirs.

The intuition for this proposition is as follows: $M_1^{TE_1}$ can be part of the reference point for two reason: (1) It minimizes 2's payoff while maximizing 1's payoff. When $M_1^{UE_1}$ leads to even lower payoffs for 2, the only way to not be dominated is by being even better for player 1 than $M_1^{TE_1}$. (2) If $M_1^{TE_1}$ doesn't maximize 1's payoff, it is actually dominated by some other action, but remains efficient due to trust. As $M_1^{UE_1}$ leads to lower payoffs for player 2, it would also be efficient due to trust if it were to make player 1 better off than $M_1^{TE_1}$. But since none of the two cases are true, it must be Pareto-dominated by $M_1^{TE_1}$.

Proof of proposition 9. When the efficient set is a singleton, that is $TE_1(\beta_2^*) = \{M_1^{TE_1(\beta_2^*)}\}$, player 2 must respond with her material-best response to $M_1^{TE_1(\beta_2^*)}$, that is $\pi_k(M_1^{TE_1(\beta_2^*)}, \beta_2^*) = \pi_k(M_1^{TE_1(\beta_2^*)}, \sigma_2^{mBR})$ for all k, and cannot act generously to any other action (lemma 6). In this case, conditional-efficiency and trust-efficiency coincide. By lemma 3, it follows that any non-conditionally efficient action must be Pareto-dominated by $M_1^{TE_1(\beta_2^*)}$ given β_2^*.

If $|TE_1(\beta_2^*)| \geq 2$ then $M_1^{TE_1(\beta_2^*)}$ is either conditionally efficient (β_2^*) or trust-efficient (β_2^{TE}).

In the first case, $M_1^{TE_1(\beta_2^*)} \in CE_1(\beta_2^*)$, by lemma 2, $M_1^{TE_1(\beta_2^*)}$ must be the action that yields player 1 his highest payoffs. Given that $M_1^{UE_1}$ induces an even lower payoffs for player 2, the only way for $M_1^{UE_1}$ to not be Pareto-dominated is when $\pi_1(M_1^{UE_1}, \beta_2^*) > \pi_1(M_1^{TE_1(\beta_2^*)}, \beta_2^*)$, in which case $M_1^{UE_1} \in CE_1(\beta_2^*)$, a violation.

In the second case, let a_1 be the action that a_1 $\mathbf{PD}(\beta_2^*)$ $M_1^{TE_1(\beta_2^*)}$ (but not using β_2^{TE}). a_1 has the property that

$\pi_1(a_1, \beta_2^*) > \pi_1(M_1^{TE_1(\beta_2^*)}, \beta_2^*) > \pi_1(a_1, \sigma_2^{mBR})$ and $\pi_2(a_1, \sigma_2^{mBR}) > \pi_2(a_1, \beta_2^*) > \pi_2(M_1^{TE_1(\beta_2^*)}, \beta_2^*)$. Hence if $M_1^{UE_1}$ isn't Pareto-dominated by $M_1^{TE_1(\beta_2^*)}$ given β_2^* then it is not β_2^{TE}-dominated by a_1 either as $\pi_1(M_1^{UE_1}, \beta_2^*) > \pi_1(M_1^{TE_1(\beta_2^*)}, \beta_2^*)$.

Moreover, there does not exists another action a_1' that a_1' $\mathbf{PD}(\beta_2^{TE})$ $M_1^{UE_1}$. Suppose there is, then by lemma 5, $a_1' \in TE_1(\beta_2^*)$.

Suppose first that a_1' is not followed by any generous action, so that the beliefs for 2's action after a_1', $\beta_2^{TE}(\cdot|a_1')$ and $\beta_2(\cdot|a_1')$ coincide. Moreover, since $M_1^{UE_1}$ induces a lower payoff for player 2 than $M_1^{TE_1(\beta_2^*)}$ and $M_1^{TE_1(\beta_2^*)}$ represents player 2's minimum efficient payoff, β_2^{TE} changes nothing (relative to β_2^*) after these actions either. a_1' then satisfies $\pi_k(a_1', \beta_2^*) > \pi_k(M_1^{UE_1}, \beta_2^*)$ for all k and therefore $\pi_1(a_1', \beta_2^*) > \pi_1(M_1^{UE_1}, \beta_2^*) > \pi_1(M_1^{TE_1(\beta_2^*)}, \beta_2^*)$. If $\pi_2(a_1', \beta_2^*) > \pi_2(M_1^{TE_1(\beta_2^*)}, \beta_2^*)$ then a_1' $\mathbf{PD}(\beta_2^{TE})$ $M_1^{TE_1(\beta_2^*)}$. If instead $\pi_2(a_1', \beta_2^*) < \pi_2(M_1^{TE_1(\beta_2^*)}, \beta_2^*)$ then $M_1^{TE_1(\beta_2^*)}$ doesn't induce the minimum efficient payoff. Both are contradictions.

Next, suppose a_1' is followed by a generous action. a_1' now satisfies $\pi_k(a_1', \sigma_2^{mBR}) > \pi_k(M_1^{UE_1}, \beta_2^*)$ for all k and thus $\pi_1(a_1', \beta_2^*) > \pi_1(a_1', \sigma_2^{mBR}) > \pi_1(M_1^{UE_1}, \beta_2^*) > \pi_1(M_1^{TE_1(\beta_2^*)}, \beta_2^*)$. If $\pi_2(a_1', \beta_2^*) > \pi_2(a_1', M_1^{TE_1(\beta_2^*)})$ then a_1' $\mathbf{PD}(\beta_2^{TE})$ $M_1^{TE_1(\beta_2^*)}$ as $\pi_2(a_1', \sigma_2^{mBR}) > \pi_2(a_1', \beta_2^*)$. If instead $\pi_2(a_1', \beta_2^*) < \pi_2(a_1', M_1^{TE_1(\beta_2^*)})$ then $M_1^{TE_1(\beta_2^*)}$ doesn't induce the minimum payoff. It follows that a_1' cannot exist, but then $M_1^{UE_1} \in TE_1(\beta_2^*)$, a violation. It follows that $M_1^{TE_1(\beta_2^*)}$ $\mathbf{PD}(\beta_2^*)$ $M_1^{UE_1}$. □

As in the $|A_1| = 2$ case, I need to make some assumptions with regards to punishment. When there are more than 2 actions, the multiple equilibrium problem becomes even more involved: Player 2 may react very differently to different unkind actions depending on whether she thinks he thinks she punishes.

Assumption 2: *Take any $a_1, a_1' \in A_1$. If player 2 has a punishing action $p \in A_{2,a_1}$ after a_1 then she also has the same punishing action available to her after a_1'. That is there exists a $p' \in A_{2,a_1'}$ with $\pi_1(a_1', p') - \pi_1(a_1', \sigma_2^{mBR}) = \pi_1(a_1, p) - \pi_1(a_1, \sigma_2^{mBR}) < 0$ and $\pi_2(a_1, p) - \pi_2(a_1, \sigma_2^{mBR}) = \pi_2(a_1', p') - \pi_2(a_1', \sigma_2^{mBR}) < 0$.*

This first assumption simply ensures that player 2 always has the same punishment actions available to her.[38]

[38] Clearly, this is the strongest possible assumption and could be relaxed.

Assumption 3: *Suppose assumption 2 holds. If* $\pi_2(a_1, \sigma_2^{mBR}) \geq \pi_2(a_1', \sigma_2^{mBR})$ *for any* $a_1, a_1' \in A_1$ *then player 2 punishes player 1 less after* a_1 *than* a_1'*; That is if she takes* $p' \in A_{2,a_1'}$ *after* a_1' *then she doesn't take an action* $p \in A_{2,a_1}$ *that* $\pi_1(a_1, p) - \pi_1(a_1, \sigma_2^{mBR}) < \pi_1(a_1', p') - \pi_1(a_1', \sigma_2^{mBR})$.

While this assumption is written in terms of what player 2 does, it could equivalently be written in terms of what player 1, and thus what player 2 believes she does. The respective assumed behavior would follow.

Proposition 10: *Let* $(\sigma^*, \alpha^*, \beta^*)$ *be an RTE. If* $M_2^{TE_1(\beta_2^*)} \neq M_2^{UE_1}$ *then* $M_2^{TE_1(\beta_2^*)}$ **PD**(β_2^*) $M_2^{UE_1}$. *Moreover if assumption 2 and 3 holds, then one of the following holds:*

1. $\pi_1(M_1^{TE_1(\beta_2^*)}, \sigma_2^{mBR}) > \pi_1(M_1^{UE_1}, \sigma_2^{mBR})$ *and* $\pi_2(M_1^{TE_1(\beta_2^*)}, \sigma_2^{mBR}) > \pi_2(M_1^{UE_1}, \sigma_2^{mBR})$, *or*
2. $\pi_1(M_1^{TE_1(\beta_2^*)}, \sigma_2^{mBR}) < \pi_1(M_1^{UE_1}, \sigma_2^{mBR})$ *and* $\pi_2(M_1^{TE_1(\beta_2^*)}, \sigma_2^{mBR}) > \pi_2(M_1^{UE_1}, \sigma_2^{mBR})$.

The proposition mirrors the binary case. The key difference between the two settings is that after $M_2^{TE_1(\beta^*)}$, player 2 may actually punish now. This is the reason why we need a more complete assumptions on punishment choices and behavior than in the simple binary-case - where player 2 doesn't punish after the only efficient choice.

Proof of proposition 10. When the efficient set is a singleton, that is $TE_1(\beta_2^*) = \{M_1^{TE_1(\beta_2^*)}\}$, the proof is identical to the binary case, $|A_1| = 2$, except with references to assumptions 2 and 3.

If $|TE_1(\beta_2^*)| \geq 2$, then player 2 may punish after $M_1^{TE_1(\beta_2^*)}$. If she doesn't, repeat the argument of the binary case, $|A_1| = 2$. If she does then $\pi_k(M_1^{TE_1(\beta_2^*)}, \sigma_2^{mBR}) > \pi_k(M_1^{TE_1(\beta_2^*)}, \beta_2^*) > \pi_k(M_1^{UE_1}, \beta_2^*)$ for all k.As player 2 doesn't punish more when her selfish payoff is larger, assumption 3, together with the availability of punishment choices assumed in 2, leads to $\pi_2(M_1^{TE_1(\beta_2^*)}, \sigma_2^{mBR}) > \pi_2(M_1^{UE_1}, \sigma_2^{mBR})$ - otherwise she would have needed to punish $M_1^{UE_1}$ more at a larger cost to herself as she would never use 'inefficient' punishment. As before, the level of player 1's payoffs determines the case. □

References

T.K. Ahn, Myungsuk Lee, Lore Ruttan, and James Walker. Asymmetric payoffs in simultaneous and sequential prisoner's dilemma games. *Public Choice*, 132:27–46, 1999.

George A. Akerlof. Labor contracts as partial gift exchange. *The Quarterly Journal of Economics*, 97 (4):543–569, 1982.

Omar Al-Ubaydli and Min Sok Lee. An experimental study of asymmetric reciprocity. *Journal of Economic Behavior & Organization*, 72:738–749, 2009.

James Andreoni, William Harbaugh, and Lise Vesterlund. The carrot or the stick: Rewards, punishments, and cooperation. *American Economic Review*, 93(3):893–902, 2003.

Robert Aumann and Adam Brandenburger. Epistemic conditions for nash equilibrium. *Econometrica*, 63(5):1161–1180, 1995.

Pierpaolo Battigalli and Martin Dufwenberg. Dynamic psychological games. *Journal of Economic Theory*, 144:1–35, 2009.

Joyce Berg, John Dickhaut, and Kevin McCabe. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122 –142, 1995.

Truman F. Bewley. A depressed labor market as explained by participants. *The American Economic Review*, 85(2):250–254, 1995.

Felix Bierbrauer and Nick Netzer. Mechanism design and intentions. *Journal of Economic Theory*, 163:557–603, 2016.

Sally Blount. When social outcomes aren't fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes*, 63(2):131–144, 1995.

Friedel Bolle and Peter Ockenfels. Prisoners' dilemma as a game with incomplete information. *Journal of Economic Psychology*, 11:69–84, 1990.

Bogachan Celen, Andrew Schotter, and Mariana Blanco. On blame and reciprocity: Theory and experiments. *Journal of Economic Theory*, 169:62–92, 2017.

Gary Charness and Matthew Rabin. Understanding social preferences with simple tests. *The Quarterly Journal of Economics*, 117(3):817–869, 2002.

Kisuk Cho and Byung-Il Choi. Cross-society study of trust and reciprocity: Korea, japan and the u.s. *International Studies Review*, 3(2):31–43, 2000.

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

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

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

James C. Cox, Rudolf Kerschbamer, and Daniel Neururer. What is trustworthiness and what drives it? *Games and Economic Behavior*, 98:197–218, 2016.

Jason Dana, Daylian M. Cain, and Robyn M. Dawes. What you don't know won't hurt me: Costly (but quiet) exit in dictator games. *Organizational Behavior and Human Decision Processes*, 100 (2):193–201, 2006.

Jason Dana, Roberto A. Weber, and Jason Xi Kuang. Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80, 2007.

Martin Dufwenberg and Georg Kirchsteiger. A theory of sequential reciprocity. *Games and Economic Behavior*, 47:268–298, 2004.

Tore Ellingsen and Magnus Johannesson. Pride and prejudice: The human side of incentive theory. *American Economic Review*, 98(3):990–1008, 2008.

Armin Falk. Gift exchange in the field. *Econometrica*, 75(5):1501–1511, 2007.

Armin Falk and Urs Fischbacher. A theory of reciprocity. *Games and Economic Behavior*, 54:293–315, 2006.

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

Ernst Fehr and Armin Falk. Wage rigidity in a competitive incomplete contract market. *Journal of Political Economy*, 107(1):106–134, 1999.

Ernst Fehr and Simon Gächter. Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3):159–181, 2000.

Ernst Fehr and Simon Gächter. Do incentive contracts crowd out voluntary cooperation? *working paper*, 2001.

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

Ernst Fehr, Georg Kirchsteiger, and Arno Riedl. Does fairness prevent market clearing? an experimental investigation. *The Quarterly Journal of Economics*, 108(2):437–459, 1993.

John Geanakoplos, David Pearce, and Ennio Stacchetti. Psychological games and sequential rationality. *Games and Economic Behavior*, 1:60–79, 1989.

Robert A. Giacalone and Jerald Greenberg. *Antisocial behavior in organizations*. SAGE Publications, Thousand Oaks, California, 1997.

Farak Gul and Wolfgang Pesendorfer. Interdependent preference models as a theory of intentions. *Journal of Economic Theory*, 165:179–208, 2016.

Werner Güth, Rolf Schmittberger, and Bernd Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4):367–388, 1982.

Nahoko Hayashi, Elinor Ostrom, James Walker, and Toshio Yamagishi. Reciprocity, trust, and the sense of control - a cross-societal study. *Rationality and Society*, 11(1):27–46, 1999.

Menusch Khadjavi and Andreas Lange. Prisoners and their dilemma. *Journal of Economic Behavior & Organization*, 92:163–175, 2013.

Alan B. Krueger and Alexandre Mas. Strikes, scabs, and tread separations: Labor strife and the production of defective bridgestone/firestone tires. *Journal of Political Economy*, 112(2):253–289, 2004.

Mark T. Le Quement and Amrish Patel. Cheap talk as gift exchange. *working paper*, 2017.

David K. Levine. Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1:429–431, 1998.

Ulrike Malmendier and Klaus M. Schmidt. You owe me. *The American Economic Review*, 107(2): 493–526, 2017.

Ulrike Malmendier, Vera L. te Velde, and Roberto A. Weber. Rethinking reciprocity. *Annual Review of Economics*, 6:849–874, 2014.

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

Nick Netzer and Armin Schmutzler. Explaining gift-exchange – the limits of good intentions. *Journal of European Economic Association*, 12(6):1586–1616, 2014.

Theo Offerman. Hurting hurts more than helping helps. *European Economic Review*, 46(8):1423–1437, 2002.

A. Yesim Orhun. Perceived motives and reciprocity. *Games and Economic Behavior*, 2018. forthcoming.

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

Marcus Roel. Sequential reciprocity and incomplete information. *working paper*, 2018b.

Julio Rotemberg. Minimally acceptable altruism and the ultimatum game. *Journal of Economic Behavior & Organization*, 66(3-4):457–476, 2008.

Alexander Sebald. Attribution and reciprocity. *Games and Economic Behavior*, 68:339–352, 2010.

Joel Sobel. Interdependent preferences and reciprocity. *Journal of Economic Literature*, 43(2):392–436, 2005.

Motoki Watabe, Shigeru Terai, Nahoko Hayashi, and Toshio Yamagishi. Cooperation in the one-shot prisoner's dilemma based on expectations of reciprocity. *Japanese Journal of Experimental Social Psychology Quarterly*, 51:265–271, 1996.

Ernst Zermelo. Über eine Anwendung der Mengenlehre auf der Theorie des Schachspiels. *Proceedings of the Fifth International Congress on Mathematics*, 1913.