The London School of Economics and Political Science

Essays in Behavioral Economics

Marcus Roel

A thesis submitted to the Department of Economics of the London School of Economics for the degree of Doctor of Philosophy. London, May 2018.

To my wife and all my parents

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of approximately 48000 words.

Statement of conjoint work

I confirm that Chapter 3 was jointly co-authored with Manuel Staab and I contributed 50% of this work.

Acknowledgements

This thesis would have been impossible without the help and support of so many people.

First and foremost, my gratitude goes out to my two advisors, Erik Eyster and Francesco Nava, for their guidance and support throughout this long process. The time they spent on me went far beyond what one could ever dream for. Their insight, intellectual rigor, and curiosity not only shaped this very thesis but tremedously improved my understanding of economics and what it means to be a researcher. Thank you so much.

I would also like to thank Nava Ashraf, Björn Bartlin, Andrew Ellis, Wouter den Haan, Gilat Levy, Matthew Levy, Nick Netzer, Ronny Razin, Armin Schmutzler, Balazs Szentes, Séverine Toussaert, Roberto Weber as well as seminar participants at the London School of Economics for many helpful comments and suggestions.

I am grateful for the opportunity to be part of the STICERD Theory Group and to learn from and with its talented PhD students: Carlo Cabrera, Krittanai Laohakunakorn, Clement Minaudier, Francesco Sannino, Konstantinos Tokis, Lisa Windsteiger, among many others.

I want to thank my co-author and deskmate Manuel Staab for our productive and truly enjoyable collaboration on chapter 3, his humor, and his companionship during the PhD. I am grateful for Rafael Hortala-Vallve's encouragement, insights, and support during the Job Market. My eternal thanks go out to my great friend, colleague, and ex-flatmate Giuseppe Rossitti. I honestly cannot imagine how I would have survived my PhD without you.

I am forever indebted to my parents Regina, Klemens, Andreas and Claudia. Without your love and unconditional support over all these years, I could have never done it. Thank you for instilling in me a curiosity that led me to embark on this long, arduous, yet deeply gratifying journey. I must also thank my wonderful eight grandparents who supported me to follow my passion. Finally, I must thank the most wonderful, intelligent, and beautiful person in my world, my wife Shen Xin. It is you who made these last years so special and so easy when life got tough. Thank you for your endless kindness and unconditional love.

Abstract

This thesis contains two theoretical essays on reciprocity and one that analyzes the effects of perception biases on learning and decision-making.

In the first chapter, 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.

The second chapter extends my theory of reciprocity to incomplete information. I outline how reciprocity can give rise to pay-what-you-want pricing schemes. In the classic bilateral trade setting, I show that sequential interactions can be more efficient than normal form mechanisms when some people are motivated by reciprocity. Reciprocity creates incentives for information sharing.

The last chapter is co-authored with Manuel Staab. We study the effects of perception biases and incorrect priors on learning behavior, and the welfare ranking of information experiments. We find that both types of biases by themselves reduce expected utility in a model where payoff relevant actions also generate informative signals, i.e. when actions constitute information experiments. However, experiments can be affected to different degrees by these biases. We provide necessary and sufficient conditions for when any binary ranking of action profiles can be reversed. Building on these findings, we show that an agent can be better off suffering from both biases rather than just one.

Contents

List of Figures

Chapter 1

A Theory of Reciprocity with Trust

1.1 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 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.

Figure 1.1 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 1.1 on page 12, 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,

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

[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.

	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 1.1: Choice data from prisoner's dilemmas

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 and Schmidt argue that players feel obligated to reciprocate. Finally, figure 1.1 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?'

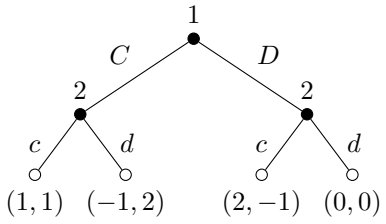
DK04, instead, suggest that a mutually beneficial action is kind as long as there exists *a strategy for player 2* for which player 1 is better off by taking the alternative choice.[4] It follows that only a mutually beneficial action that is also player 1's (payoff) dominant action cannot be considered 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

[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 1.1, 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.

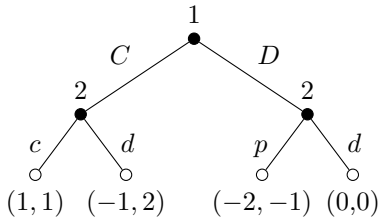
[4] In the prisoner's dilemma, this is satisfied for the strategies 'always cooperate', 'always defect', and 'defect after cooperation and cooperate after defection'.

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 1.1) to behavior in a prisoner's dilemma with punishment (game 1.2). 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.



Game 1.1: Sequential prisoner's dilemma



Game 1.2: 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

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

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).

The rest of this chapter is organized as follows: In the next section, I discuss the related literature. In section 1.3, I present my model and show that an equilibrium exists. Sections 1.4 and 1.5 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 1.6, 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 1.7. The paper ends with concluding remarks, section 1.8. Proofs, as well as the mathematical detail for most examples can be found in Appendix A.

1.2 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. 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

[6] A formal description of each reference point can be found in section 1.4 and 1.5.

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.

1.3 The Model

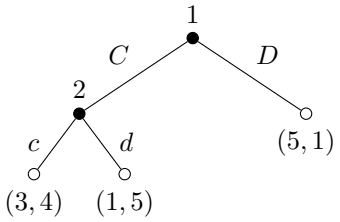
Basic idea of reciprocity. Reciprocity models allow for utility to depend on one's own as well on another player's payoff. In a two player game, player 2's utility takes the simple form of

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

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 1.3 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 1.3: A simple example of kindness

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

[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.

[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.

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 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.

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

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 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

[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.

[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.

(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. 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 1.2, 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 1.2. 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]

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.

[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 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 1.4.

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*

$$U_i\left(\sigma_i, \alpha_j, \beta_i | h\right) = \pi_i(\sigma_i, \alpha_j | h) + \gamma_i \cdot \kappa_j\left(\alpha_j, \beta_i | h\right) \cdot \pi_j(\sigma_i, \alpha_j | h) \tag{1.1}$$

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* $a_i^* \in \underset{a_i \in A_{i,h}}{\arg\max}\ U_i\left(\sigma_i^* \backslash a_i, \alpha_j^*, \beta_i^* | h\right)$
- $\alpha_i^* = \sigma_i^*$
- $\beta_i^* = \sigma_i^*$

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 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.

[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.

Example. Sequential prisoner's dilemma, game 1.1. 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 $\sigma_2(c|C) = \beta_2(c|C) = 2 - \frac{1}{2\gamma_2}$. 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.

1.4 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

[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.

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 1.3, 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.

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:

[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 1.10.1, game 1.11. The example underlines that RTE is best used as an equilibrium model.

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

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

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

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

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

[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.

[22] In general it does not need to be true that $CE_1(\beta_2) \subseteq TE_1(\beta_2)$. Game 1.12 in Appendix 1.10.2 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.

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.*

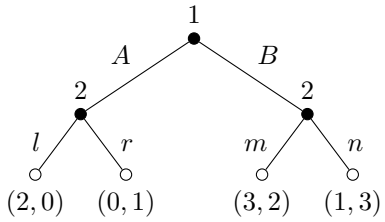
Whenever trust plays a role, an RTE cannot be an equilibrium in the Rabin model. To understand this proposition, recall that when player 2 conditionally cooperates in the prisoner's dilemma, cooperation is the only efficient choice for player 1 given the notion of conditional-efficiency. In this case, 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.

1.4.1 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 1.4 illustrates this idea.

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 1.4: Trust doesn't imply efficiency

1.5 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 1.3, but replaces trust-efficiency with unconditional efficiency.

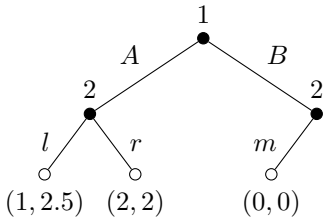
Returning once again to the prisoner's dilemma, it should be clear that both actions are unconditionally efficient, $UE_1 = \{C, D\}$. If player 2 always defects, C is better for 2, while D is better for 1. Consequently, the equilibrium predictions of RTE and unRE are identical. In general, unconditional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

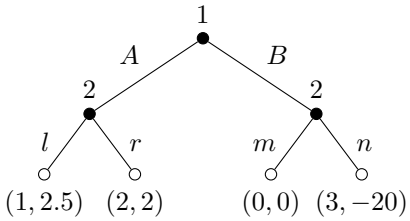
$$\min_{a_1 \in UE_1} \pi_2(a_1, \beta_2) \leq \min_{a_1 \in TE_1(\beta_2)} \pi_2(a_1, \beta_2).$$

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 1.5 and 1.6. In game 1.5, A Pareto-dominates B regardless of player 2's action; A is not kind. In game 1.6, 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 1.5



Game 1.6

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.

1.5.1 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.

[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).

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 1.6. 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 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.

[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.

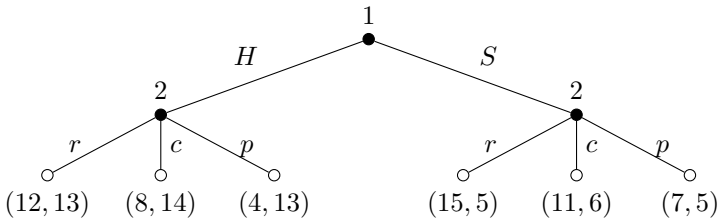
[25] While the assumption is written in terms of behavior, it could have also been written in terms of the equivalent beliefs, which implies such behavior.

[26] See Appendix A, Lemma 10 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'.

1.6 Applications

This section revisits games in the literature where player 2 can reward and punish. Game 1.7 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 1.7: Offerman (2002)

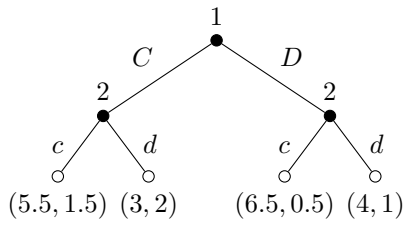
The key observation for this game is that when player 2 punishes after S, H Pareto dominates S for both $\beta_2 = rp$ and $\beta_2 = cp$. As a result, only H is trust-efficient and therefore not kind. Player 2's optimal response to H is c.

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. 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

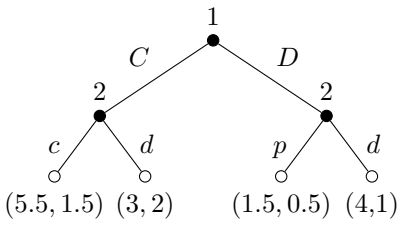
[28] Intentional harm hurts player 2's self-esteem and therefore induces punishment. However, she attributes being treated well by players or nature to being 'a good person deserves help'. As a result, there is no need to reciprocate.

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 1.8 and 1.9. In the game 1.8, player 2 has the usual option to cooperate, whereas in game 1.9 she can punish player 1 instead.



Game 1.8: Sequential prisoner's dilemma



Game 1.9: 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 1.8 to 35% in game 1.9.[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 (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]

[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.

[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.

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 1.10, 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.

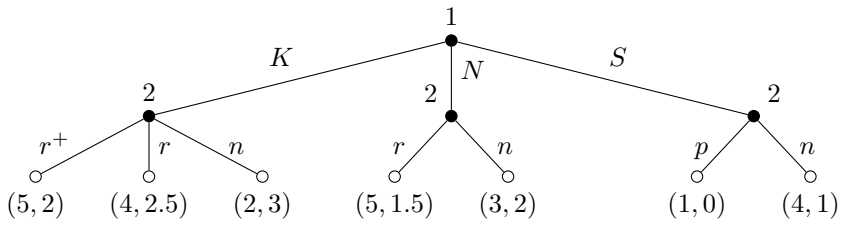
From our discussion in the previous paragraphs, it should be clear by now that this is exactly the behavior RTE predicts. When the most selfish offers are met with punishment, they become inefficient, giving rise to an increase in the reference point. As a result, more generous offers appear less kind 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 1.10, 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]

For a selfish player 1 and a reciprocal player 2, with $\gamma_2 = 1$, the RTE is (K, rnp), the unRE is

[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 after S.



Game 1.10

(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]

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

1.7 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 1.2 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 1.10 is an example of this.

1.8 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

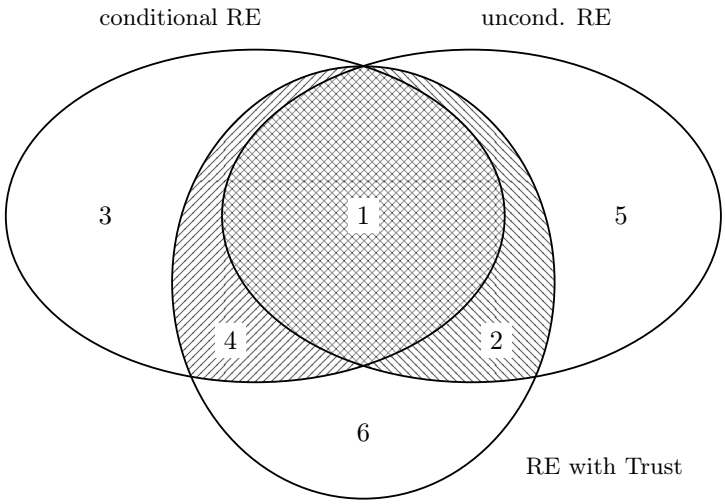


Figure 1.2: Equilibrium predictions overview

achieve this in the sequential version. This is consistent with experiments on the prisoner's dilemma, see figure 1.1. To 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

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

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, it would be interesting to test games like 1.5 and 1.6 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.

[37] Game 1.6 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 Chapter 2.

1.9 Appendix A: Proofs

1.9.1 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. □

1.9.2 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 $\epsilon \to 0$.

essential lemmas and properties of PD$(\cdot)$

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

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

Lemma 1: PD(β_2) *is transitive.*

Proof of lemma 1. If a_1 **PD**(β_2) a_1' and a_1' **PD**(β_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 **PD**(β_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' **PD**(β_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. □

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 8: *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 8. 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 9: *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 9. 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 8 and lemma 9. □

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 $M_1^{CE_1(\beta_2^*)}$ $\mathbf{PD}(\beta_2^*)$ $M_1^{TE_1(\beta_2^*)}$.
If $|A_1| = 2$ and thus $|CE_1(\beta_2^*)| = 1$, clearly $M_1^{CE_1(\beta_2^*)}$ $\mathbf{PD}(\beta_2^*)$ $M_1^{TE_1(\beta_2^*)}$ 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 $M_1^{CE_1(\beta_2^*)}$ $\mathbf{PD}(\beta_2^*)$ $M_1^{TE_1(\beta_2^*)}$. In this case there must exist (at least) another action $a_1 \in CE_1(\beta_2^*)$ (lemma 3) that satisfies a_1 $\mathbf{PD}(\beta_2^*)$ $M_1^{TE_1(\beta_2^*)}$. 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 $M_1^{CE_1(\beta_2^*)}$ $\mathbf{PD}(\beta_2^*)$ $M_1^{TE_1(\beta_2^*)}$.

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 $M_1^{CE_1(\beta_2^*)}$ $\mathbf{PD}(\beta_2^{TE})$ $M_1^{TE_1(\beta_2^*)}$. By lemma 6, this cannot occur in a conditional Reciprocity Equilibrium. □

1.9.3 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 10: *When player 1 has only 2 actions, $A_1 = \{a_1, a_1'\}$, and assumption 1 holds, an RTE exists.*

Proof of lemma 10. 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). □

1.9.4 Applications

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

Game 1.7.

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 1.10.

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 $\mathbf{PD}(\cdot \cdot p)$ 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.

1.9.5 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 9, 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.

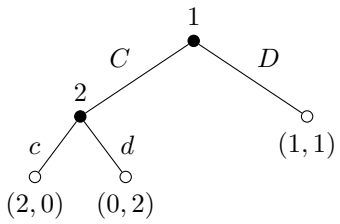
If instead, $\min_{a_1 \in CE_1(\beta_2^*)} \pi_2(a_1, \beta_2^*) < \min_{a_1 \in UE_1} \pi_2(a_1, \beta_2^*)$, yet player 2 prefers the same actions, then she must also prefer the same action given a reference point in between. If this isn't immediately obvious, simply take the utility difference of 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. □

1.10 Appendix B: Further Detail

1.10.1 RTE is an Equilibrium Concept

Game 1.11 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.

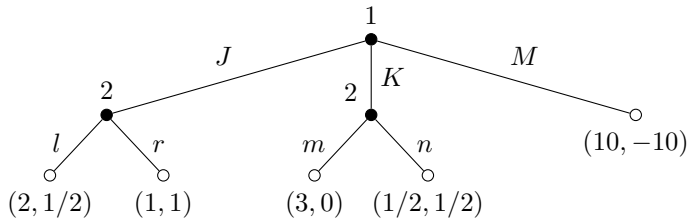
Clearly, this can never be an equilibrium belief and isn't very sensible. But without imposing a more restrictive utility function, such beliefs are indeed rationalizable. This suggests that imposing equilibrium is often needed - which I have done throughout the paper. It should be clear that when there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



Game 1.11

1.10.2 Relationship Between Efficient Sets

Game 1.12 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 1.12

1.10.3 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]

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 $M_2^{TE_1(\beta_2^*)}$ $\mathbf{PD}(\beta_2^*)$ $M_2^{UE_1}$. 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

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

$\pi_2(M_1^{TE_1(\beta_2^*)}, \sigma_2^{mBR}) > \pi_2(M_1^{UE_1}, \sigma_2^{mBR})$ - 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. □

Chapter 2

A Theory of Reciprocity with Trust, Incomplete Information

2.1 Introduction

In many social interactions, people do not have full information about other people's social motivations or their material payoffs. A person, who is motivated by reciprocity, may worry that his counterpart is not reciprocal or altruistic, but instead selfish or potentially spiteful. Upon observing company-wide wage cuts during a recession, workers don't necessarily know whether their company is taking advantage of its bargaining position, or whether such cuts are necessary for the company's survival. In case of the former, a reciprocal worker would feel obliged to punish the firm, whereas in the latter, he may keep on working normally. To incorporate such concerns, this chapter extends the theory of reciprocity with trust, which was developed in chapter 1, to incomplete information.

While intention-based reciprocity is frequently used to analyze games of complete information, it is rarely applied to incomplete information environments. Bierbrauer and Netzer (2016) introduce intention-based reciprocity to mechanism design by adapting Rabin (1993) to normal form games with incomplete information. Related work focuses on ex-post implementation (Netzer and Volk (2014)) and mechanisms that are robust to social-preferences (Bierbrauer et al. (2017), Bartling and Netzer (2016)). Von Siemens (2013) analyzes a simple two-stage game where a principal chooses how much to control a worker and introduces incomplete information about the worker's type. Le Quement and Patel (2017) examine how reciprocity can improve information

transmission in cheap talk.

In contrast to type-based models, intention-based reciprocity such as Rabin (1993) and Dufwenberg and Kirchsteiger (2004) requires more care when moving from complete to incomplete information. Type-based models assume heterogeneity in people's level of altruism and model reciprocity by assuming that people care more about those who are more altruistic (Levine (1998), Gul and Pesendorfer (2016)).[1] In this type of models, generous actions are rewarded because they signal a high level of altruism. Adding private information about material payoffs simply adds another dimension to such signalling games.[2] Intention-based models, instead, rely on an endogenous reference point to model kindness. The key challenge when moving from complete to incomplete information is defining an appropriate reference point for this new environment.

In this chapter, I adopt the perspective that players with different information are different individuals. Consequently, I define the reference point for each type of player separately, incorporating the idea of trust that was developed in Chapter 1. This paper contributes to the literature by providing the first general model of intention-based reciprocity for sequential games with incomplete information. After defining the model, I apply it to a pricing game with incomplete information about the buyer's valuation. When the buyer is sufficiently reciprocal, the seller, who makes a take-it-or-leave-it offer, suggests a price below the price p^s that would be profit maximizing if the buyer acted selfishly. High-valuation buyers reciprocate by voluntarily paying more than the suggested price, although less than p^s. By suggesting a low price, the seller can sell to more low-valuation customers, without suffering the full revenue loss from high-valuation customers. This application highlights how reciprocity can give rise to *pay-what-you-want* pricing schemes, where the seller allows the buyer to choose the price she wants to pay (Kim et al. (2009), Gneezy et al. (2012)).

For the next application, I revisit the classic bilateral trade problem. With discrete types, I show how simple sequential interactions can achieve full efficiency in cases where normal form mechanisms that satisfy incentive compatibility, individual rationality and budget balance, cannot. The sequential interaction takes the following simple form: the (selfish) seller informs the (reciprocal) buyer about his cost and the buyer responds by deciding whether to trade or not, and if she does, at what cost. In this setting, reciprocity creates incentives for information sharing. When the seller reveals that he has low costs, high-valuation buyers reward his kindness by

[1] Alternatively, people may simply care how they are perceived (as in self- or social image concerns, Bénabou and Tirole (2006)) or how they are perceived by people they care for (Ellingsen and Johannesson (2008)).

[2] See, for example, Bassi et al. (2014), who examine a work-place screening problem. Sally (2002) revisits Akerlof's (1970) lemon problem in his framework of sympathy.

sharing the surplus. The resulting free-flow of information improves efficiency by also enabling low-valuation buyers, who would not trade if everyone were to use their selfish strategies, to trade. More generally, this type of reciprocal information sharing can explain how cheap talk improves efficiency in bargaining (Valley et al. (2002)) without having to rely on honest types (Saran (2011), Saran (2012)) or lying aversion (Abeler et al. (2016)). Le Quement and Patel (2017) make a similar observation in the one-sided incomplete information environment of cheap talk (Crawford and Sobel (1982), Cai and Wang (2006)).

In the final section, I discuss how my model relates to alternative models of intention-based reciprocity in general and comment on different approaches to modelling kindness from an interim perspective.

The rest of this chapter is organized as follows. Section 2.2 begins by introducing two simple decision problems highlighting what additional aspects a theory of reciprocity needs to address in incomplete information environments. The model is formally developed in section 2.2.2. Section 2.3 features the applications on pricing with incomplete information about the buyer's valuation (2.3.1) and bilateral trade (2.3.2). Differences to the literature are discussed in section 2.4. Section 2.5 concludes.

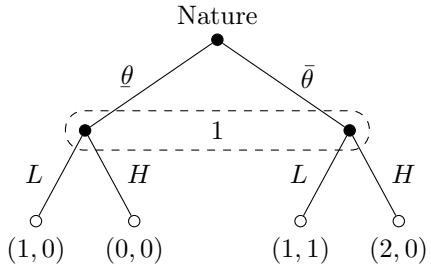
2.2 The Model

2.2.1 Kindness and Incomplete Information

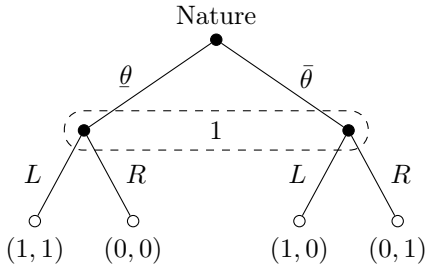
I begin this section by introducing two examples to outline the new features that incomplete information adds to kindness considerations. The examples come in the form of a decision problem, in which an uninformed player 1 chooses between two actions.[3] The payoffs for player 1 and for a passive player 2 depend on the state of nature. In contrast to player 1, player 2 is aware of nature's state. In view of reciprocity models for complete information, I will suggest how player 2 may perceive player 1 and will discuss several modelling approaches.

Game 2.1 can be interpreted as the reduced-form game, in which player 1 makes player 2 a take-it-or-leave-it offer. Player 2's valuation for the good is either high ($\bar{\theta}$) or low ($\underline{\theta}$) and she buys whenever her valuation is equal to or greater than the price.

[3] If player 1 also had private information, his kindness towards player 2 is conditional on his information, i.e. I adopt an interim perspective to kindness.



Game 2.1



Game 2.2

When players have private information, I adopt the perspective that players with different information are different individuals.[4] A high-valuation player 2 may thus perceive player 1 as kind (towards her type) when he selects the low price L. This is subject to one caveat, however: in simple choice problems with complete information, all intention-based reciprocity models posit that a person who chooses an allocation that maximizes the other person's payoffs isn't kind when it also maximizes his own payoffs, i.e. he chooses $(\pi_1, \pi_2) = (1, 1)$ over the only other alternative $(0, 0)$. Translating the same idea to incomplete information means that if action L maximizes player 1's expected payoff, it cannot be kind. That is, for $\Pr(\theta_2 = \bar{\theta}) \leq 1/2$, the high-valuation player 2 doesn't consider player 1 kind. A low-valuation player 2 considers player 1 neither kind nor unkind as her payoffs are independent of his actions.

While kindness towards each type is computed separately, any particular player 2 may appreciate the fact that player 1's behavior could also have benefited a different type of player 2 (either as a hypothetical case or due to player 1 repeatedly matching with short-lived player 2s). For example, suppose that player 1 had a third choice, vL, which lowered prices even further. In this case, not only would the high-valuation type $\bar{\theta}$ be even better off, she would also understand that player 1 is kind towards $\underline{\theta}$. Although I won't model this type of overall kindness in the main section, I will comment on it when discussing the applications (section 2.3). For now, the focus is on the kindness towards each type.

Game 2.1 has the feature that player 1's action affects player 2's payoff in similar ways: If an action a_1 improves θ_2's payoff relative to a_1', it also (weakly) improves a player 2's payoff whose

[4]While the literature on social preferences has investigated settings of more than 2 players (Güth and van Damme (1998), Zizzo and Oswald (2001), Charness and Rabin (2002), Engelmann and Strobel (2004), Bolton and Ockenfels (2006), among others) there is very general data with regards to what people perceive as kind or unkind in these settings, apart from a few specific settings. Charness and Rabin (2002) feature two 3-player games where a third player strongly punishes the greed of the first player. Fehr and Fischbacher (2004) show the predominance of third-party punishment both in the dictator game and the prisoner's dilemma. Engel and Zhurakhovska (2014) show that when cooperation in a prisoner's dilemma inflicts harm on a passive outsider, people cooperate less. Although experimental subjects cooperate more when they expect others to cooperate more, they find that the greater the cost to outsider, the weaker this effect becomes. In Malmendier and Schmidt (2017), gifts induce reciprocity at the detriment to a third party.

type is $\theta_2' \neq \theta_2$. Clearly, this does not need to be true in general. In game 2.2, for instance, action L improves $\underline{\theta}$'s payoff relatively to R, and vice versa for $\bar{\theta}$. Player 1's payoffs are independent of θ. Notice how this example nests two very different decision problems of complete information, $\Pr(\theta_2 = \bar{\theta}) \in \{0, 1\}$. When $\Pr(\theta_2 = \bar{\theta}) = 0$, R is unkind and L is neither kind nor unkind. In contrast, when $\Pr(\theta_2 = \bar{\theta}) = 1$, R is kind and L is unkind.[5]

There are three possibilities to define the 'efficient set' in this context: defining the efficient set (1) based on player 1's and player 2's expected material payoffs, (2) jointly for all types of player 2 that have a non-zero probability, and (3) separately for each type of player 2.

The first approach is employed in Netzer and Volk's (2014) (and the online appendix of Bierbrauer and Netzer (2016)) interim kindness definition. It can be useful when player 2 cares about player 1's kindness towards all types of player 2. As this paper aims to describe kindness towards each type first (and only then potentially averages across types), it appears natural to also build the reference point at the type level.[6] In section 2.4.2, I will revisit the differences between theirs and my approach.

Approach (2) closely resembles the idea of Pareto-efficiency. An action is part of E_1 as long as it is not dominated by another action, for at least one type. This would imply that for all $\Pr(\theta_2 = \bar{\theta}) \in (0, 1)$, a type $\underline{\theta}$ would perceive L as kind - even as the likelihood of $\bar{\theta}$ becomes vanishingly small. Not only is the discontinuity at 0 not a great technical feature, it also doesn't appear very plausible.[7]

The third approach adopts the viewpoint that when determining the immediate kindness of player 1 towards a θ_2-type player, only such type's payoffs are relevant, and not the payoffs of other types. This means that the presence of $\bar{\theta}$ does not affect the kindness perceptions of a $\underline{\theta}$-type. It also avoids extreme discontinuities around zero-probability beliefs.[8] A possible disadvantage of this approach is that R is judged more harshly by a $\underline{\theta}$ type as the reference point is only based on the efficient action L, even when action R appears to be a sensible strategy, i.e. especially when $\bar{\theta}$ is rather likely. As mentioned previously, it should be highlighted in this regard that player

[5] Specifically, the efficient set in the former is $E_1 = \{L\}$ since L Pareto dominates R, and $E_1 = \{L, R\}$ in the latter case since both actions are Pareto-efficient.

[6] Using both approaches at the same time appears less sensible. For instance, in game 2.2 only L is efficient in approach (1) when $\Pr(\theta_2 = \bar{\theta}) < 1/2$. In this case, the *inefficient* action R is perceived as unkind by type $\underline{\theta}$ ($0 < 1$) but kind by type $\bar{\theta}$ ($1 > 0$).

[7] This version is closely related to DK04's efficient set definition for more than 2 players. It would be interesting to empirically explore kindness in three player games further. For example, what happens to the classic 2 player example, where a action a_1, that induces payoffs of $(\pi_1, \pi_2) = (1, 1)$, isn't seen as kind if the only alternative action a_1' induces $(0, 0)$, when a third player is introduced. It would appear that if a_1 induces payoffs of $(1, 1, 0)$ vs. $(0, 0, 1/2)$ (for a_1'), player 2 would still not perceive player 1 as kind when he chooses a_1.

[8] Such discontinuities could also be used to blow up kindness perceptions, i.e. by introducing actions that are dominated for player 1 and almost all types of player 2, with arbitrarily low payoffs for most types. This is reminiscent of the original argument for restricting the set of payoffs in complete information settings, in order to avoid the influence of irrelevant, Pareto-dominated actions.

1's kindness towards other types can still be incorporated by averaging his kindness towards all player 2 types. In this paper, I will pursue this approach when modelling kindness.[9]

2.2.2 The Formal Model

Game. Let the game be a 2-player, finite, multi-stage game, with perfect but incomplete information and finite actions.

Players, information sets, and strategies. Let $N = \{1, 2\}$ be the set of players. Each player has a realized type $\theta_i \in \Theta_i$ which captures their *private* knowledge about payoff relevant aspects of the game. It may include information about player i's 'selfish' utility over an outcome or about her (own) concern for reciprocity.[10] Let Θ_i be a finite set and let $\theta = (\theta_1, \theta_2) \in \Theta = \Theta_1 \times \Theta_2$ be distributed according to $F(\cdot)$. For simplicity, I assume that players do not get further information about the state of nature θ.

Denote player i's information sets by H_i, with a typical information set being called $\mathbf{h} \in H_i$. The set of decision nodes are X and the set of terminal nodes are Z. $A_{i,\mathbf{h}}$ describes the (possibly empty) set of actions for player $i \in N$ at $\mathbf{h} \in H_i$. 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 $(1 \leq t \leq l)$. At times, it is useful to explicitly refer to the information set $\mathbf{h} \in H_i$ by using both history and type, $(\theta_i, h) = \mathbf{h} \in H_i$.

A system of beliefs μ specifies the beliefs (about decision nodes) at each information set $\mathbf{h}$.[11] Instead of writing beliefs in terms of abstract decision nodes, it is more useful to express them as conditional probabilities about the other person's type. That is, let $\mu_i(\theta_j|\mathbf{h})$ describe player i's belief that j has type θ_j at information set $\mathbf{h}$ instead of $\mu(x)$ when $x = ((\theta_i, \theta_j), h) \in \mathbf{h}$. Let the set of all possible belief systems be M.

Player i's behavior strategy is denoted by $\sigma_i \in \times_{\mathbf{h} \in H_i} \Delta(A_{i,\mathbf{h}}) =: \Delta_i^H$. It assigns each information set $\mathbf{h} \in H_i$ a probability distribution $\sigma_i(\cdot|\mathbf{h})$ over the set of pure actions. Define $\Delta^H := \prod_{i \in N} \Delta_i^H \ni \sigma$. Given σ let $P^\sigma(x)$, $P^\sigma(z)$, and $P^\sigma(\mathbf{h})$ denote the respective probabilities that node x, z, and information set $\mathbf{h}$ is reached.[12] Define the conditional probabilities

[9] Fehr and Schmidt (2006) aptly point out that for n-player games (n>2) it is often not immediately clear what the correct reference group is, highlighting a lack of theoretical and empirical work with regards to this issue. Recently, Mcdonald et al. (2012) suggest that the reference group may vary, resulting in non-monotonic behavior in ultimatum games when a third party with fixed endowment is present. Clearly, more empirical research is needed with regards to this fundamental question.

[10] It may also describe concerns for alternative other-regarding preferences such as altruism, spitefulness, etc.

[11] Formally, $\mu : X \to [0, 1]$ and satisfies $\sum_{x \in \mathbf{h}} \mu(x) = 1$ for each $\mathbf{h} \in H$.

[12] As usual, I omit the prior belief for these terms.

$P^\sigma(x|\mathbf{h}', \mu)$, $P^\sigma(z|\mathbf{h}', \mu)$, and $P^\sigma(\mathbf{h}|\mathbf{h}', \mu)$ in the same fashion.

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. Finally, define the expected utilities as usual: $\pi_i(\sigma|\mu_i, \mathbf{h})$ refers to player i's expected utility at information set $\mathbf{h}$ given belief μ_i, while $\pi_i(\sigma|\mu_i, \theta_i)$ refers to i's initial expected utility knowing θ_i.[13]

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. In addition to the player's beliefs about her opponent's type, she also forms a belief about her opponent's strategy (first order belief) and what she thinks her opponent thinks of her strategy (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 $\beta_i \in \Delta_i^H$. Player i also forms beliefs about her opponent's system of belief $\mu_j \in M_j$, which is called $\tilde{\mu}_{ij} \in M_j$.[14]

As in the complete information setting, the updating of beliefs about strategies is required. Beliefs are updated to match the observed actions for all types that a player considers likely. A history $h = (a^1, \ldots, a^l)$ is said to be an immediate predecessor to h', $h \succ_1 h'$, if and only if $h' = (a^1, \ldots, a^l, a^{l+1})$.[15]

Definition 1: *For any $\alpha_j \in \Delta_j^H$ and $\mathbf{h} = (\theta_i, h) \in H_i$, let $\alpha_j|\mathbf{h} \in \Delta_j^H$ be the updated first-order belief about strategies. It updates α_j as follows: First, take history h and its immediate predecessor h', $h' \succ_1 h = (a^1, \ldots, a^l)$, and set $\alpha_j(a_j^l \mid (\theta_j, h')) = 1$ (and respectively to 0 for all other actions) for any θ_j with $\mu_i(\theta_j|(\theta_i, h)) > 0$. Do the same for h' and its immediate predecessor h'', etc. β_i is updated in the same fashion, using the respective belief system $\tilde{\mu}_{ij}$.*

Just like in the complete information setting, a player adjusts her beliefs about the other person's strategy to match realized play. The only addition is that they only do so for types they consider likely.

To illustrate this idea, suppose that in a sequential prisoner's dilemma, player 1 is either altruistic ($\theta_1 = a$) or spiteful ($\theta_1 = s$) and that player 2 initially believes that both types cooperate, $\alpha_1(C|\theta_1) = 1$ for all $\theta_1 \in \{a, s\}$. Upon observing defection, if she believes that he is

[13] $\pi_i(\sigma|\mu_i, \mathbf{h}) = \sum_{z \in Z} P^\sigma(z|\mu_i, \mathbf{h})\pi_i(z)$.

[14] As in the complete information setting, I opt not to use belief hierarchies as seen in Battigalli and Dufwenberg (2009). One reason for this is to keep models consistent across chapters. A much more relevant reason is that both my theory of reciprocity with trust for complete information and this extension to incomplete information are equilibrium models. As a result, epistemological questions that require belief hierarchies are simply not a concern. Overall, the benefits of a more complicated, cumbersome model along the line of Battigalli and Dufwenberg (2009) appear to outweigh the cost.

[15] Alternatively, define it via the predecessor relationship of nodes. For $x = (\theta, h), x' = (\theta, h') \in X$, $h \succ_1 h'$ if $x \succ_1 x'$.

the spiteful type, $\mu_2(\theta_1 = s|D) = 1$, she updates her belief about his actions, leaving her beliefs about the altruist's action unchanged.

This type of updating, once again, implies that players must give up on probabilistic beliefs. If player 2, for instance, were to think that both types randomize and therefore considers each type possible after a_1, $\mu_2(\theta_1|a_1) > 0$ for all $\theta_1 \in \{a, s\}$, she will update her beliefs about strategies for each to match observed play.

Whenever a term features multiple updated beliefs, or also conditions on $\mathbf{h}$, I simply condition once at the *very* end i.e. $\pi_i(\beta_i, \alpha_j|\mu_i, \mathbf{h}) := \pi_i(\beta_i|\mathbf{h}, \alpha_j|\mathbf{h} \mid \mu_i, \mathbf{h})$.[16]

Perceived Kindness. As in the complete information setting, player i forms beliefs about j's kindness by comparing the payoff she thinks she will obtain against a reference point. The reference point is a convex combination of her minimal and maximal payoff in the trust-efficient set. Since this model computes kindness at the type level, that is between a type θ_j towards a type θ_i, each of the following terms will be defined for such a pair of types.

The trust-efficient set relies on the classification of actions into material best-responses, generous actions, and punishing actions. Define player i's material best-response as the behavior strategy that maximizes her payoff at every information set, that is

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

In case $\sigma_i^{mBR}(\alpha_j, \mu_i)$ is not unique, abusing notation, let it refer to a pure strategy that also maximizes j's payoff at every $\mathbf{h} \in H_i$ (among $\sigma_i^{mBR}(\alpha_j, \mu_i)$). To simplify notation, I will omit μ_i from this term whenever it is sensible. Denote the optimal choice at each $\mathbf{h}$ that makes up this pure strategy by $a_{i,\mathbf{h}}^{mBR}(\alpha_j, \mu_i)$, that is $\sigma_i^{mBR}(a_{i,\mathbf{h}}^{mBR}(\alpha_j, \mu_i)|\mathbf{h}) = 1$. Finally let $\sigma_i \backslash x_{\mathbf{h}}$ refer to the behavior strategy that replaces the local choice at $\mathbf{h}$ in σ_i by $x_{\mathbf{h}} \in \Delta(A_{i,\mathbf{h}})$.[17]

Deviations from the material best response are defined next. An action is called generous (punishing) towards j with type θ_j if it gives such type more (less) than what he would get as a result of the material best-response.

Definition 2: *Player i's action $a_i \in A_{i,\mathbf{h}}$ at $\mathbf{h} \in H_i$ is **generous** towards θ_j if*

$$\pi_j(\sigma_i^{mBR}(\alpha_j, \mu_i) \backslash a_i, \alpha_j|\theta_j, \mathbf{h}) > \pi_j(\sigma_i^{mBR}(\alpha_j, \mu_i), \alpha_j|\theta_j, \mathbf{h}).$$

*Action $a_i \in A_{i,\mathbf{h}}$ is **punishing** towards θ_j if $\pi_j(\sigma_i^{mBR}(\alpha_j, \mu_i) \backslash a_i, \alpha_j|\theta_j, \mathbf{h}) <$*

[16] Whenever $\mathbf{h}$ is not at the very end, i.e. if $h(\beta_i|\mathbf{h}, \sigma_j)$, it solely refers to the updating of the respective term.

[17] When $x_{\mathbf{h}} \in A_{i,\mathbf{h}}$ represents a pure action, it is implicitly understood that it refers to $\sigma_i(x_{\mathbf{h}}|\mathbf{h}) = 1$ and $\sigma_i(a_{\mathbf{h}}|\mathbf{h}) = 0$ for all other actions.

$\pi_j(\sigma_i^{mBR}(\alpha_j,\mu_i),\alpha_j|\theta_j,\mathbf{h})$. *Denote player i's set of generous actions towards θ_j at* $\mathbf{h}$ *by* $A_{i,\mathbf{h}}^{G,\theta_j}(\alpha_j,\mu_i)$ *and the respective set of punishing actions by* $A_{i,\mathbf{h}}^{P,\theta_j}(\alpha_j,\mu_i)$.

Consequently, define the trust-adjusted second order belief towards θ_j as follows:

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

for all $\mathbf{h} \in H_i$, $a_i \in A_{i,\mathbf{h}}$.

$\beta_i^{TE(\theta_j)}$ adjusts i's second order belief to the hypothetical belief that i takes her material best-response instead of a generous action.[18] While an action is usually (weakly) generous towards all types, this does not need to be true in general. Defining the idea of trust towards θ_j thus captures the most extreme case where i never takes the action that is generous towards θ_j.[19] Player j's trust-efficient strategies for the pair (θ_i,θ_j) are his Pareto-efficient strategies given $\beta_i^{TE(\theta_j)}$.

Definition 3 (Trust Efficiency)**:** *A behavior strategy $\sigma_j \in \Delta_j^H$ is trust-efficient towards θ_i for θ_j if there doesn't exist a $\sigma_j' \in \Delta_j^H$ that*

$$\pi_j(\sigma_j',\beta_i^{TE(\theta_j)}|\tilde{\mu}_{ij},\theta_j) \geq \pi_j(\sigma_j,\beta_i^{TE(\theta_j)}|\tilde{\mu}_{ij},\theta_j) \text{ and } \pi_i(\sigma_j',\beta_i^{TE(\theta_j)}|\theta_i,\theta_j) \geq \pi_i(\sigma_j,\beta_i^{TE(\theta_j)}|\theta_i,\theta_j)$$

with a strict inequality for at least one player. The set of trust-efficient strategies for the pair $\theta = (\theta_i,\theta_j)$ is denoted by $TE_j^{\theta}(\beta_i,\alpha_j,\mu_i,\tilde{\mu}_{ij})$.

Notice that in definition 3, player j's payoffs are expected payoffs, evaluated by i via $\tilde{\mu}_{ij}$, while i's payoffs are conditional on the pair of types. It implies that any action that is dominated in expectation for player j, which is also worse for θ_i, is not efficient. For example, in game 2.1, this means that L is the only efficient action if $\Pr(\theta_2 = \bar{\theta}) \leq 1/2$, in which case L cannot be kind. By conditioning i's payoff on (θ_i,θ_j), R is never efficient for $\theta_2 = \underline{\theta}$ in game 2.2, while for $\theta_2 = \bar{\theta}$, both actions are efficient.

[18] I opt for the simpler specification that relies on μ_i, interpreting β_i^{TE} as simply reflecting i's thought process of 'what would happen if she took her material best response'. One could also define it using the more involved second order belief about types, i.e. i's belief about $\tilde{\mu}_{ji}$, $\tilde{\mu}_{iji}$, instead of μ_i.

[19] In general, generosity depends on the other person's response. In the ultimatum game, for instance, an offer of 30% is generous to someone who accepts all offers but punishing to another person who is known to only accept offers of at least 50%.

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

Definition 4: *Let player i's reference point for a given pair $\theta = (\theta_i, \theta_j)$ at $\mathbf{h} = (\theta_i, h) \in H_i$ be*

$$\pi_i^r(\beta_i|\mathbf{h}, \alpha_j, \mu_i, \tilde{\mu}_{ij}, \theta) := \lambda \cdot \max_{\sigma_j \in TE_j^\theta(\beta_i|\mathbf{h}, \alpha_j, \mu_i, \tilde{\mu}_{ij})} \pi_i(\beta_i|\mathbf{h}, \sigma_j\,|\theta) + (1-\lambda) \cdot \min_{\sigma_j \in TE_j^\theta(\beta_i|\mathbf{h}, \alpha_j, \mu_i, \tilde{\mu}_{ij})} \pi_i(\beta_i|\mathbf{h}, \sigma_j\,|\theta)$$

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.[20] Combining all terms yields i's perception of j's kindness:

Definition 5: *Player i with θ_i perceives the kindness of a θ_j-player from strategy σ_j at history $\mathbf{h} = (\theta_i, h) \in H_i$, with $\theta = (\theta_i, \theta_j)$, according to*

$$\kappa_j(\alpha_j, \beta_i, \mu_i, \tilde{\mu}_{ij} \mid \theta, \mathbf{h}) := k\left(\pi_i(\alpha_j, \beta_i|\theta, \mathbf{h}),\ \pi_i^r(\beta_i|\mathbf{h}, \alpha_j, \mu_i, \tilde{\mu}_{ij}, \theta)\right)$$

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, j's kindness perceptions reduces to the usual $\kappa_j(\cdot|\,\theta, \mathbf{h}) = \pi_j(\cdot) - \pi_j^r(\cdot)$ for the pair θ. This function will be used in all examples.

Definition 6: *The utility of player i at $\mathbf{h} = (\theta_i, h)$ is*

$$U_i\left(\sigma_i, \alpha_j, \beta_i, \tilde{\mu}_{ij} \mid \mu_i, \mathbf{h}\right) = \pi_i(\sigma_i, \alpha_j|\mu_i, \mathbf{h}) + \gamma_i \sum_{\theta_j \in \Theta_j} \mu_i(\theta_j|\mathbf{h}) \cdot \kappa_j(\alpha_j, \beta_i, \mu_i, \tilde{\mu}_{ij}|(\theta_i, \theta_j), \mathbf{h}) \cdot \pi_j(\sigma_i, \alpha_j|\theta_j, \mathbf{h})$$

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

Player i's expected utility combines her usual expected material payoff with her expected utility from reciprocity. A type θ_j's payoff is weighted by how kind i perceives him to be. The total expected utility from reciprocity averages across all possible θ_j types. This averaging treats types

[20] 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 how to smooth out the reference point, see Appendix A in Rabin (1993).

as separate players and ensures that i wants to be kind to the 'right' person, namely the type that is kind towards her.[21]

Equilibrium. The equilibrium uses the familiar concept of sequential equilibrium and further requires the following to be correct: first and second order beliefs about strategies, beliefs about types, and beliefs about the other person's beliefs about types.

Definition 7: *An assessment* $(\sigma, \alpha, \beta, \tilde{\mu}, \mu,)$ *is sequentially rational if for all* $i \in N$*, for each* $\mathbf{h} \in H_i$*, and for any* $a_i \in A_{i,\mathbf{h}}$ *it holds that*

$$\text{if } \sigma_i(a_i|\mathbf{h}) > 0 \text{ then } a_i \in \arg\max_{a_i' \in A_{i,\mathbf{h}}} U_i\left(\sigma_i \backslash a_i', \alpha_j, \beta_i, \tilde{\mu}_{ij} \mid \mu_i, \mathbf{h}\right).$$

Let $\Delta^{H,0}$ denote the set of all completely mixed behavioral profiles. If $\sigma \in \Delta^{H,0}$, then any decision node $x \in X$ is reached with strictly positive probability, so that beliefs over $x \in \mathbf{h}$ can be defined in the usual way: $\mu_i(x) = \Pr^{\sigma}(x)/\Pr^{\sigma}(\mathbf{h})$.

Definition 8: *An assessment* $(\sigma, \alpha, \beta, \tilde{\mu}, \mu,)$ *is consistent if* $\mu = \lim_{n\to\infty} \mu^n = \Pr^{\sigma^n}(x)/\Pr^{\sigma^n}(\mathbf{h})$*,* $\sigma = \lim_{n\to\infty} \sigma^n$ *for* $\sigma^n \in \Delta^{H,0}$*, and* $\sigma_i = \alpha_i = \beta_i$*,* $\mu_i = \tilde{\mu}_{ji}$ *for all* $i \in N$*.*

Definition 9: *An assessment* $(\sigma, \alpha, \beta, \tilde{\mu}, \mu,)$ *is a sequential Reciprocity with Trust Equilibrium (sRTE) if it is sequentially rational and consistent.*

Sequential rationality requires that at each information set $\mathbf{h} \in H_i$ player i maximizes her utility, taking her future behavior as given. Consistency ensures that she holds correct beliefs.

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

The proof proceeds according to the usual logic: First it is shown that a perfect equilibrium with reciprocity preferences exists (Selten (1975)). This in turn implies the existence of a sequential Reciprocity with Trust Equilibrium.

[21] If player i, instead, forms expectation about j's kindness and expectations about j's payoffs, and then multiplicatively connects both, it can lead to questionable interactions across types of players. To see this, suppose that j is either selfish or reciprocal, and that player i moves first in a sequential prisoner's dilemma. Given that the selfish-type's payoff is large when i cooperates, even a relatively small average kindness due to the reciprocal-player may induce player i to cooperate. After all, the overall utility from reciprocity is relatively large. This goes against the idea that people want to be kind to those who are kind to them, and possibly punish those they consider unkind.

2.3 Applications

2.3.1 Pricing and Incomplete Information

For the first application, I analyze a simple pricing game with incomplete information about the buyer's valuation. In addition to accepting or rejecting the seller's offer, the buyer can reward the seller with a tip. We will see that if the buyer is sufficiently reciprocal, the seller effectively allows the buyer to set her own price.

Let there be a selfish seller and a reciprocal buyer, whose reference point is her minimum efficient payoff, i.e. $\lambda = 0$. The seller makes a take-it-or-leave-it offer for a single, indivisible good at a price $p \geq 0$. His cost of producing the unit is $c = 0$. The buyer has private information about how much she values the object. Her valuation is either high or low, $v \in \{v_l, v_h\}$, with $\Pr(v = v_h) = q$ and $v_h > v_l$. In addition to rejecting or accepting the offer ($a \in \{0, 1\}$), she can also tip the seller, $t \in [0, \bar{t}]$ with $\bar{t} > v_h$. Her material payoff is $\pi_B(a, t, p|v) = a \cdot (v - p) - t$.

If both buyer and seller were purely motivated by material payoffs, then the buyer will never tip and will accept the offer if and only if $p \leq v$. The seller sets the price at the high valuation, $p = v_h$, if $q \geq \frac{v_l}{v_h} =: \bar{q}$ and at the low valuation, $p = v_l$, otherwise.

The very same selfish behavior by the buyer is clearly also a trust-efficient response. It follows that when $q < \bar{q}$, a low price is never kind since it is in the seller's material interest. In this case, the seller sets $p = v_l$, which is accepted by any buyer, with tips being zero for any $\gamma_B > 0$.[22]

For the remainder of this section, assume that there are sufficiently many high types, that is $q \geq \bar{q}$. Clearly, the seller never sets the price above v_h and the high valuation buyer accepts if and only if $p \in [0, v_h]$. This is commonly understood by both players.[23] Denote the second order belief about tips by $t''(p) \in [0, \bar{t}]$. The utility of a high-valuation buyer from tipping t after observing a price $p \in [0, v_h]$ is

$$U_B(t, t''(p)|p, v_h) = v_h - p - t + \gamma_B \cdot \kappa_S(t''(p) \mid p, v_h) \cdot (p + t - c)$$

with $\kappa_S(t''(p) \mid p, v_h) = v_h - p - t''(p) - 0$.[24] Since $U_B(t, p|v_h)$ is linear in t (and tips being weakly positive), it must be that in equilibrium $\gamma_B \cdot (v_h - p - t''(p)) \leq 1$. This results in two cases: first,

[22] The example highlights the potentially strong informational requirement that the equilibrium model places on the buyer. It requires the buyer to fully understand the incomplete information environment the seller is facing in order to infer the seller's reasonable, efficient actions. When the seller takes an inefficient action, the buyer does not revise her beliefs about the seller's view on the environment, but treats the action as inefficient.

[23] I disregard the potential alternative equilibrium, where a high-price is inefficient due to being punished, i.e. rejected, which in turn can motivate the buyer to reject it.

[24] Notice that the reference point, which is the minimum efficient payoff, is 0 and arises from $p = v_h$. It takes as given that the buyer doesn't tip in response.

if the seller's kindness from p is too low even when $t''(p) = 0$, $v_h - p \leq 1/\gamma_B$, the buyer will never tip. If instead $v_h - p > 1/\gamma_B$, the optimal tip $t''(p)$ will make the buyer indifferent between any tip.[25] In equilibrium, the tip is

$$t(p) = \max\{0, v_h - p - 1/\gamma_B\}.$$

When the seller sets a sufficiently low price and the buyer's concern for reciprocity (γ_B) is large enough, she rewards him with a tip. The tip is decreasing in the price. The seller's total revenue from a high-valuation buyer is $\text{Rev}(p|v_h) = p + t(p) = \max\{p, v_h - 1/\gamma_B\}$ if $p \leq v_h$ and 0 otherwise. If the seller sets the price too high, the buyer will accept the offer but won't consider him kind. If the price is low enough, any decrease in price is matched one-for-one with an increase in tip. In this case, the seller gets a constant total payment which is below the price $p = v_h$ that is profit maximizing when everyone acts selfishly. The total payment from a high type must necessarily be below v_h for otherwise the seller isn't kind. However, he may be able to improve his expected revenue as we will see next.

Proposition 2: *The expected revenue is maximized at* $p = v_l$ *if* $v_h - 1/\gamma_B > v_l$ *and* $(1-q)\cdot v_l \geq q/\gamma_B$.

The first condition is a necessary condition. It requires that the high-valuation buyer is sufficiently reciprocal to tip the seller at a price of $p = v_l$. The second condition ensures that the expected gain from selling to the low-valuation buyer, $(1 - q)v_l$, exceeds the expected loss, q/γ_B, from selling the good to the high-valuation buyer at v_l instead of v_h.[26] Reciprocity concerns among high-valuation buyers allow the seller to profitably sell to more buyers. By lowering the price, the seller sells to more types of customers without having to bear the full loss of revenues from 'existing' high-value clients, who voluntarily pay more.[27]

This simple game captures three observations that are consistent with the empirical literature on *pay-what-you-want* pricing. First, most people pay more than the minimum required price, which is usually set at 0. Second, most people tend to pay less than the previously used fixed price. Third, companies tend to serve more customers (Kim et al. (2009), Regner and Barria (2009), Gneezy et al. (2010), Gneezy et al. (2012)). Consistent with proposition 2, a *pay-what-you-want* strategy also appears to be both profitable and sustainable over longer periods in some

[25] To see why, suppose $t''(p) = 0$. In this case, she wants to give the maximal tip $\bar{t}$. But then consistency requires $t''(p) = \bar{t} > v_h$, in which case she doesn't want to tip at all.

[26] Clearly the necessary condition is also sufficient at $q = \bar{q}$. At this prior, the seller is indifferent between selling at v_l or v_h even without any tips. With reciprocal types, he makes an additional tip on the high-value type.

[27] Notice that in the limit, $\gamma_B \to \infty$, the seller extracts all surplus.

environments.[28]

Before concluding this simple application, I would like to quickly explore an extension of the kindness function. The payment data in the empirical studies on *pay-what-you-want* pricing actually reveals that a small fraction of customers pay more than the original price when they get to choose their own price; for instance, see Kim et al. (2009). While this can be due to many reasons beyond reciprocity, it can be captured by allowing consumers to care not only about the seller's kindness towards them, but also about his kindness towards other types of customers who are possibly less well-off.[29]

This type of reciprocity preferences would incentivize the seller to lower the price even further, enabling poorer customers to participate at a per-unit price below his marginal cost.[30] To explore these ideas in more detail, a continuous type distribution would be useful. Clearly, the above analysis extends to this setting as well, yielding a very similar equilibrium condition for tips. Having buyers who also care about other buyers will push the seller's price further towards 'zero'. Similar to my simple, discrete setting, the continuous setup is unlikely to yield an actual price of zero. This, however, shouldn't be too surprising. The real reason behind pricing an item at $p = 0$ instead of a small but positive price is likely due to a discontinuity in kindness and trust perceptions at 0. Moreover, *pay-what-you-want* pricing schemes also tend to have some restrictions in practise: restaurants require drinks to be paid separately (Kim et al. (2009)) and music labels require consumers to pay for the cost of shipping the CD (Regner and Barria (2009)).

2.3.2 Bilateral Trade

After looking at a game with one-sided incomplete information, I now turn to two-sided incomplete information in the classic bilateral trade problem. I focus on a particular binary-type example that is taken from Bierbrauer and Netzer (2016) (henceforth BN16).

There is one indivisible object. The seller's cost is $\underline{c} = 0$ or $\bar{c} = 80$, the buyer's valuation is

[28] The longest evidence comes from Gneezy et al. (2012) in form of a two-year time series on payments, customers, and revenue of an Austrian restaurant following the introduction of a *pay-what-you-want* pricing scheme. While average payments decrease, customers and revenue increases over time. Hence, behavior doesn't seem to be driven by a novelty effect - which was likely an important factor in the famous example of the band Radiohead, who initially allowed fans to download their 2007 album *In Rainbows* and pay as much as they want. It is noteworthy that they did require fans to cover the credit-card fee of 45p, recovering their marginal download costs from selfish types.

The profitability of these schemes tends to vary across settings. In Kim et al. (2009), total revenues increased for a lunch-buffet at a restaurant, but fell for a movie theatre as well as for beverages in the cafe of a delicatessen shop. Gneezy et al. (2010) reports lower profits for souvenir pictures in an amusement park. When 50% of the *pay-what-you-want* price was given to charity, however, it dwarfed profits of the regular fixed price (with and without payments to charity). Lastly, also in Gneezy et al. (2012), profits from souvenir pictures on a tour boat are the same for posted prices and the *pay-what-you-want* scheme.

[29] See, for example, reporting for *The Guardian* by Carroll (2018).

[30] For instance, the high-value type could weight the seller's payoff with the new kindness function $\kappa_S(t''(p) \mid p, v_h)^{new} := \kappa_S(t''(p) \mid p, v_h) + \delta\kappa_S(a'', t''(p) \mid p, v_l)$ for some constant $\delta \in (0, 1)$.

$\underline{v} = 20$ or $\bar{v} = 100$. Types are independent and equally likely. If players are fully selfish, then it can be shown that Myerson and Satterthwaite's famous impossibility result holds - there is no social choice function f which is materially Pareto-efficient, satisfies incentive compatibility (IC) and the participation constraint (PC), and is budged balanced (BB).

BN16 show that when both player are strictly reciprocal, i.e. no one is selfish, there are simultaneous mechanisms that satisfy IC, PC, and BB, and trade occurs whenever it is materially efficient. In addition to sending a message about their type, each player is given a button that, if pressed, allows one to profit at the expense of the other. This button generates kindness whenever it is not used. For reciprocal players, not pressing the button becomes mutually reinforcing and furthermore can incentivise them to maximize joint payoffs, i.e. to honestly reveal their types.[31]

Clearly, a selfish player would always use such button in a simultaneous game. As a result, BN16's mechanism is limited to environments where the fraction of selfish types is small. When firms and consumers interact, this requirement is likely violated. In this section, I will explore whether sequential interactions may improve efficiency relatively to simultaneous moves, without relying on artificial buttons.

Suppose the buyer and the seller interact in the simplest sequential form. First, the seller makes a claim about his cost, $m_S \in \{\underline{c}, \bar{c}\}$. After observing his message, the buyer decides whether to trade or not and if so at what predetermined price. To do so, she announces her valuation, $m_B \in \{\underline{v}, \bar{v}\}$, and trade occurs if and only if $m_B \geq m_S$. Let the probability of trade and the resulting price *(prob of trade, price)* be as shown in Figure 2.1.

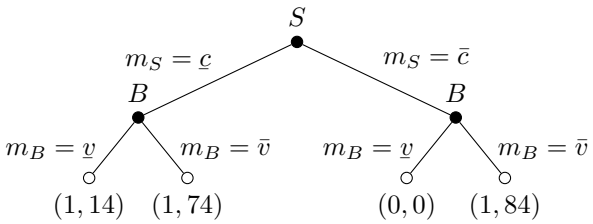


Figure 2.1: Probability of trade and prices

If players were selfish, both types of buyers respond to a low-cost message by trading at the lowest possible price, i.e. $m_B(v \mid m_S = \underline{c}) = \underline{v}$ for all v. After a high-cost message, only high-valuation buyers trade, $m_B(\bar{v} \mid m_S = \bar{c}) = \bar{v}$ and $m_B(\underline{v} \mid \bar{c}) = \underline{v}$. In equilibrium, all types of

[31] Observe that the revelation principle no longer holds in this setting.

sellers pool by claiming to have high-costs.[32] Due to the lack of trade among the pair $(\underline{c}, \underline{v})$, this particular sequential interaction is inefficient.

The buyer's material best-response can be used to determine the trust-efficient actions. In this setting all actions are efficient: given the buyer's material best-response, each seller prefers $m_S = \bar{c}$ while each buyer is better off when $m_S = \underline{c}$.[33]

For simplicity, assume that the seller is selfish and the buyer is reciprocal with obligation preferences, i.e. $\lambda = 0$.[34]

Proposition 3: *For $\gamma_B \in [1/10, 1/6]$, truth-telling is a sequential Reciprocity with Trust Equilibrium.*

For the high-cost seller, announcing his type is a dominant strategy, $m_S(\bar{c}) = \bar{c}$, since any price after the alternative $m_S = \underline{c}$ is less than his cost and trade always takes place. Given that a high cost message minimizes the buyer's payoffs, it is never kind. Consequently, any type of buyer will respond with her material best-response, i.e. she will announce her type: $m_B(\bar{v} \mid \bar{c}) = \bar{v}$ and $m_B(\underline{v} \mid \bar{c}) = \underline{v}$. In contrast, a low-cost message is indeed perceived as kind as it maximizes the buyer's payoff. A sufficiently reciprocal high-valuation buyer, $\gamma \geq 1/10$, is therefore willing to reveal her type, $m_B(\bar{v} \mid \underline{c}) = \bar{v}$. The low-value buyer sends an honest message, $m_B(\underline{v} \mid \underline{c}) = \underline{v}$, if her concern for reciprocity is not too large, ie. $\gamma_B < 1/6$. If γ_B is too large, a low-valuation buyer wants to exaggerate her type in order to reward the seller.[35] Whenever the buyer honestly reveals her type, it is also in the low-cost seller's best interest to be honest ($(14 + 74)/2 \geq 84/2$). It follows that truth-telling is a sequential Reciprocity with Trust Equilibrium for $\gamma_B \in [1/10, 1/6]$. The buyer's concern for reciprocity gives rise to full efficiency.[36]

While this is a game of incomplete information, the overall mechanism is, not surprisingly, identical to the one in games of complete information - see the sequential prisoner's dilemma in Chapter 1 for instance. By revealing his low costs, the low-cost seller improves the buyer's payoff irrespective of her type. Given the buyer's equilibrium response, it is also in his own material best interest to be honest, which makes his action mutually beneficial. The buyer perceives him to be kind since his action exposes vulnerability: if the high-value buyer were to take her material

[32] The expected profit of a low-cost seller is $(84 - 0)/2 > 14 - 0$.

[33] Notice that even for a high-cost seller, a low cost message is Pareto-efficient as it makes the buyer better off. One way to avoid such classifications would be to require that any player's payoff is at least equal to their outside option.

[34] As usual, this aims to capture a setting where many sellers (buyers) are selfish (reciprocal).

[35] Notice that such behavior would quickly lower the seller's kindness towards her.

[36] Indeed, this is true for any $\gamma_B \geq 1/10$. As in the extension for the previous pricing game, a high-value type may also appreciate the fact that the low-cost seller enables the low-valuation buyer to trade, increasing her kindness perception from $m_S = \underline{c}$ even further. Such preferences would make it even easier to motivate the buyer to reveal the truth.

best-response, the seller would be better off claiming to have a high cost instead. Consequently, she is happy to share some of the surplus with him. The sequential interaction exploits the asymmetry between the selfish seller and the reciprocal buyer by requiring the selfish seller to act first, which provides him with the opportunity to signal kindness and trust. Reciprocity once again ensures that the buyer doesn't exploit the seller's trust.

This example highlights how reciprocity creates incentives for information sharing. More generally, this type of reciprocal information sharing can explain how cheap talk improves efficiency in bargaining (Valley et al. (2002)) without having to rely on honest types (Saran (2011), Saran (2012)) or lying aversion (Abeler et al. (2016)). Le Quement and Patel (2017) make a similar observation in the one-sided incomplete information environment of cheap talk (Crawford and Sobel (1982), Cai and Wang (2006)).

There is one critical dimension in which reciprocity in settings of information sharing differs from reciprocity in games of complete information: in many games of complete information, negative reciprocity can be a tremendous force to ensure compliance by punishing selfish actions. Punishment becomes more difficult to justify if it is unknown whether someone's claim is truthful or not. After all, punishment may inflict harm not only on someone who lies for selfish gains but also on an honest person.[37] As people tend to give others the benefit of the doubt (Mitzkewitz and Nagel (1993), Rapoport et al. (1996), Güth and Huck (1997)) it is doubtful that punishment can be a driving force for creating efficiency - unless lies are likely to be discovered.

2.4 Discussion

In this section, I first discuss how my model relates to alternative models of intention-based reciprocity. Afterwards, I comment on some of the finer details of modelling kindness from the ex-ante and interim perspective.

2.4.1 Alternative Intention-Based Reciprocity Models

The first chapter featured an extensive comparison of the different intention-based reciprocity models. The essential difference to Rabin's (1993) conditional-efficiency model, extended to incomplete information, remains to be the fact that a person i does not consider another person j to be kind if j's action also benefits himself.[38]

[37] Indeed, to dissuade people from lying, one must be willing to punish honesty of others.

[38] For an example, see Le Quement and Patel (2017). One way to extend Rabin's model to sequential games of incomplete information is simply by using the actual belief β_i instead of β_i^{TE} in the definition of a trust-efficient

For the remainder of this subsection, I will explore why extending Dufwenberg and Kirchsteiger's (2004) model to incomplete information leads to arguably implausible behavior. This is ultimately due to the fact that a dominated action by player 2 may make an action by player 1 efficient. We will see that this can give rise to kindness perceptions that are independent of the prior belief over types.

Take a simple pricing game between a buyer and a seller. There are two types of buyers with valuation $v \in \{v_l, v_h\}$, with $\Pr(v = v_h) = q$. A selfish seller makes a take-it-or-leave-it offer for a single unit at $p \in \{v_l, v_h\}$ and has zero cost, $c = 0$. Upon observing the offer, the buyer decides whether to accept or not.

Suppose the buyer acts in her best material interest. A buyer with high valuation always buys; a buyer with low valuation buys if and only if $p = v_l$. As a result, the seller sets a price $p = v_h$ if $q > v_l/v_h$ and $p = v_l$ otherwise. Given that the buyer can only accept or reject the offer, it seems natural that a high-valuation buyer would only view the seller as kind when he sets a low price $p = v_l$ and acts against his own interest, i.e. when $q > v_l/v_h$. Indeed, this is exactly what my model predicts (for more details, see the pricing game section in 2.3.1).

When the seller's efficient actions are modelled via the unconditional efficient set, this is no longer true, however. To see this, recall that a price p is *unconditionally-efficient* if it is Pareto efficient for at least one strategy of player 2, $\sigma_2 \in \Delta_2^H$. According to this definition, both prices are efficient. $p = v_h$ is Pareto efficient due to the strategy where all types of buyers accept a price of $p = v_h$ (and all accept $p = v_l$). Notice that since a high-price is never kind, no low-valuation buyer would use such a strategy. Just like in game 1.6 (Chapter 1) this strategy is non-rationalizable, yet it affects the efficient set. Since $p = v_h$ is efficient for any $q \in [0, 1]$, a high-valuation buyer always views the seller as kind when he charges a low price: kindness no longer depends on the prior belief. This is particularly implausible when q is close to 0.

An efficiency notion based on material best-replies, actual choices (or second order beliefs thereof), or trust, avoids this problem and is thus more useful for incomplete information settings.

2.4.2 Modelling Choices for Kindness Perceptions

For their application of intentions to mechanism design, Bierbrauer and Netzer (2016) define kindness for normal-form games from an ex-ante perspective. A player's kindness is the difference between the ex-ante expected payoff and the ex-ante reference point. Efficiency is similarly defined from the ex-ante perspective, in the spirit of Rabin (1993). Netzer and Volk (2014) model kindness

behavior strategy.

(still for normal-form games) at the interim stage.[39] The kindness of player j with θ_j towards player i describes the difference between the expected payoff he gives to player i and her reference point. Both the reference point and the efficient set are defined in terms of i's expected payoff (conditional on θ_j). Player i perceives j's kindness towards her by taking expectation over j's types.

The key difference between their notion of interim-kindness and mine (apart from the obvious difference that their model is defined for normal form games and in terms of conditional efficiency) is how a particular type θ_i perceives j's kindness. In their model, kindness perceptions are independent of i's realized type, whereas in mine, j's action is only judged in terms of how it affects θ_i's (perceived) payoff. This difference leads to the testable implication of whether a type θ_i rewards or doesn't reward an action that is only kind to her hypothetical alternative selves.A more minor difference between the two models arises from the fact that their efficient set is defined in terms of interim expected payoffs, and not in terms of the pair (θ_i, θ_j). In many cases, there won't be any difference between both approaches. A notable exception is game 2.2, section 2.2.1, which describes a game where player 1's actions affect the payoffs for the two types of player 2 in opposite directions.[40]

For many applications, the private information encoded in θ_i will capture not only material payoffs but also social preferences. By treating different types as different players, a selfish player and a reciprocal player are treated as separate beings. As a result, player i's utility correctly weights her action's consequences on the selfish and the reciprocal player's payoffs by their respective kindness towards her and each type's likelihood. This captures the original idea of reciprocity that a person wants to help (hurt) the person she considers kind (unkind). This is not the case if i perceives j's kindness as the average of his behaviors across types and then uses this kindness term to weight j's expected payoff.[41]

[39] The interim-stage approach can also be found in the online appendix B.1 of Bierbrauer and Netzer (2016).

[40] Le Quement and Patel's (2017) model on cheap-talk only features one-sided information at the side of the sender. The sender's kindness is modelled conditional on his type. Their model is similar to mine in the sense that it is sequential and the receiver updates her beliefs about the state of the world based on the sender's message. She then computes the expected kindness given her updated belief, which they call ex-post perceived kindness. They also define an ex-ante kindness that is independent of the message by averaging across the states and the expected messages.

[41] To see this, suppose that player i moves first in a sequential prisoner's dilemma. Given that the selfish-type's payoff is large when i cooperates, even a relatively small average kindness due to the reciprocal player may induce player i to cooperate as her overall utility from reciprocity is relatively large.

2.5 Conclusion

This chapter extends the theory of reciprocity with trust, which was developed in chapter 1, to incomplete information. It captures the realistic feature that in many cases one does not have full knowledge about others' material payoffs or their social motivations. In two applications of this model, I showed how reciprocity can give rise to *pay-what-you-want* pricing schemes and how it can foster information sharing in bilateral trade.

Incomplete information about material payoffs gives rise to a particularly interesting question: If player j is genuinely trying to help player i, does i consider j kind even when j's help has no positive effect on her, or more extremely, his help actually hurts her? This paper adopted a 'selfish' approach to kindness perceptions: player i only considers player j's impact on her particular payoffs when she evaluates his kindness, neglecting any impact that he could have had on other hypothetical types of her. My approach is similar to the consequentialism of intention-based reciprocity models in complete information, in which one is judged by his actual actions. How a person actually reacts to such actions, and thus how reciprocity should be modelled in incomplete information, remains an empirical question.There has been very little research on what people consider kind in these games. I hope that my model inspires future empirical research on this fundamental topic.

2.6 Appendix: Proofs

Proof of proposition 1. The proof follows the usual logic of first showing that there exist an ϵ-constrained (reciprocity) equilibrium, that requires totally mixing strategies (with minimum probabilities of ϵ). Taking limits ($\epsilon \to 0$) results in the perfect equilibrium (with reciprocity preferences). From there, I use the usual argument to construct a sequential (Reciprocity with Trust Equilibrium) equilibrium.

Define the minimum trembles at $\mathbf{h} \in H_i$ for player i for the local strategy by $\epsilon_{i,\mathbf{h}}$. An *ϵ-constrained reciprocity equilibrium* is the totally mixed strategy profile σ^ϵ that is sequentially rational, with correct beliefs, but requires that at each $\mathbf{h}$, each action ($a_{i,\mathbf{h}} \in A_{i,\mathbf{h}}$) is played at least with probability $\epsilon_{i,\mathbf{h}}$, that is $\sigma_i\left(a_{i,\mathbf{h}}|\mathbf{h}\right) \geq \epsilon_{i,\mathbf{h}}$. Denote the restricted strategy space by $\Delta^{H,\epsilon}$.

A (trembling hand) *perfect reciprocity equilibrium* is any limit of an ϵ-constrained reciprocity equilibrium σ^ϵ as $\epsilon \to 0$. (Fudenberg and Tirole, 1991, ch. 8, def 8.5A)

1. I first argue that a ϵ-constrained reciprocity equilibrium σ^ϵ exists:
Define the local best response correspondence $r_{i,\mathbf{h}} : \Delta^{H,\epsilon} \to \Delta(A_{i,\mathbf{h}})^\epsilon$ by

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

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

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

Along the sequence of totally mixed strategies, μ is uniquely pinned down; it is without loss to drop it from the definition of $r(\cdot)$.

As $\prod_{(i,\mathbf{h}) \in N \times H}$ and $\Delta^{H,\epsilon}$ are topologically equivalent, I can simply define an equivalent function $\tilde{r} : \Delta^{H,\epsilon} \to \Delta^{H,\epsilon}$ and look for a fixed point. A fixed point under $\tilde{r}$ satisfy the conditions of the *ϵ-constrained reciprocity equilibrium* as it maximizes each player's expected utility of player at $\mathbf{h}$, and first and second order beliefs about strategies are correct (the same holds true for μ_i, μ_{ij}).

Kakutani's fixed point theorem applies in this setup. To see this, notice that at any information set $\mathbf{h}$ the local choice set $\Delta(A_{i,\mathbf{h}})^\epsilon$ is a subset of a simplex. As a simplex is compact,

convex and non-empty, so is the ϵ-restricted subset (for ϵ sufficiently small). $r_{i,\mathbf{h}}$ is non-empty as U_i is continuous in her own local choice, the set is compact and hence attains a maximum. $r_{i,h}$ is convex as U_i is indeed linear in i's own choice. Upper hemi-continuity of $r_{i,\mathbf{h}}$ follows from the fact that U_i is continuous. Since these properties extend from $r_{i,\mathbf{h}}$ to $\tilde{r}_{i,\mathbf{h}}$, and $\tilde{r}$, all conditions of Kakutani's fixed point theorem are satisfied, which completes the first step of the proof that $\tilde{r}(\sigma^\epsilon)$ has a fixed point.

2. A (trembling hand) *perfect reciprocity equilibrium* exists:
By compactness of the strategy space, it it follows by the Bolzano-Weierstrass theorem that the sequence $\{\sigma^\epsilon\}_\epsilon$ has a convergent subsequence as $\epsilon \to 0$, which proves the existence of a *perfect reciprocity equilibrium.*

3. A *perfect reciprocity equilibrium* implies a *sequential Reciprocity with Trust Equilibrium.*
This follows immediately from the usual argument. In a sRTE σ must be sequentially rational given the set of beliefs $(\alpha, \beta, \tilde{\mu}, \mu)$, and the assessment $(\sigma, \alpha, \beta, \tilde{\mu}, \mu)$ must be consistent. All that is left to do is to construct a sequence of beliefs $\mu^n \to \mu$. Along the convergent subsequence of totally mixed strategies σ^n, beliefs μ^n are uniquely defined by Bayes' rule. μ simply refers to the limit of the respective convergent subsequence from earlier. By construction, $(\sigma, \alpha, \beta, \tilde{\mu}, \mu)$ is consistent. As each player is taking her best response at each information set along the sequence, together with the fact that U_i is continuous, $(\sigma, \alpha, \beta, \tilde{\mu}, \mu)$ is sequentially rational. □

Proof of proposition 2. To improve upon the expected revenue of $p = v_h$, $\mathbb{E}\,\mathrm{Rev}(p = v_h) = q \cdot v_h$, it is necessary that both types buy since the total payment $p + t(p)$ from the high type is lower than v_h. As a result, any pice $p \in (v_l, v_h)$ cannot be optimal.

Since $q \cdot v_h > v_l$, it is necessary that the total payment from the high-type exceeds the low valuation, that is $v_h - 1/\gamma_B > v_l$. (This condition directly implies that for any $p \leq v_l$, the high-value buyer will tip the seller.)

Selling below $p < v_l$ cannot be optimal as the seller loses revenue from the low-valuation buyer (the transfer from the low valuation seller is $t(p) = \max\{0, v_l - p - 1/\gamma_B\}$) relatively to $p = v_l$ without affecting the revenue from the high-valuation buyer.

Lastly, expected revenue from v_l exceeds v_h if $\mathbb{E}\,\mathrm{Rev}(p = v_l) = q \cdot (v_h - 1/\gamma_B) + (1-q) \cdot v_l \geq q \cdot v_h$,

which can be rewritten as $(1-q)\cdot v_l \geq q/\gamma_B$. □

Proof of proposition 3. Kindness of a low-cost seller: Suppose the buyer holds the second order belief that she honestly reveals her type at each node. Denote this belief by β_B^{honest}. Moreover, she holds the correct belief that $\mu_B(\underline{c}|m_S=\underline{c})=1$. Since all messages are trust-efficient for the seller, and the buyer is assumed to have obligation preferences, her reference point is equal to her minimum payoff, which is induced by the high cost message. In particular $\pi_B^r(\beta_B^{\text{honest}}|(\underline{c},\underline{v}))=0$ and $\pi_B^r(\beta_B^{\text{honest}}|(\underline{c},\bar{v}))=100-84=16$. Hence each kindness term is

$$\kappa_S(\beta_B^{\text{honest}}|m_S=\underline{c},(\underline{c},\underline{v}))=20-14-0=6$$
$$\kappa_S(\beta_B^{\text{honest}}|m_S=\underline{c},(\underline{c},\bar{v}))=100-74-16=10.$$

The buyer's utility from each message is

$$U_B(m_B=\bar{v},\beta_B^{\text{honest}}|m_S=\underline{c},v)=v-74+\gamma_B\cdot\kappa_S(\beta_B^{\text{honest}}|m_S=\underline{c},(\underline{c},v))\cdot(74-0)$$
$$U_B(m_B=\underline{v},\beta_B^{\text{honest}}|m_S=\underline{c},v)=v-14+\gamma_B\cdot\kappa_S(\beta_B^{\text{honest}}|m_S=\underline{c},(\underline{c},v))\cdot(14-0).$$

She prefers $m_B=\bar{v}$ over $m_B=\underline{v}$ if $\gamma_B\cdot\kappa_S(\beta_B^{\text{honest}}|m_S=\underline{c},(\underline{c},v))\geq 1$. It follows that truth-telling is an equilibrium if $\gamma_B\in[1/10,1/6]$.

□

Chapter 3

The Benefits of Being Misinformed[1]

3.1 Introduction

We commonly observe situations in which people take actions in line with hypotheses that seem to be contradicted by empirical evidence. Furthermore, these actions often produce additional information that should highlight the initial mistake. For example, a significant number of people refuse even essential vaccinations despite the very strong evidence of their benefit and despite the measurable increase in outbreaks of the related disease as a consequence of this refusal.[2]

For this purpose, we revisit the comparison of experiments as in Blackwell (1951) under the assumption that information processing is not always flawless and might be impeded by systematic mistakes. We do this in a setup that captures the fundamentals of Bayesian updating and its consequences on utility: First an agent takes an action that directly affects his payoff but also provides information about the payoff relevant state of the world. The agent then takes another action before payoffs are realized. Motivated by the psychological and experimental literature on beliefs and perceptions, the information processing can be 'imperfect' in two ways: an agent might initially hold an incorrect prior affecting the type of signals he receives and the agent might misinterpret the signals.[3] We find that both types of biases by themselves are welfare decreasing. The potential welfare loss can be ordered according to the magnitude of the bias. Next, we provide necessary and sufficient conditions under which a given binary ranking of action profiles can be reversed by a perception bias. Building on these findings, we can show that it is not

[1] This chapter is joint work with Manuel Staab.
[2] See, for instance, Wallace (2009).
[3] For example: Bruner and Potter (1964), Darley and Gross (1983), Fischoff et al. (1977) and Lichtenstein et al. (1982).

always true that adding another type of bias to a pre-existing one makes an agent even worse-off. In particular, if the agent tends to misperceive signals, then pushing the agent's prior further away from the truth can beneficial if it causes the agent to prefer an action whose signal is less ambiguous and thereby less likely misinterpreted. Our setting thus provides a novel channel for the long known observation by Hirshleifer (1971) that information may not always be welfare improving.

On the other hand, when the agent's prior is incorrect and signals are useful in the sense that they inform decision-making, the agent will always be worse off from misperception. Only in the extreme case, where it's optimal for an agent to always take a fixed action irrespective of any signal, increasing the agent's degree of misperception can be beneficial. It follows that no straight-forward welfare ordering can be established if both types of misperception are present. We further explore the implications of an agents awareness over the bias. We provide conditions under which sophistication regarding the misperception of signals can help or harm the agent's welfare.

To provide an illustration, let's look at a student who decides whether to become an entrepreneur. The literature highlights that entrepreneurship is ex-ante not very profitable. Landier and Thesmar (2009) show that entrepreneurs tend to be overly optimistic with respect to their company's future growth prospects. Moreover, their optimism is positively correlated with higher-education, and more optimistic entrepreneurs tend to choose more short-term debt. Suppose the payoffs from being an entrepreneur depend on the student's ability. The student can either quit university immediately in order to start his business, or wait and finish his degree first. While graduating is inherently useful, it also provides him with a signal about his abilities. Finally, when he starts his business, the student also has to decide how much to borrow.

A student who is very overconfident in his abilities will find it optimal to start his business immediately, yet may not be overconfident enough to borrow much. In contrast, a slightly less overconfident agent will pursue the more 'sensible' path of completing his studies. When the student also suffers from biases in perception, he may (a) misinterpret the signal as confirming his superior ability or (b) over-interpret the signal strength. As a result, the student may come out of university feeling even more confident in his ability, which can persuade him to start a business as well as taking on larger risks, i.e. a bigger amount of debt. While both businesses fail with equal probability, the initially less over-confident agent may end up being worse off in expectation. The more over-confident agent managed to avoid the signal and thereby was not in a position to misinterpret it.

The chapter is organized as follows. In section 3.2, we summarize the relevant literature. This is followed by a basic description of our model in section 3.3 and the characterization of the unbiased choice problem in section 3.4. Section 3.5 introduces biased perception and biased priors into the model and relates our setting to Blackwell (1951). This is followed by our main results in section 3.6. We conclude with a short investment example in section 3.7 and a final discussion in part 3.8.

3.2 Literature

Blackwell (1951) formalizes when an information experiment is more informative than another. Marschak and Miyasawa (1968) transfer these statistical ideas to the realm of economics. The key conclusion is that no rational decision maker would choose to 'garble' his information, i.e. voluntarily introduce noise into his experiments.

Having more information, however, may not always be better in an economic settings. Hirshleifer (1971) highlights that public information may destroy mutually beneficial insurance possibilities. The recent behavioral literature takes this idea further: Papers on overconfidence have shed light on how holding incorrect beliefs can be useful. These benefits arise from strategic interaction between agents. Ludwig et al. (2011) show that overconfidence can improve the agents relative and absolute performance in contests by inducing higher efforts. De La Rosa (2011) studies the effect of overconfidence on incentive contract in a moral hazard problem. When agents overestimate the probability of success as well as their marginal contribution to success, it becomes easier for the principal to induce effort. It turns out that the efficiency gains from slight levels of overconfidence can improve the agent's welfare.

Carrillo and Mariotti (2000) show that Blackwell garbling of information may increase the current self's payoff when individuals are time-inconsistent. Benabou and Tirole (2002) develop this idea further. In their model, a time-inconsistent agent can take on a project with deferred and uncertain benefits but immediate losses. Due to time-inconsistency a time-0 self prefers that the project is undertaken in the next period, but anticipates a lack of motivation by her next period's incarnation. Having the opportunity to either perfectly learn the true success rate or to stay uninformed, she may prefer to stay uninformed if her prior motivates the time-1 agent to work. While these two papers are phrased as a decision problem, an agent with time-inconsistency plays a game among his different selves, which is the fundamental driver of these results.

Other papers analyze how various behavioral shortcomings can be improved upon by overcon-

fidence. These papers aim to provide a motivation why overconfidence exists in the first place. In Compte and Postlewaite (2004), failing to recall past failures can improve welfare as it counteracts the fear of failure. Brunnermeier and Parker (2005) highlight that agents prefer to hold incorrect, too optimistic beliefs about the future, when they derive immediate benefits from these expectations.

Our paper is also related to the enormous literature on mistakes in updating expectations. Given our focus on how basic biases in information processing interact, we shall keep references to specific biases short. In our model, agents have the tendency to misperceive information, which represents a generalisation of the confirmatory bias in Rabin and Schrag (1999). In their paper, they show how confirmatory bias, the tendency to misinterpret new information as supportive evidence for one's currently held hypothesis, can not only lead to overconfidence in the incorrect hypothesis, but even cause someone to become fully convinced of it.

There have been many studies that suggest people hold incorrect beliefs. On average, people tend to have unrealistically positive views of their traits and prospects. To mention a few, see Weinstein (1980) for health and salaries, Guthrie et al. (2001) for rates of overturned decisions on appeal by judges, and Fischoff et al. (1977) as well as Lichtenstein et al. (1982) for estimates of ones' own likelihood to answer correctly. Recent papers include Landier and Thesmar (2009), for entrepreneurs, as well as Malmendier and Tate (2005) linking CEO overconfidence to a higher likelihood of pursuing risky actions, for instance acquisitions. Benoit and Dubra (2011) argue that a lot of the empirical evidence is also consistent with Bayesian updating under correct priors and thus may not demonstrate overconfidence. In response, laboratory experiments robust to this criticism were carried out by Burks et al. (2013), Charness et al. (2014), and Benoit and Moore (2015), again documenting overconfidence.

3.3 The Setting

We consider a two period model with two states of the world $\Omega = \{A, B\}$. The agent has a finite set of actions $X = X_1 \times X_2$. In the first period, he chooses an action from X_1 and subsequently receives a signal about the state. The quality of the signal depends on the action he chooses. He then decides on a second action from X_2 and receive a payoff determined by the action profile as well as the state of the world. The probability that a particular state materializes is $\Pr(\omega = A) = p \in (0, 1)$ and $\Pr(\omega = B) = 1 - p$ respectively. Denote the agent's prior belief that the state is A by μ. This belief may or may not coincide with the correct p. Let $u(x_1, x_2|\omega)$

denote the payoff if x_1 is taken at $t = 1$, x_2 at $t = 2$ and the state is ω, with $u(x_1, x_2|\omega) \in \mathbb{R}$.

We make the following assumptions on payoffs which guarantee that in each state there is a unique best action profile:

Assumption 1: *There exist action profiles:*

1. $(x_1^A, x_2^A) \in X$ *such that* $u(x_1^A, x_2^A|A) \geq u(x_1, x_2|A)$ *for any* $(x_1, x_2) \in X$.
2. $(x_1^B, x_2^B) \in X$ *such that* $u(x_1^B, x_2^B|B) \geq u(x_1, x_2|B)$ *for any* $(x_1, x_2) \in X$.

To avoid trivial scenarios, we exclude cases in which those two profiles have any common component. It will also be implicitly assumed that X only contains actions that are not completely payoff equivalent and ties are broken deterministically. Additionally, payoffs are assumed to be bounded:

Assumption 2: *There exists a* $K \in \mathbb{R}$ *such that* $K > \max\{u(x_1^A, x_2^A|A), u(x_1^B, x_2^B|B)\}$.

After period one, an agent receives a private signal $s \in S$. For each action $x_1 \in X_1$, there is a binary set of potential signals $S(x_1) = \{a(x_1), b(x_1)\} \subset S$. The probability distribution over those signals is conditional on the action as well as the state of the world meaning that x_1 defines a probability measure on $S \times \Omega$. Let $\pi(x_1, \omega) \equiv Pr(s = a(x_1)|\omega, x_1)$ and consequently $1 - \pi(x_1, \omega) \equiv Pr(s = b(x_1)|\omega, x_1)$. Let the signal structure be symmetric between states, i.e. $\pi(x_1, A) = 1 - \pi(x_1, B)$, and let signals be weakly informative in the sense that a-signals are more likely than b-signals in state A. We call $(x_1, \{x_a, x_b\})$ an *action profile* where $x_a, x_b \in X_2$ represent the actions taken after an $a(x_1)$ or $b(x_1)$ signal. Such an action profile is said to be *signal-sensitive* if $x_a \neq x_b$. A *simple action profile* is such that $x_a = x_b$ meaning the second-period choice is not conditional on the signal. It is for brevity denoted by (x_1, x_2).

We can think of the first action as an experiment that delivers information about the realized state. The agent can react accordingly and adjust his action in the second period. Notice, however, that an experiment is only useful if it is not too costly. For instance, it could perfectly reveal the state but reduce the attainable utility to an extent that it would be better to choose a noisier experiment.

3.4 Unbiased Choice Problem

We reverse-engineer the agent's decision problem step by step which serves to clarify the later results. First we look at optimal actions in the second period and then focus on the optimal

choice of experiments in the first period. The discussion is rather technical and the descriptions necessarily somewhat dry. But it allows for more constructive proofs later. Readers impatient for key results may skip ahead and return to this section as needed for later understanding.

3.4.1 Period 2 Cutoff-Strategy

The expected utility of an action $x_2 \in X_2$ from the second period's perspective depends on the posterior after receiving the signal. Furthermore, it can also depend directly on the action in the previous period. Let $\mu(s)$ be the posterior after receiving a type-s signal. An agent chooses action x_2 at $t = 2$ given a posterior $\mu(s)$ and an action x_1 at $t = 1$ if

$$E\left[u(x_1, x_2)|\mu(s)\right] \geq E\left[u(x_1, x)|\mu(s)\right] \qquad \forall x \in X_2$$

The expected payoff from any given action in period 2 is monotonic in beliefs. Starting from some $0 < \mu < 1$, an increase in μ strictly increases the expected payoff from (x_1^A, x_2^A) and decreases the one from (x_1^B, x_2^B). In fact, for any $(x_1, x_2) \in X$ with $u(x_1, x_2|A) \neq u(x_j, x_k|B)$, the expected payoff is either strictly in- or decreasing in μ. We can order actions at $t = 2$ according to expected payoffs based on $\mu(s)$. This gives rise to a cutoff-type decision rule.

To illustrate this, consider the period 2 expected payoff of x_2^A given the action x_1^B in period 1 and some posterior $\mu(s)$:

$$\mu(s)u(x_1^B, x_2^A|A) + (1 - \mu(s))u(x_1^B, x_2^A|B).$$

Similarly, the expected payoff from x_2^B is

$$\mu(s)u(x_1^B, x_2^B|A) + (1 - \mu(s))u(x_1^B, x_2^B|B).$$

If $u(x_1^B, x_2^A|A) > u(x_1^B, x_2^B|A)$, there exist some $\mu^*(s) \in (0, 1)$ such that for all $\mu(s) > \mu^*(s)$, the expected payoff from choosing x_2^A is larger than from x_2^B. Equally, for all $\mu(s) < \mu^*(s)$, the expected payoff from x_2^B exceeds the expected payoff from x_2^A as $u(x_1^B, x_2^A|B) < u(x_1^B, x_2^B|B)$. An equivalent argument shows that the same is true for any other $x_2 \in X_2$ such that $u(x_1^B, x_2|A) > u(x_1^B, x_2^B|A)$. For μ large enough, the expected payoff from x_2 exceeds the one from x_2^B and vice versa for μ small enough. Iterating this argument over all available actions at $t = 2$ and each $t = 1$ action allows us to conclude that the region in which a given x_2 is chosen must be connected.

Result 12 in Appendix C shows this formally.

We can now derive the conditions under which a given action is chosen in period 2 for at least some beliefs. For any $x_1 \in X_1$ there is a $x_a \in X_2$ such that $u(x_1, x_a|A) \geq u(x_1, x_k|A)$ for all $x_k \in X_2$. x_a is optimal for $\mu(s) = 1$ and will be chosen if $\mu(s)$ is high enough. The equivalent is true for some action in state B. If the actions are identical across states then this is the best possible action for all $\mu(s)$. Otherwise, there might be different optimal actions for intermediate beliefs.

Consider the threshold belief $\mu_{a,j}$ for which the expected payoff from (x_1, x_a) is exactly equal the one from (x_1, x_j):

$$\mu_{a,j}u(x_1, x_a|A) + (1 - \mu_{a,j})u(x_1, x_a|B) = \mu_{a,j}u(x_1, x_j|A) + (1 - \mu_{a,j})u(x_1, x_j|B)$$

We can rearrange the equation to

$$\frac{\mu_{a,j}}{(1 - \mu_{a,j})} = \frac{u(x_1, x_j|B) - u(x_1, x_a|B)}{u(x_1, x_a|A) - u(x_1, x_j|A)}.$$

Since by definition $u(x_1, x_a|A)$ is the highest utility in state A for x_1, the equation highlights that $u(x_1, x_j|B) > u(x_1, x_a|B)$ is a necessary and sufficient condition for this threshold to exist. We can then simply order actions according to payoffs in both states and ignore actions that are dominated in both states. Starting from $\mu_a = 1$, for which x_a is the best action, we can compare the potential cutoffs for all actions $x_i \in X_2$ that are not strictly dominated. The action with the highest $\mu_{a,i}$ will be the one chosen for some range of beliefs. We then continue to iterate this process from this action until there is no more action that has a higher payoff in state B.

Result 1: *For every $x_1 \in X_1$, there exists a partition $\mathcal{P}_{x_1}$ of $[0, 1]$ such that for every two consecutive elements p_i and p_{i+1} of the partition, there is an action $x_2 \in X_2$ such that $E[u(x_1, x_2)|\mu] \geq E[u(x_1, x)|\mu]$ for all $x \in X_2$ and $p_i < \mu < p_{i+1}$.*

Result 1 follows directly from the previous argument and Result 12. It highlights how the choice in period 2 depends on the posterior, which in turn is determined by the signal. Differences in the posterior are only welfare relevant if they fall in different elements of the partition. For binary signals, there are at most 2 different choices in period 2 and thus potentially 4 different utility outcomes. Since those are pinned down for every $x \in X_1$ by $\mathcal{P}_x$, we can collapse the problem to a comparison of experiments in period 1, fixing the corresponding optimal period 2 choices.

3.4.2 Choice in Period 1

In the first period, the decision maker simply maximizes his expected utility:

$$\max_{x_1 \in X_1} E\left[u\big(x_1, \{x^*_{a(x_1)}, x^*_{b(x_1)}\}\big)|\mu\right]$$

where $\{x^*_{a(x_1)}, x^*_{b(x_1)}\}$ are the optimal period 2 actions following $x_1 \in X_1$. The period 1 choice balances the information value of an experiment as well as the immediate utility derived from it. A very informative action leads to very different posteriors and, keeping in mind the partition, to different actions in period 2. When the experiment is uninformative, the posterior $\mu(s)$ equals the prior μ and falls into the same bracket of the partition.

Taking x_1 as given, let $\mu(s)$ be the posterior after receiving a type-s signal and x^*_s be the corresponding optimal action. We can write the expected utility as

$$\begin{aligned}&[\mu(a)u(x_1, x^*_a|A) + (1-\mu(a))u(x_1, x^*_a|B)]Pr(s=a|x_1,\mu)\\ +\ &[\mu(b)u(x_1, x^*_b|A) + (1-\mu(b))u(x_1, x^*_b|B)]Pr(s=b|x_1,\mu).\end{aligned}$$

Recall that the probability of receiving an $a(x_1)$ signal in state A is $\pi(x_1, A) \geq \frac{1}{2}$ which is also equal the probability of receiving the signal $b(x_1)$ in state B. Using Bayes' rule, we can rewrite the expected utility expression as:

$$\begin{aligned}\mu \quad &[\pi(x_1,A)u(x_1,x^*_a|A) + (1-\pi(x_1,A))u(x_1,x^*_b|A)]\\ +(1-\mu) \quad &[\pi(x_1,B)u(x_1,x^*_a|B) + (1-\pi(x_1,B))u(x_1,x^*_b|B)]\end{aligned} \tag{3.1}$$

This is the objective function for the utility maximization problem at μ. It is a weighted average of receiving the "correct" signal and thus choosing the correct action and the probability of receiving the "incorrect" signal and thus choosing the action that yields the lower utility in the realized state. A higher informativeness in the sense of $\frac{\pi(x_1,A)}{1-\pi(x_1,A)}$ reduces the likelihood of such a mistake. An agent might then be confident enough in the signals that he chooses actions that have a higher variation between states.

We finish this section with a key property of the unbiased agent's problem. Some of the later results arise from a violation of this.

Result 2: *The maximum expected utility at $t = 1$ as a function of μ is convex in μ.*

The previous discussion is best illustrated with a simple example:

Example 1: In each period, there are three actions $\{x_A, x_I, x_B\} = X_1 = X_2$. The actions x_A, x_B yield uninformative signals with $\pi(x_A, A) = \pi(x_B, A) = 0.5$. Action x_I provides an informative signal with $\pi(x_I, A) = 0.75$. Utilities are symmetric in states such that $u(x_A, x_A|A) = 5 = u(x_B, x_B|B)$, $u(x_I, x_A|A) = u(x_I, x_B|B) = 4$ while all other combinations yield 0 utility. x_I represents a pure information experiment that is only useful because it indicates the true state and thus the appropriate action at $t = 2$.

In this setting, the agent never chooses a combination of x_A and x_B. If the agent takes the information experiment in period 1 then he will choose between x_A and x_B in period 2 depending on whether his posterior is greater or smaller than $\frac{1}{2}$.

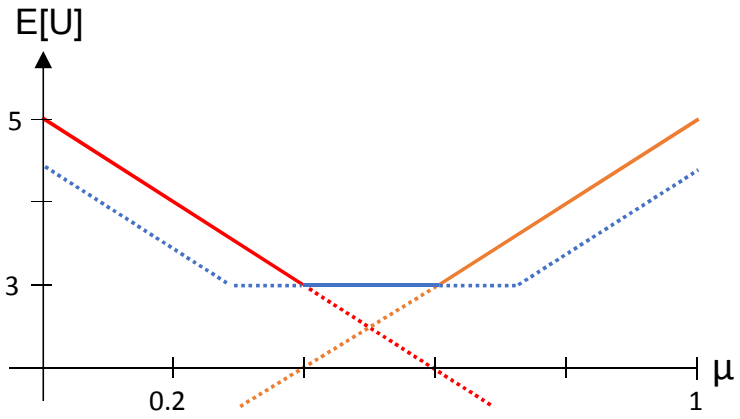


Figure 3.1: Utility frontier

Figure 3.1 illustrates the expected utility outcomes. Actions x_A (orange) and x_B (red) are optimal for extreme enough beliefs. For intermediate beliefs, the informativeness of x_I (blue) is more valuable. The posterior can fall into both partition elements such that x_A or x_B is chosen at $t = 2$ conditional on the signal. His maximum expected utility, indicated by the bold line, is clearly convex in μ.

3.5 Biased Perception

We now turn to our main area of investigation: scenarios in which the agent might misinterpret parts of his environment. In particular, he might not always perceive information accurately or might commit systematic mistakes when observing signals. We call this a *perception bias*. Furthermore, the agent could have a prior that is deviating from the true probability distribution

over states. We call this somewhat loosely a *bias in prior*.

The perception bias adds noise to the signal; a form of Blackwell garbling. It weakens the correlation between the signal and the state and thus reduces the information value of experiments in period 1. The agent arrives at incorrect posteriors following the initial signal and may choose suboptimal actions in the second period. The bias in prior might lead to a suboptimal choice of experiment in period 1 with potential consequences for the subsequent choices.

The perception bias could be due to an agent's unwillingness to change their preferred hypothesis or a systematic mistake in interpreting information. Initially, we take no stand regarding the source and exact form of the bias. Unless stated otherwise, we make the crucial assumption that the agent is naive in the sense that he is unaware of this perception problem. He computes his posterior and chooses his actions as if he was receiving signals according to the true underlying structure. In a later section, we discuss how this is related to a setting where agents have no perception bias but simply a distorted view about the accuracy of signals. We argue that they overlap but are not identical. We also explore the implication of a sophisticated agent, who is aware of his perception problem.[4]

To illustrate the setting, imagine the following thought experiment. A doctor orders a medical test. The outcome can be either negative or positive. On top of any inaccuracies of the test itself, suppose there is a certain chance the lab technician enters the incorrect result in the patient's file. The attending doctor never misperceives information if the technician does his job perfectly. However, there will be an information loss if, for instance, the technician mistakenly enters a positive result if the test actually came back negative. It is as if the doctor misperceives the signal. The distortion does not have to be balanced but is independent of the state of the world: the lab technician does not have any knowledge about the truth other than through the test result he observes. If the doctor is unaware of his technician's potential mistake, we refer to him as naive; if he takes his mistakes into account, we call him sophisticated. As mentioned before, we mostly maintain the assumption of a naive doctor. An incorrect prior regarding the condition of the patient might simply lead to an unnecessary test (or failure to conduct a necessary one).

Let s^t be the signal observed by the technician conducting the experiment x_1, and s the signal received by the doctor. Then

Assumption 3: $\Pr(s, s^t|\omega, x_1) = \Pr(s^t|\omega, x_1) \cdot \Pr(s|s^t, x_1) \qquad \forall \omega \in \{A, B\}$.

Define the signal probability distortion function d as a function that converts the true signal

[4]Note that a distinction between a sophisticated and naive agent would be meaningless for a bias in prior.

probabilities into (potentially different) probabilities that capture the likelihood with which the agent actually perceives the signals. In particular, $d((\pi, 1-\pi), S(x_1), \mu)$ gives the vector of probabilities for a and b signals for a belief μ, signal structure $S(x_1)$, and the undistorted probability structure $(\pi, 1-\pi)$. We can break this down into two functions, d_a and d_b, one for each type of signal. Let the probability of correctly transmitting an a signal when performing the experiment x_1 be $k_a(x_1, \mu)$, and let $k_b(x_1, \mu)$ be the corresponding probability for a b signal. Notice that the likelihood of mistakes may depend on the prior of the agent as well as the type of signal. We can then write the signal probability distortion function as:

$$d((\pi, 1-\pi), S(x_1), \mu) = \begin{pmatrix} d_a(\pi, x_1, \mu) \\ d_b(1-\pi, x_1, \mu) \end{pmatrix} = \begin{pmatrix} k_a(x_1, \mu)\pi + (1 - k_b(x_1, \mu))(1-\pi) \\ k_b(x_1, \mu)(1-\pi) + (1 - k_a(x_1, \mu))\pi \end{pmatrix} \tag{3.2}$$

for any $\pi \in [0, 1]$ with the obvious restriction that $d_a(\pi, x_1, \mu) + d_b(1-\pi, x_1, \mu) = 1$. An increase in the *magnitude of the bias* means a reduction of $k_a(x_1, \mu)$, or $k_b(x_1, \mu)$, or both.

Result 3: *The signal probability distortion function d takes the form of equation* (3.2)*. It is state independent in the sense that $k_a(x_1, \mu)$ and $k_b(x_1, \mu)$ are state independent.*[5]

For convenience, we can represent the actual experiment as well the distorted one by a Markov matrix. In this notation, an experiment is characterized by a 2×2 Markov matrix K, where element K_{ij} refers to the probability of receiving a signal $s = j$ in state ω_i. Denote the unbiased, information experiment arising from x_1 by P_{x_1} with the structure:

$$P_{x_1} = \begin{bmatrix} \pi & 1-\pi \\ 1-\pi & \pi \end{bmatrix}$$

The agent's perception bias is represented by a Markov matrix $M_{P_{x_1}}$:

$$M_{P_{x_1}} = \begin{bmatrix} k_a(x_1, \mu) & 1 - k_a(x_1, \mu) \\ 1 - k_b(x_1, \mu) & k_b(x_1, \mu) \end{bmatrix}$$

The resulting probability structure of the experiment $P_{x_1} M_{P_{x_1}}$ captures exactly the idea behind

[5] Notice that the true probability of receiving a certain signal is obviously still conditional on the state. If we want to think of this distortion as confirmation bias along the lines of Rabin and Schrag (1999), with an agent being biased towards hypothesis B and thus misinterpreting a signals as b signals with probability δ, but never the other way around, we set: $k_a(x_1, \mu) = 1 - \delta$, $k_b(x_1, \mu) = 1$.

the signal probability distortion function:

$$P_{x_1} M_{P_{x_1}} = \begin{bmatrix} d_a(\pi, x_1, \mu) & d_b(1-\pi, x_1, \mu) \\ d_a(1-\pi, x_1, \mu) & d_b(\pi, x_1, \mu) \end{bmatrix}$$

3.5.1 Blackwell '51

We now take a small detour to Blackwell's seminal research on the comparisons of experiments (Blackwell (1951), Blackwell and Girshick (1954)).[6] This digression is useful as it highlights the general principles behind information experiments. We will also see how Blackwell's general results translate to settings where agents are biased - both naive, or sophisticated with respect to their perception mistakes.[7]

In Blackwell's comparison of experiments, an agent compares statistical information experiments. Our setting is slightly more general in the sense is that experiments not only yield information but may also have direct payoff consequences. Upon receiving a signal from an information experiment, the agent maximizes his utility. Denote the value function of experiment P by $V(P)$.

Definition 1: *Let P and Q be two experiments. We say that experiment P is more informative than Q, denoted by $P \supset Q$, if there is a 2×2 Markov matrix M with $PM = Q$.*

This definition highlights how the information experiment P is garbled into Q by the matrix M; it is as if noise was added to P to create a less informative experiment Q. From this statistical notion, Blackwell derives his famous result:

Result B1: *Let P and Q be two experiments. $V(P) \geq V(Q)$ if and only if there is a 2×2 Markov matrix M such that $PM = Q$.*

Simply put, a more informative experiment implies a higher utility. Surprisingly, the converse is also true: If one experiment yields a higher expected utility than another, it must be more informative.

In Blackwell's setup the agent is rational. He fully understands the signal generating structure of the two experiments and correctly perceives each signal. The garbling matrix doesn't represent the agent's misperception but is a tool to order the experiments by informativeness.

[6] Both are phrased very much in the language of a statistician. For an economist's take on this see Marschak and Miyasawa (1968) who translated the setup into a utility-framework.

[7] For further details about the general setting of Blackwell, please see the Appendix B. All statements generalise to n-signals and n-states. These can also be found in the Appendix.

In order to talk about a Blackwell setting when agents misperceive information, we need a new function representing the agents' expected utility from the experiments. Suppose the unbiased, true information experiment is P but the agent misperceives some signals according to matrix M. As a result, the agent faces an information experiment PM. We denote the expected utility of a naive agent, who is unaware of his perception bias, by $V_n(P, M)$. The first argument always indicates the original signal generating process and the second argument the agent's misperception. This agent believes the information experiment is P and thus chooses his action as if he was rational facing P. If the agent is sophisticated, that is he is cognisant of his misperception M, we write his utility as: $V_s(P, M)$.

At this point, all tools and notation are developed, the stage is set for agents who are unbiased, sophisticated or naive with respect to their signal processing.

3.6 Biases and Their Implications

We start with the statement that the agent is always worse off with a perception bias than without. This observation follows from the original Blackwell result on the ordering of information experiments.

Result B2: *Take two experiments O and P, with $O \supset P$, and let M be a Markov matrix with $OM = P$. Then $V(O) \geq V(OM) = V_s(O, M) \geq V_n(O, M)$.*

The first inequality shows the welfare effect for sophisticated agents, which follows immediately from the Blackwell result. In the Blackwell setting, the agent has a choice between the original and the garbled experiment. Since it is impossible to reverse one's bias, a biased agent does not actually have this choice and so will always be worse off. The second inequality says that a naive agent is (weakly) worse off than a sophisticated given the same bias. This arises from two observations. One, both the sophisticated and the naive receive the same signals with the same likelihoods. Two, for each signal, the sophisticated is aware of its lower information value, which allows him to make better decisions. It follows that:

Proposition 1: *Take any information experiment P. A sophisticated agent's utility from P is (weakly) larger than the naive agent's utility, but weakly worse than an agent that doesn't suffer from any misperception.*

For our specific choice problem, a stronger statement is possible. We say that the magnitude of the bias increases if k_a, k_b, or both, decrease, i.e. when the off-diagonal of the garbling matrix M

increases.

Result 4: *Any signal distortion function where* $d_a(\pi(x_1,\omega),x_1,\mu) \neq \pi(x_1,\omega)$ *for some* $\omega \in \Omega$ *(weakly) reduces the expected utility if* $x_1 \in X_1$ *is chosen at* $\mu = p$*. It strictly reduces the expected utility if the action profile is signal sensitive. The welfare loss is increasing in the magnitude of the bias.*

Any perception bias is bad news for an agent who conditions his choices on signals. It leads to more "mistakes" in period 2 choices. As the misperception increases, i.e. d_a moves further away from π, the mistakes increase and the expected utility falls. The welfare consequences of a perception bias can thus be ordered for a given direction of bias.

Looking at this from the perspective of the distortion function d, the following Result 5 shows that we can interpret an increase in the magnitude of the bias as a reduction in the differences of the signal probabilities across states. As the bias increases, the likelihood of receiving a given signal converges across states. Given the previous result, this leads to a welfare loss.

Result 5: *An increase in the magnitude of the bias for some* $x_1 \in X_1$ *and* μ *leads to decrease in*

$$d_a(\pi(x_1,\omega),x_1,\mu) - d_a(1-\pi(x_1,\omega),x_1,\mu).$$

Using the earlier example, we can illustrate the welfare consequences of a bias for a signal sensitive action profile:

Example 2: Take the setting from Example 1 and suppose that the agent has a confirmation bias towards B. This implies that he perceives $b(x_I)$ too often, i.e. $k_a(S(x_I),\mu) = 1-\delta$, but never misperceives $a(x_I)$, i.e. $k_b(S(x_I),\mu) = 1$. The following graph plots the utility frontier for $\delta = 0.2$ which increases the likelihood of receiving a b signal in state B from $1-\pi(x_I,B)$ to $1-\pi(x_I,B)+0.2\pi(x_I,B)$ and in state A accordingly.
The utility frontier, as seen in Figure 3.2, is strictly lower for any interval in which x_I is actually chosen and identical otherwise. It does not affect the expected utility for more extreme priors, even if x_I was chosen as the posterior will always remain in the same partition regardless of the signal realization. In contrast, the effect is strictly negative for the interval in which the agent conditions the choice on the signal.

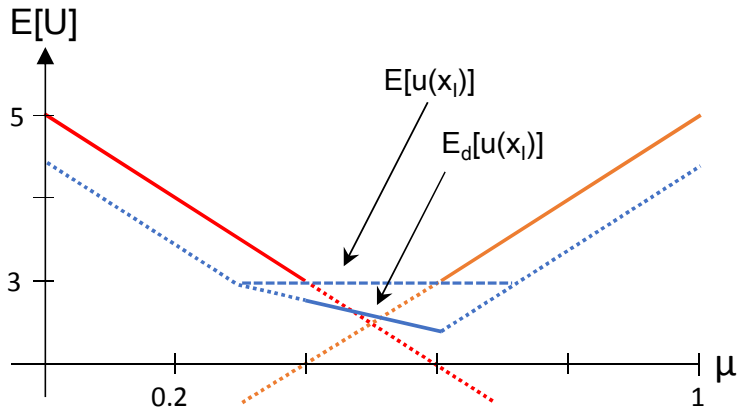


Figure 3.2: Utility frontier with confirmation bias

3.6.1 Biased Prior

We now introduce an alternative source of suboptimal choice: the agent holds an incorrect prior, namely $\mu \neq p$. The prior may be wrong as a consequence of a variety of behavioral and or updating mistakes in the past. If $\mu = p$ the agent will take optimal choices. If, however, $\mu \neq p$, then the true expected utility does not necessarily match the one evaluated by μ.

Equation (3.1) illustrates how a different μ can lead to different choices. An inaccurate μ puts too much weight on one of the states and thus favours actions appropriate for that state. As it realizes with a different probability p, the choice might be suboptimal. It follows:

Result 6: *For any $\mu \geq p$, expected utility is (weakly) decreasing in μ. Equivalently, for $\mu \leq p$ expected utility is increasing in μ.*

3.6.2 Interaction of Biases

So far, we have learned that, in isolation, both biases have a negative effect on outcomes. Additionally, an increase in the magnitude of the biases exacerbates the problem. One might expect that the interaction of both reduces expected utility even further, a 'double whammy'. However, it turns out that this is not the case. To the contrary, an agent with both, a biased perception and a biased prior, can be better off than an agent who only suffers from one of the two. The fundamental reason behind this is that the perception bias shuffles the otherwise straight-forward ranking of experiments at different priors to a different degree. We analyse both directions, i.e. fixing one bias and increasing the other - and determine when a welfare ordering can be established.

Adding a Biased Priors to a Biased Perception

In this section, we assume that the agent's perception is biased 'from the start'. This can either be specific to the experiment - i.e. some experiments generate signals that are more likely to be misperceived than others - or the same across all experiments. We investigate how a bias in prior interacts with the misperception. Breaking the established pattern, we first present an example and then generalize this insight.

Example 3: Reconsider the setting of Example 2 with an agent exhibiting confirmation bias. If the agent is unaware of the bias, the utility frontier is not convex in μ. This arises due to the merely perceived indifference between action x_B and x_I - given optimal period 2 actions - at $\mu = 0.4$. Both achieve an expected utility of 3. The "true" expected utility from x_I is only 2.88 because the agent perceives $b(x_I)$ too frequently and thus takes the period 2 action x_B more often in state A than he would without the bias. The agent should strictly prefer action profile (x_B, x_B). The equivalent situation applies to (x_A, x_A) and μ close to 0.6.

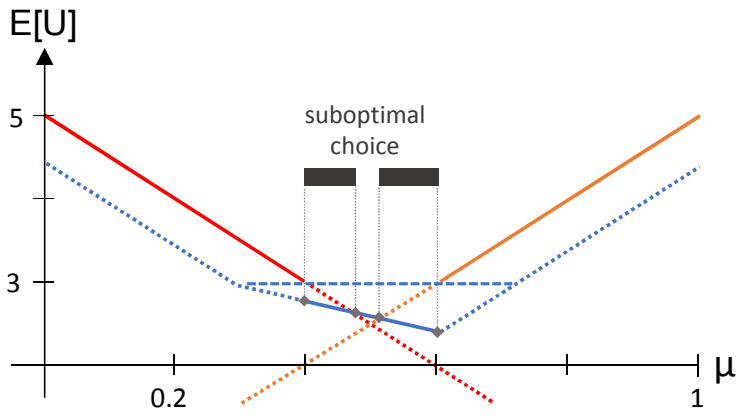


Figure 3.3: Non-convex utility frontier

Now consider the following: if the true p is just below 0.6, then the actually attainable expected utility of an agent with any $\mu > 0.6$ is strictly higher compared to his expected expected utility with $\mu = p$. The choice (x_A, x_A) which would be suboptimal for a fully rational agent, actually yields a strictly better outcome than the "rational choice" x_I. The agent with a biased prior and perception bias fares better than an agent with only perception bias because his suboptimal choice has a signal structure that is less (or not at all) affected by the perception bias.

To understand the conditions under which such a situation can arise, we introduce the

following relation between action profiles:

Definition 2 (Worst-case dominance)**:** *An action profile* $(y_1, \{y_a, y_b\})$ *worst-case dominates* $(x_1, \{x_a, x_b\})$ *at* $\mu \in (0,1)$ *if*

$$\min_{s \in \{a,b\}} \mu u(x_1, x_s)|A) + (1-\mu)u(x_1, x_s|B) < E[u(\mathbf{y})|\mu].$$

In words, worst-case dominance compares an action profile $\mathbf{y}$ to the scenario where the agent gets the 'worst' signal realisation for the profile $\mathbf{x}$; meaning the signal that leads him to make the worst choice in his contingent plan $\{x_a, x_b\}$ evaluated by the expected utility. Clearly, this worst signal depends on the prior. If the prior is sufficiently close to 1, the worst signal is b and for a prior close to 0, the worst signal is a. Notice that $\mathbf{y}$ may or may not be preferred to $\mathbf{x}$ when worst-case dominance holds.

The usefulness of this definition lies in the fact that the concept is independent of the state of the world and enables us to identify which action profiles can be made worse relative to another profile.

Result 7: *For any signal sensitive action profile* $\mathbf{y}$ *that worst-case dominates* $\mathbf{x}$ *at* $\mu \in [0,1]$*, there exists a distortion* d *with associated distortion matrices* M_x*,* M_y *such that*

$$E_d[u(\mathbf{x})|\mu] < E_d[u(\mathbf{y})|\mu].$$

If an action profile is worst-case dominated, it is possible to create a distortion that would rank another profile strictly higher in terms of expected utility. Nevertheless, the result is quite weak since it might require a distortion that targets a specific type of signal: the signals $S(x_1)$ generated by experiment x_1. Depending on the scenario, this is hard to justify in practice in a literal way. But we can think of it as a particular experiment yielding harder-to-interpret results while another experiment has very straight-forward outcomes. In this context, worst-case dominance tells us something about how the ranking of alternatives can change when one experiment generates very ambiguous signals. We can now state the result that tells us the precise condition when adding a bias in prior can make an already perception-biased agent better off.

Proposition 2: *For any signal sensitive action profile* $\mathbf{x}$ *that is chosen at* μ *and any action profile* $\mathbf{y} \neq \mathbf{x}$ *that is chosen at* $\mu' \neq \mu$*, there exists a distortion* d *with associated distortion matrices* M_x, M_y *such that* $E_d[u(\mathbf{y})] > E_d[u(\mathbf{x})]$ *when* $p = \mu \neq \mu'$ *if and only if* $\mathbf{y}$ *worst-case dominates* $\mathbf{x}$ *at* μ*.*

At least for a relatively unrestricted class of distortions d, we can always find one that makes an agent with an incorrect belief $\mu' \neq p$ better-off than an agent with a correct belief as long as there is sufficient potential variation in payoffs. The next corollary shows that this does not hinge on distorting only one experiment. With a notion very similar to worst-case dominance, we can extended the result to a setting where all signals are distorted equivalently. In the context of our example of a lab technician that might enter incorrect results, the tendency to make such a mistake is now equalized across tests rather than one test being more prone to transcription errors than another.

Definition 3 (Simple dominance)**:** *An action profile* $(y_1, \{y_a, y_b\})$ *simply dominates* $(x_1, \{x_a, x_b\})$ *at* $\mu \in (0,1)$ *if*

$$\mu u(x_1, x_a)|A) + (1-\mu)u(x_1, x_a|B) < \mu u(y_1, y_a)|A) + (1-\mu)u(y_1, y_a|B), \text{ or}$$
$$\mu u(x_1, x_b)|A) + (1-\mu)u(x_1, x_b|B) < \mu u(y_1, y_b)|A) + (1-\mu)u(y_1, y_b|B), \text{ or both.}$$

This definition breaks each action profile down into the simple profiles that can be derived from it and compares their expected utility. While simple dominance is not directly comparable to worst-case dominance, the idea is similar. Simple dominance compares the expected utility at μ of each of the two simple profiles derived from the action profiles that are being contrasted. Worst-case dominance compares the expected utility of an action profile with the expected utility of the worst simple profile generated from the action profile it is contrasted against.

Returning to our example, simple dominance compares two treatment methods fixing a particular test result without taking into account the information value. If one method is inferior in at least one of those cases, we can find a type of mistake the lab technician could commit equally across those tests such that the treatment optimal in the absence of mistakes is inferior if the mistake is present. In contrast, worst-case dominance compares a test under ideal conditions with one where the lab technician commits the worst possible type of mistake from an ex-ante perspective.

Corollary 8: *For any signal sensitive action profile* $\mathbf{x}$ *that is chosen at* μ *and any action profile* $\mathbf{y} \neq \mathbf{x}$ *that is chosen at* $\mu' \neq \mu$*, there exists a distortion* d *with associated matrices* $M_x = M_y$ *such that* $E_d[u(\mathbf{y})] > E_d[u(\mathbf{x})]$ *when* $p = \mu \neq \mu'$ *if and only if* $\mathbf{y}$ *simply dominates* $\mathbf{x}$ *at* μ*.*

Proposition 2 and Corollary 8 provide a tight characterisation for utility reversals, accounting for first period action's direct utility effect. The reason behind these results is that the naive agent

has an incorrect view about the information value from x_1 relatively to some alternative y_1. This is captured in the following result:

Result B3: *Take two experiments O and P, with perception biases M_O, M_P. For the naive agent, neither the ordering (according to informativeness) of O and P nor the ordering of OM_O and PM_P determines, or is determined by the agents expected utility ranking of the experiments, $V_n(O, M_O)$ and $V_n(P, M_P)$.*

To understand this statement, recall that with rational agents, Blackwell's result B1 says that if O is more informative than P, then the agent is better off from O. For naive, the fundamental underlying experiment O can be more informative than P, yet may yield lower utility as he ultimately faces a different experiment due to his perception bias. However, even if we take into account what the naive's effective experiments are, OM_O and PM_P, it is still possible that the more informative experiment OM_O can have lower utility. This is due to the fact that the agent does not recognize the true information value of the experiments. It follows that Blackwell's main result on the comparison of experiments does not holds for biased agents.

Given the fact that there is no first period action in Blackwell, this result highlights when the worst-case dominance definition is trivially satisfied, and how first period actions' instrumental- and direct utility value can be separated. First, suppose that the information experiment from x_1 is more informative than from y_1, $P_{x_1} \supset P_{y_1}$. If the direct utility from x_1 is lower than from y_1 at p, then worst-case dominance will always be satisfied by Result B3. However, if the utility from x_1 is higher at p, then we need to check whether worst-case-dominance is satisfied. Similarly, when x_1 is less informative, $P_{x_1} \subset P_{y_1}$, but it is precisely chosen for it's direct utility, we again need to check worst-case dominance.

To understand the role of naivety, we present the Blackwell result for sophisticated agents:

Result B4: *Take two experiments O and P, with perception biases M_O, M_P. $OM_O \supset PM_P$ if and only if $V_s(O, M_O) \geq V_s(P, M_P)$.*

The result follows from the observation that the sophisticated but biased agent's utility of O is equal to the rational agent's utility, who faces OM_O: $V_s(O, M_O) = V_s(OM_O, I) = V(OM_O)$. Clearly, this statement has no implication for the ordering of O and P itself. Observing an agent who takes an 'objectively' less informative experiments does therefore not imply that he is making an incorrect choice.[8] But this also implies that any sophisticated agent cannot be better off from

[8] To see this take a fully informative $n \times n$ vector P and some informative O of equal dimension. Let $M_P = [1/n]_{n \times n}$ i.e. the matrix filled only with $1/n$ which will turn P completely uninformative and keep the information value of O by setting $M_O = I$, so that $O \subset P$ but $OM_O \supset PM_P$.

an incorrect prior. The incorrect prior may cause him to choose an alternative action y_1 with a different underlying experiment. Since he correctly accounted for the lower-information value arising from his perception bias for any information experiment, this simply makes him choose worse first period actions.

Adding a Biased Perception to a Biased Priors

We now turn the previous analysis around. Taking the agent's biased prior as given, we show under which conditions adding or increasing a perception bias can make him better off. This result turns out to be more clear-cut. Adding any type of perception bias makes an agent worse-off if his chosen action profile is not 'too suboptimal' given the true p. Otherwise, there exists some type of perception bias that can improve the agent's welfare.

Proposition 3: *Let $\boldsymbol{x} = (x_1, \{x_a, x_b\})$ be a signal-sensitive action profile chosen at $\mu \neq p$.*

1. *If $E[u(\boldsymbol{x})|p] > \max\{E[u(x_1, x_a)|p], E[u(x_1, x_b)|p]\}$ then increasing the degree of perception bias strictly decreases expected utility for any type of perception bias.*
2. *If $E[u(\boldsymbol{x})|p] < \max\{E[u(x_1, x_a)|p], E[u(x_1, x_b)|p]\}$, then increasing the degree of perception bias strictly increases expected utility for some type of bias.*

Statement (1) says that the signal-sensitive $\mathbf{x}$ is preferred by an agent with the correct prior p over choosing the simple profile where the agent always chooses x_a or x_b in the second period. Adding noise to the signals is equivalent to putting more weight on a simple profile and strictly lowers μ's payoffs. This is exactly what any distortion function does.

Statement (2) states that one of the simple action profiles is preferred to the contingent one by the agent with belief p. In that case, the agent is strictly better off the more often he receives the respective signal realization that makes him choose such a profile.

We can also interpret Proposition 3 in terms of how severe the bias in prior is. To see this, recall that the action profile is signal sensitive at μ for exactly the reason that the posteriors following the signal fall into different parts of the partition that prescribes optimal second-period behavior. If we push up the prior, then after any signal action x_a is relatively more appealing compared to x_b than before. Pushed far enough the agent only wants to take x_a regardless of the signal realization; x_a is, after all, strictly better in state A. At this point, a perception bias can help the agent. If the prior is too close to p he is always worse off due to losing information. Finally, it should also be clear form this discussion that for every signal sensitive profile, there always exists a region for either case.

3.6.3 Martingale Bias

In the previous sections, the bias could generate a drift in beliefs whenever the perception bias was unbalanced between signals. A rational observer aware of the bias would form an ex-ante expectation over the posterior that is different from the prior. We now restrict the perception bias to what we call a 'martingale bias': a bias such that the agent as well as the observer form the same expectation over posteriors. The martingale bias represents a simple random error in perception, that occurs indiscriminately of the signal type.[9] This section will help to qualify our previous Propositions. In contrast to Proposition 3, we show that under such a mistake, the agent with a biased prior can never be better off when a martingale perception bias is added. Changing the prior for an agent with such a perception bias can still be beneficial, however.

Consider a naive agent that thinks the signal strength is the true undistorted $\pi(x_1, \omega)$. Let $k_a(x_1, \mu) = k_b(x_1, \mu) \equiv \kappa \in [\frac{1}{2\pi(x_1,A)}, 1]$. We can write the agent's expected utility as

$$\begin{aligned} \mu \quad & [\kappa\pi(x_1, A)u(x_1, x_a|A) + (1 - \kappa\pi(x_1, A))u(x_1, x_b|A)] \\ +(1-\mu) \quad & [(1 - \kappa(1 - \pi(x_1, B)))u(x_1, x_a|B) + \kappa(1 - \pi(x_1, B))u(x_1, x_b|B)] \end{aligned} \tag{3.3}$$

From state independence, we know that $\kappa \leq 1$. Since the martingale bias represents a pure random error, in the extreme the signal should be pure noise and thus $\kappa \geq \frac{1}{2\pi(x_1,A)}$.
As can be seen from Equation (3.3), if the true probability of receiving an a signal in state A is $\kappa\pi(x_1, A)$, but the agent believes this to be $\pi(x_1, A)$, there is too much weight on $u(x_1, x_a|A)$ and $u(x_1, x_b|B)$. The agent expects to achieve the high outcomes too often which overstates his perceived expected utility of the action profile. It follows that as the perception bias becomes worse, i.e. κ decreases, the agent is worse off. Moreover, this is true for any true p and any prior belief μ, so that there is no way that adding a martingale perception bias to a bias in prior is welfare improving:

Corollary 9: *For any $\kappa \in [\frac{1}{2\pi(x_1,A)}, 1]$ and any signal sensitive action profile $\{x_1, \{x_a, x_b\}\}$ chosen at μ, a decrease in κ decreases the expected utility for any μ and $p \in (0, 1)$.*

Hence the restriction to a martingale bias yields a stronger result than Proposition 3. The martingale bias adds noise and thus reduces the signal strength equally for all signal types. Independent of the true state, this leads to strictly worse period 2 choices. However, it is still possible that under such random error perception bias, the agent can benefit from an incorrect

[9] i.e. the garbling matrix is symmetric $M = \begin{bmatrix} m & 1-m \\ 1-m & m \end{bmatrix}$

prior:

Corollary 10: *For any signal sensitive action profile* $\mathbf{x}$ *that is chosen at* μ*, and any action profile* $\mathbf{y} \neq \mathbf{x}$ *that is chosen at* $\mu' \neq \mu$*, there exists a* κ *such that* $E[u(\mathbf{y})] > E_d[u(\mathbf{x})]$ *when* $p = \mu \neq \mu'$ *if and only if*

$$\frac{\mu}{2}\left[u(x_1, x_a)|A) + u(x_1, x_b)|A)\right] + \frac{1-\mu}{2}\left[u(x_1, x_a)|B) + u(x_1, x_b)|B)\right] < E[u(\boldsymbol{y})|p]. \tag{3.4}$$

Like Proposition 1 this assumes that the bias applies only to the chosen profile $\mathbf{x}$. We can also apply the bias symmetrically as long as we respect the condition that we can at most generate pure noise (rather than achieve a negative correlation). This means the experiment with the less precise signals determines the limit up to which we can garble symmetrically. We then get the following result:

Corollary 11: *For any signal sensitive action profile* $\mathbf{x}$ *that is chosen at* μ*, and any signal-sensitive action profile* $\mathbf{y} \neq \mathbf{x}$ *that is chosen at* $\mu' \neq \mu$ *and for* $\kappa_{min} \equiv \max\{\frac{1}{2\pi(x_1,A)}, \frac{1}{2\pi(y_1,A)}\}$*, there exists a* κ_{max} *with* $1 > \kappa_{max} > \kappa_{min}$ *such that* $E_d[u(\mathbf{y})] > E_d[u(\mathbf{x})]$ *when* $p = \mu \neq \mu'$ *for all* $\kappa \in [\kappa_{min}, \kappa_{max})$ *if and only if*

$$E[u(x)|p, \kappa_{min}] < E[u(y)|p, \kappa_{min}].$$

In other words, if we want to check whether an agent with prior μ' could be better off than one at the true p, we have to compare $\mathbf{y}$ to $\mathbf{x}$ at the extreme point where one of the two information experiments yields just noise. If at this point $\mathbf{y}$ is indeed preferred, such a reversal in the ordering of $\mathbf{x}$ and $\mathbf{y}$ is possible. But of course, this could already happen at a less extreme distortion.

In this section, and indeed in all previous ones, the agent is overconfident in the signal as he fails to discount the perception bias. The martingale property makes this quite similiar to a related major updating mistake found in the literature: in inference problems agents tend to sometimes over- and sometimes under-infer from a given sample of signals. For instance in the classical paper by Griffin and Tversky (1992) agents over-infer from small sample sizes but under-infer from large ones.[10] Indeed one can find a corresponding martingale bias setup for someone who over-infers: The two agents perceive the signal to be of the same strength and will receive the same signal with the same likelihood. However, the key difference between the two mistakes is that over- / under-inference takes the true signal process as given and only changes the perception

[10]For a current, extensive summary of the experimental evidence on this topic, see Appendix B of Benjamin et al. (2013).

of signal strengths, whereas the distorted transmissions fixes the perception of signal quality but alters the actual probabilities with which signals are perceived. By changing the perceived signal strength, anything can happen in terms of welfare consequences. This lack of structure leads us to relegate the discussion to a footnote.[11]

3.6.4 Sophisticated vs Naive Agents

When the prior is correct, we have seen that a sophisticated agent is always weakly better off than a naive agent. We finish this section with a short example that highlights that this no longer needs to be true when μ differs from p. A sophisticated agent is aware of the reduced information value of each experiment and thus may want to take a different choice than a naive agent. If the true ordering of experiments under μ is different than under p, such behavioral adjustments can be suboptimal.

Example 4: Example 1 serves again as the basis. Suppose that the agent suffers from a martingale bias and tends to misperceive both a and b signals with probability $1 - \kappa = 0.7$. Suppose further that $\mu = 0.57$. Compared to the naive agent, the sophisticated is aware of the lower information value of x^I and opts for the simple profile (x^A, x^A). If $\mu = p$ he clearly does better than the naive who still chooses the information experiment in the first period. The naive agent, however, can do better than the sophisticated when the actual p is lower, take for example $p = 0.4$. His naivety causes him to pick the information experiment and subsequently to take x^B whenever he receives an b signal. For a low enough p this is always better than the sophisticated's choice of (x^A, x^A).

[11] With over- or under-reaction to information, the agent can almost trivially achieve a higher utility regardless of which bias is adjusted:
First, suppose the sets of actions are $X_1 = \{x_1\}$ and $X_2 = \{x_a, x_b\}$. First, take the wrong prior as given: (1) Let the signal sensitive action profile $\mathbf{x}$ be optimal at p whereas the agent with prior $\mu \neq p$ prefers the simple profile (x_1, x_a). Then the agent is better off by exaggerating the signal quality so that his second period profile becomes signal sensitive. (2) Similarly, suppose the simple action profile (x_1, x_a) is optimal at p. Then an agent with $\mu \neq p$, who prefers the signal sensitive profile, can be better off by inferring too little if he prefers (x_1, x_a) over (x_1, x_b) in the absence of any signal.
Next take information processing bias as given: (3) Suppose that due to over-inference the agent prefers the conditional profile when it is optimal to choose (x_1, x_a). Changing the correct prior to some $\mu > p$ where the agent prefers (x_1, x_a) improves his utility.
The only non-trivial case (4) is under-inference and changing the correct prior. Suppose $\mathbf{x}$ is optimal at p but due to under-inference the agent with the correct prior prefers $\mathbf{z}$. Let $\mathbf{y}$ be the agent's preferred choice at some $\mu \neq p$. If the actual signal strength of $S(y_1)$ is sufficiently larger than of $S(z_1)$, then shifting the prior from p to μ helps the agent. To see why, notice due to under-inference his response to the signal is attenuated. However, the likelihood of receiving the correct signal under y_1 is higher (than he thinks) which will 'too often' select the 'correct' action from his conditional profile.

3.7 Example: Investment with Information Acquisition

We finish this paper with a simple investment problem that captures the fundamentals of the entrepreneur example outlined in the introduction.

An agent can invest his income w in either a risk-free or a risky asset, both having a unit price. There are two states of the world $\Omega = \{A, B\}$. The risk-free asset's return in any state of the world is normalised to zero. The risky asset yields a positive return of r_h in state A but a negative return of r_l in state B. The agent has a prior $\mu_0 = \Pr(\omega = A)$ and is risk averse with $U(c) = \ln(c)$. Before investing, the agent has the opportunity to purchase a signal $s \in \{a, b\}$ at a cost $c > 0$ that fully reveals ω. He then decides on the fraction of income that he invests in the risky asset, denoted by x. Finally, let the agent's information perception bias be of the following structure: Whenever he receives a b signal - a signal that tells him not to invest - he might misinterpret it as an a signal with a probability δ. On the contrary, a signals are always perceived correctly.

We now show how holding an incorrect prior can help mitigate the bias in information perception through avoiding information. First, we solve for x as a function of the posterior μ:

$$x(\mu, r_h, r_l) = \begin{cases} 0 & \text{if} \quad \mu r_h + (1-\mu) r_l \leq 0 \\ 1 & \text{if} \quad \mu \geq -r_l(r_h+1)/(r_h - r_l) \\ \frac{\mu r_h + (1-\mu) r_l}{-r_l r_h} & \text{otherwise} \end{cases}$$

Let $\Pr(\omega = A) = p = 0.55$. Let the risky asset's return be $r_h = 60\%$ in state A and $r_l = -95\%$ in state B. The agent's wealth is $w = 1.2$. The cost of acquiring information is $c = 0.25$. Furthermore, the agent's information bias is $\delta = 0.4$.

Notice that the expected return of the asset is $0.55 \cdot 0.6 - 0.45 \cdot 0.95 = -0.0975 < 0$. When $\mu = p$, an agent, who does not acquire information, would never invest in the risky asset. If the agent acquires the fully informative signal, he will invest everything into the risky asset when he observes an a signal, and nothing at all following a b signal.

An unbiased agent (with the true prior) will acquire information since the information is cheap enough, $\mathbb{E}\, U(\text{acquire info}) = p \ln((1+r_h)(w-c)) + (1-p)\ln(w-c) = 0.21 > \ln(1.2) = \mathbb{E}\, U(\text{not acquire})$.

A naive agent with misperception bias makes the same choices as an unbiased one. When his prior is correct, his true utility from acquiring information is only $\mathbb{E}\, U_{true}(\text{acquire info}|\mu = 0.55) =$

-0.33 since he sometimes invests his whole income when the state is B.[12] If he instead holds a rather incorrect prior, i.e. $\mu = 0.9$, the information becomes of little value.[13] He consequently chooses not to acquire it and actually does better than if he had the correct prior.[14] While holding an initially more extreme belief, his posterior belief is lower than the naive agent who observed a (possibly wrong) a signal. This makes him invest only 78% of his total income, which is useful when the negative returns of an asset are large. On top of that, he benefits from saving on the information cost.

If we interpret the cost of acquiring information either as the direct cost of going to university or the opportunity cost of not working, this investment problem become analogous to the entrepreneur example that was outlined at the beginning. If the would-be entrepreneur holds a very inflated view about his abilities, he starts his business immediately (invests without information). If not, he goes to university. Suffering from a perception bias, he might then become even more convinced about himself, leading him to gamble even more on his endeavour. In expectation, the possibly useful information hurts him if the probability of misinterpreting it is too high.

3.8 Discussion

We have analyzed the interaction of two fundamental mistakes in information processing. We have seen how and when one bias can alleviate the problems caused by the other. The recent worrying rise of the anti-vaccination movement, with their tendency to disregard surmounting scientific evidence in favour of circumstantial observations that autism happens after (one of many) childhood infections, is just one example for the potential importance of these biases.

The idea that a mistake on one dimension may counteract some failure on another is clearly not new (Compte and Postlewaite (2004), De La Rosa (2011), etc.), yet to our knowledge it has never been explored in this basic updating setting. In contrast, the recent literature on confirmatory bias (Clements (2013), Roland G. Fryer et al. (2016)) as well as biases in information processing (Andreoni and Mylovanov (2012), Kartik et al. (2016)) focuses solely on beliefs, how beliefs may converge or diverge in group, or whether information can be aggregated in social settings. It does not focus on the dynamic aspect of payoff-relevant decisions as information experiments and

[12] This is true for any $\mu \in [0.47, 0.89]$

[13] The results hold for any extreme prior in the interval $[0.89, 0.925]$. An even higher prior causes too much investment, lowering utility.

[14] His perceived expected utilities are: $\mathbb{E}\,U_{perceived}(\text{acquire info}|\mu = 0.9) = -.39 < -0.37 = \mathbb{E}\,U(\text{not acquire}|\mu = 0.9)$ while is actual utility (evaluated at the true prior) is $\mathbb{E}\,U_{true}(\text{not acquire}|\mu = 0.9) = -0.22$.

their consequences on future payoffs. In that sense, we think of this paper as an extension of Blackwell (1951) to behavioral mistakes in information processing, for which we provide a full characterisation.

3.9 Appendix A: Proofs

Proof of Result 2. Convexity follows directly from the fact that maximum expected utility is convex in μ. This result extends to two period model with intermediate signals.

Let $\vec{\mu}$ be row vector of belief about the different states Ω, $\vec{\mu} = (\Pr(\omega = A), \Pr(\omega = B))$. Define $\vec{u}(x)$ as the state contingent vector adjusted by signal probability and signal response:[15]

$$\vec{u}(x) = \begin{pmatrix} \pi(x_1, A)u(x_1, x_a|A) + (1 - \pi(x_1, A))u(x_1, x_b|A) \\ \pi(x_1, B)u(x_1, x_a|B) + (1 - \pi(x_1, B))u(x_1, x_b|B) \end{pmatrix}$$

Define maximum expected utility as a function of $\vec{\mu}$ as $g(\vec{\mu}) = \max_{x \in X} \vec{\mu} \cdot \vec{u}(x)$ and notice how this expression coincides with the rewritten expected utility formulation of equation 3.1.

To show $g(\mu)$ is convex, let x' (x'') optimal choice given the belief $\vec{\mu}'$ ($\vec{\mu}'$), and $\hat{x}$ the optimal choice for the convex combination $\vec{\hat{\mu}} = \lambda\vec{\mu}' + (1 - \lambda)\vec{\mu}''$ with $\lambda \in (0, 1)$. Then

$$\begin{aligned} g(\lambda\vec{\mu}' + (1 - \lambda)\vec{\mu}'') &= (\lambda\vec{\mu}' + (1 - \lambda)\vec{\mu}'') \cdot u(\hat{x}) \\ &\leq \lambda\vec{\mu}' \cdot u(u') + (1 - \lambda)\vec{\mu}'' \cdot u(x'') \\ &= \lambda g(\vec{\mu}') + (1 - \lambda)g(\vec{\mu}'') \end{aligned}$$

proving convexity. □

Proof of Result B2. By Blackwell's Result B1 it must be that $V(O) \geq V(OM)$. A rational (non-biased) agent would never choose the distorted over the undistorted experiment. A sophisticated agent is not able to choose the undistorted experiment, however. All he can do is to adjust his second period action profile anticipating the lower information value. His utility from O given M is equivalent to a rational agent facing the experiment OM. As a result the following inequalities hold: $V_s(O, M) = V_s(OM, I) = V(OM)$. This establishes the first inequality.

The naive agent acts as if he was fully rational, i.e. he applies the rationally optimal decision rule for P to OM. Consequently, he holds an incorrect posterior after observing the signal - which is equivalent to having wrong beliefs about the likelihood with which he receives a and b signals. He makes (weakly) worse decisions than a sophisticated agent; if not, the sophisticated agent could always imitate him. □

[15] compare this to the simple 1 period convexity proof, where $\vec{u}(x)$ is defined as the state contingent utility vector

Proof of Result 4. Let the (unbiased) information experiment following action x_1 be of the form

$$P = \begin{bmatrix} \pi & 1-\pi \\ 1-\pi & \pi \end{bmatrix}$$

with $\pi = \pi(x_1, A)$. Let experiment P' be the experiment arising from the distortion function with the same action x_1, namely

$$P' = \begin{bmatrix} d_a(\pi, x_1, \mu) & d_b(1-\pi, x_1, \mu) \\ d_a(1-\pi, x_1, \mu) & d_b(\pi, x_1, \mu) \end{bmatrix}$$

If follows that $P \supset P^{'}$, as we can find an M such that $PM = P^{'}$. The corresponding M is:

$$M = \begin{bmatrix} k_a(x_1, \mu) & 1-k_a(x_1, \mu) \\ 1-k_b(x_1, \mu) & k_b(x_1, \mu) \end{bmatrix}$$

where $d_a(\pi, x_1, \mu) = k_a(x_1, \mu)\pi + (1-k_b(x_1, \mu))(1-\pi)$ etc.

According to Result B1, a rational (non-biased) agent would never choose the distorted over the undistorted experiment. The utility arising from the distorted experiment is (weakly) lower.

To show the effect of an increase in the magnitude of the bias, we write the expected utility from x_1 as

$$p\left[(k_a(x_1,\mu)\pi + (1-k_b(x_1,\mu))(1-\pi))\, u(x_1, x_a|A) + (k_b(x_i,\mu)(1-\pi) + (1-k_a(x_i,\mu))\pi)\, u(x_1, x_b|A)\right]$$
$$+(1-p)\left[(k_a(x_1,\mu)(1-\pi) + (1-k_b(x_1,\mu))\pi)\, u(x_1, x_a|B) + (k_b(x_i,\mu)\pi + (1-k_a(x_i,\mu))(1-\pi))\, u(x_1, x_b|B)\right]$$

Suppose that initially k_b decreased. If we decrease it further to $k_b' < k_b$, the linearity property of the expected utility in probabilities implies that the expected utility must also decrease.

The same argument holds for any further decrease in the alternative parameter k_a.

Since there are no cross-interactions between changes of k_a and k_b, the same holds true for any joint change. □

Proof of Result 5. Note that we can write the matrix of the signal probabilities as a garbling of the matrix containing the true signal probabilities:

$$\begin{bmatrix} d_a(\pi, x_1, \mu) & d_b(1-\pi, x_1, \mu) \\ d_a(1-\pi, x_1, \mu) & d_b(\pi, x_1, \mu) \end{bmatrix} = \begin{bmatrix} \pi & 1-\pi \\ 1-\pi & \pi \end{bmatrix} \begin{bmatrix} k_a & 1-k_a \\ 1-k_b & k_b \end{bmatrix}$$

where the matrix on the right is the garbling matrix. Since the garbling matrix needs to be a Markov matrix, we know that $k_a, k_b \in [0,1]$.

Now suppose we increase the magnitude of the bias resulting in a distortion function d' and, contrary to the statement in the result, we have $d'_a(\pi, x_1, \mu) - d'_a(1-\pi, x_1, \mu) \geq d_a(\pi, x_1, \mu) - d_a(1-\pi, x_1, \mu)$. This implies that:

$$\pi k'_a + (1-\pi)(1-k'_b) - (1-\pi)k'_a - \pi(1-k'_b) \geq \pi k_a + (1-\pi)(1-k_b) - (1-\pi)k_a - \pi(1-k_b)$$

where k'_a, k'_b are the elements in the modified garbling matrix reflecting the increase in the magnitude of the bias. This can be rearranged to:

$$(2\pi - 1)(k'_a - (1-k'_b)) \geq (2\pi - 1)(k_a - (1-k_b))$$

But this requires that since $\pi > 1/2$, either $k'_a > k_a$, or $k'_b > k_b$, or both which contradicts the increase in the magnitude of the bias. □

Proof of Result 6. Take the action profile $(x_1, \{x_a, x_b\})$. It's actual expected utility is

$$\begin{aligned} p \quad & [\pi(x_1, A)u(x_1, x_a|A) + (1-\pi(x_1, A))u(x_1, x_b|A)] \\ +(1-p) \quad & [\pi(x_1, B)u(x_1, x_a|B) + (1-\pi(x_1, B))u(x_1, x_b|B)] \end{aligned}$$

If we evaluate this with some $\mu > p$, then the weight on the outcome in state A increases. This has two potential effects. Firstly, according to Result 12, a different action might be chosen in period 2 as the potential posteriors are now different. Furthermore, as follows from Result 1, a different x_1 might become optimal. Denote the potentially altered choice by $(y_1, \{y_a, y_b\})$. Any such alternative choice has the property that:

$$\begin{aligned} &\pi(y_1, A)u(y_1, y_a|A) + (1-\pi(y_1, A))u(y_1, y_b|A) \\ &\quad > \pi(x_1, A)u(x_1, x_a|A) + (1-\pi(x_1, A))u(x_1, x_b|A) \end{aligned}$$

and

$$\begin{aligned} &\pi(y_1, B)u(y_1, y_a|B) + (1-\pi(y_1, B))u(y_1, y_b|B) \\ &\quad < \pi(x_1, B)u(x_1, x_a|B) + (1-\pi(x_1, B))u(x_1, x_b|B) \end{aligned}$$

As otherwise either the choice $(y_1, \{y_a, y_b\})$ would not be optimal as it yields lower expected utility when evaluated by μ or the choice under p would not be optimal since it results in lower utility in both states. It then immediately follows that if any choice of action is different when comparing optimal choices under μ and p, such a μ leads to a utility loss because of suboptimal choices. Furthermore, fixing some $\mu > p$ and iterating this argument for $\mu' > \mu$ yields the result. The equivalent argument applies to $\mu < p$. □

Proof of Result 7. If the agent prefers $\mathbf{y}$ over $\mathbf{x}$, i.e. $E[u(\mathbf{x})] < E[u(\mathbf{y})]$, the statement is trivially true. If it does not hold, we show that there is such a d that makes

$$E_d[u(\mathbf{x})|\mu] < E_d[u(\mathbf{y})|\mu]$$

According to definition 2, we have that $\mu u(x_1, x_s|A) + (1-\mu)u(x_1, x_s|B) < E[u(\mathbf{y})|\mu]$ for some $s \in \{a(x_1), b(x_1)\}$. Suppose wlog that $s = a(x_1)$. This implies that small enough $\epsilon > 0$, we have

$$\begin{aligned}&(1-\epsilon)\left[\mu u(x_1, x_a|A) + (1-\mu)u(x_1, x_a|B)\right] + \epsilon\left[\mu u(x_1, x_b|A) + (1-\mu)u(x_1, x_b|B)\right] \\ &= \mu\left[(1-\epsilon)u(x_1, x_a|A) + \epsilon u(x_1, x_b|A)\right] + (1-\mu)\left[(1-\epsilon)u(x_1, x_a|B) + \epsilon u(x_1, x_b|B)\right] < E[u(\mathbf{y})|\mu]\end{aligned} \tag{3.5}$$

which quite resembles the maximum expected utility representation (3.1).

Now set $k_a(x_1, \mu) = 1$ and $k_b(x_1, \mu) = \frac{\epsilon}{\pi(x_1, A)}$. This causes the agent to receive the worst-case signal a in state B with probability $d_a(\pi(x_1|B), S(x_1), \mu) = 1 - \epsilon$. Clearly, this also makes him receive the (correct) worst-case signal a in state A with probability $d_a(\pi(x_1|A), S(x_1), \mu) > 1 - \epsilon$. Consequently, the actual expected utility following this distortion is slightly larger than the left-hand side of Equation 3.5 above, but as ϵ becomes sufficiently small, $E_d[u(\mathbf{x})|\mu] < E[u(\mathbf{y})|\mu]$. Finally let there be no distortion for the signals $S(y_1)$, i.e. $d_a(\pi, S(y_1), \mu) = \pi$ for all $\pi \in (0, 1)$. The result follows. □

Proof of Proposition 2. Sufficiency is an almost immediate consequence of Result 7. We know there exists d, such that $E_d[u(\mathbf{y})|\mu] > E_d[u(\mathbf{x})|\mu]$ by worst-case dominance at μ and that an agent with belief μ' chooses $\mathbf{y}$ compared to $\mathbf{x}$. Finally simply set $\mu = p$.

In the other direction, suppose such an improvement is possible. Since any distortion garbles the information, $E[u(\mathbf{y})|p] \geq E_d[u(\mathbf{y})|p]$. But if worst-case dominance does not hold there

is no marginal distribution over $S(x_1)$ such that $E_d[u(\mathbf{x})|p] < E[u(\mathbf{y})|p]$ since $E_d[u(\mathbf{x})|p] \geq \min_{s \in S(x_1)} E[u(x_1, x_s)] > E[u(\mathbf{y})|p]$. A contradiction. □

Proof of Corollary 8. Sufficiency: Similar to Proposition 2, consider the extreme distortion d that only sends the signal

$$\underline{s}_x = \arg\min_s \mu u(x_1, x_s)|A) + (1-\mu)u(x_1, x_s|B)$$
$$\overline{s}_x = \arg\max_s \mu u(x_1, x_s)|A) + (1-\mu)u(x_1, x_s|B)$$

If simple dominance holds, at least one of those signals must yield lower expected utility under x than under y.

Necessity: Suppose such a distortion exists. Then

$$E[u(y)|p] < E[u(x)|p]$$

while

$$E_d[u(y)|p] > E_d[u(x)|p].$$

We can write the actual expected utility explicitly as a linear combination between the a and the b signal.

$$\begin{aligned} p \quad & \pi(x_1, A)u(x_1, x_a|A) \quad + \quad (1-p)\pi(x_1, B)u(x_1, x_a|B) \\ +p \quad & (1-\pi(x_1, A))u(x_1, x_b|A) \quad + \quad (1-p)(1-\pi(x_1, B))u(x_1, x_b|B) \end{aligned}$$

and similarly for y. Any distortion can either lower terms or shift weight from one to the other. If the distortion is such that weight is added in the sense that both $\pi(x_1, A)$ and $\pi(x_1, B)$ are increased or decreased then if $E_d[u(y)|p] > E_d[u(x)|p]$ it must be that either

$$pu(x_1, x_a|A) + (1-p)u(x_1, x_a|B) < pu(y_1, y_a|A) + (1-p)u(y_1, y_a|B)$$

or

$$pu(x_1, x_b|A) + (1-p)u(x_1, x_b|B) < pu(y_1, y_b|A) + (1-p)u(y_1, y_b|B)$$

or both which guarantees simple. If the distortion lowers both terms i.e. by garbling both type

of signals equally, then $\pi(x_1, B)$ and $(1-\pi(x_1, A))$ increase. If this increase leads to a reversal in the expected utility ranking of $\mathbf{x}$ and $\mathbf{y}$ in the sense that $E_d[u(\mathbf{y})|p] > E_d[u(\mathbf{x})|p]$, then because of linearity this must also be true for the most extreme case with $d_a(\pi(x_1, A)) = d_b(\pi(x_1, B)) = \frac{1}{2}$. But then

$$\frac{p}{2}u(x_1, x_a|A) + \frac{(1-p)}{2}u(x_1, x_a|B) < \frac{p}{2}u(y_1, y_a|A) + \frac{(1-p)}{2}u(y_1, y_a|B)$$

or

$$\frac{p}{2}u(x_1, x_b|A) + \frac{(1-p)}{2}u(x_1, x_b|B) < \frac{p}{2}u(y_1, y_b|A) + \frac{(1-p)}{2}u(y_1, y_b|B)$$

or both which means at least one of them is also true for $d_a(\pi(x_1, A)) = d_a(\pi(x_1, B)) = 1$ or $d_b(1-\pi(x_1, B)) = d_b(1-\pi(x_1, A)) = 1$ which again guarantees simple dominance. □

Proof of Result B3. To see take two separate experiments O and P, with $O \supset P$. First, let M_O be such that $OM_O = P$ and let P be undistorted, i.e. $M_P = I$. It follows that the naive agent is better off from P than O as $V_n(P, I) = V(P) > V_n(O, M_1)$ (the inequality is strict if the agent is signal sensitive). But by continuity this also implies that there exist some M_O, M_P such that $OM_O \supset PM_P$ yet $V_n(P, M_P) > V_n(O, M_O)$. As a result, the naive's expected utility can neither be ordered by the perceived information experiments, nor the actual ones. In the other direction, we know that for $V_n(P, M_P) > V_n(O, M_O)$ it is possible that $OM_O \supset PM_P$, as well as $O \supset P$. □

Proof of Proposition 3. Statement (1) says that the signal-sensitve $\mathbf{x}$ is preferred by an agent with the correct prior p over choosing the simple profile where the agent chooses either always x_a or always x_b in the second period. Adding noise weakens the correlation with the state and the dominated outcomes occur more frequently.Hence the agent with prior μ can never be better off.

Statement (2) highlights that one of the simple action profiles is preferred to the contingent one by the agent with belief p. Assume wlog that the preferred action profile is (x_1, x_a). Any distortion function that increases the likelihood of an a signal will make the agent better off as he will take action x_a as a result, i.e. set $k_a(x_1, \mu) = 1$ and $k_b(x_1, \mu) \in (0, 1)$. Clearly, the agent's expected utility is increasing as $k_b(x_1, \mu)$ decreases to 0. The extreme $k_b(x_1, \mu) = 0$ represents a perception bias where all b signals are interpreted as a signals and the agent always takes the (second) best simple action profile. □

Proof of Corollary 9. A decrease in κ leads to a less informative signal.

$$\begin{aligned} \mu \quad & [\kappa\pi(x_1, A)u(x_1, x_a|A) + (1 - \kappa\pi(x_1, A))u(x_1, x_b|A)] \\ +(1-\mu) \quad & [(1 - \kappa(1 - \pi(x_1, B)))u(x_1, x_a|B) + \kappa(1 - \pi(x_1, B))u(x_1, x_b|B)] \end{aligned}$$

Trivially, the expected utility at μ as written above decreases as there is more weight on $u(x_1, x_b|A)$ and $u(x_1, x_a|B)$. The agent is unaware of this and his choice in unaffected by κ. For any κ, he evaluates the profile as follows:

$$\begin{aligned} \mu \quad & [\pi(x_1, A)u(x_1, x_a|A) + (1 - \pi(x_1, A))u(x_1, x_b|A)] \\ +(1-\mu) \quad & [(1 - \pi(x_1, B))u(x_1, x_a|B) + (1 - \pi(x_1, B))u(x_1, x_b|B)] \end{aligned}$$

The true p could be different from μ, so we just need to show that for any p, a decrease in κ has a negative welfare effect. The actual expected utility is:

$$\begin{aligned} p \quad & [\kappa\pi(x_1, A)u(x_1, x_a|A) + (1 - \kappa\pi(x_1, A))u(x_1, x_b|A)] \\ +(1-p) \quad & [(1 - \kappa(1 - \pi(x_1, B)))u(x_1, x_a|B) + \kappa(1 - \pi(x_1, B))u(x_1, x_b|B)] \end{aligned}$$

Simply notice since $u(x_1, x_b|A) < (x_1, x_a|A)$, as well as $u(x_1, x_a|B) < u(x_1, x_b|B)$, the random noise indeed lowers expected utility for any p. □

Proof of Corollary 10. If: Suppose the inequality in Equation 3.4 holds. Then a distortion of $\kappa = \frac{1}{2\pi(x,A)}$ turns the signals into complete noise. If this applies to the profile $\mathbf{x}$ only, $\mathbf{y}$ yields higher expected utility at p even though without any distortion $E[\mathbf{x}|p] > E[\mathbf{y}|p]$.

Only if: Suppose the property does not hold. The for any distortion $\kappa \in [\frac{1}{2\pi(x,A)}, 1]$, the expected utility from $\mathbf{x}$ at p exceeds the one from $\mathbf{y}$ as for any such κ,

$$E_d[\mathbf{x}|p] \geq \frac{\mu}{2}\left[u(x_1, x_a)|A) + u(x_1, x_b)|A)\right] + \frac{1-\mu}{2}\left[u(x_1, x_a)|B) + u(x_1, x_b)|B)\right]$$

□

Proof of Corollary 11. If: Follows directly as it is the case for $\kappa = \kappa_{min}$ and continuity guarantees that for small enough $\epsilon > 0$ such that it is also the case for $\kappa = \kappa_{min} + \epsilon$

Only If: Note that for $\mathbf{x}$ to be chosen at p, we need that $E[\mathbf{x}|p] > E[\mathbf{y}|p]$ as the agent is unaware of the perception bias. The effect of a decrease in κ is

$$\begin{aligned} -\frac{\partial E_d[u(\mathbf{x})|p]}{\partial \kappa} = \quad & -p\pi(x_1, A)[u(x_1, x_a|A) - u(x_1, x_b|A)] \\ & -(1-p)\pi(x_1, A)[u(x_1, x_b|B) - u(x_1, x_a|B)] \end{aligned}$$

This does not depend on κ itself. Therefore, if

$$\left[\frac{\partial E_d[u(\mathbf{x})|p]}{\partial \kappa}\right]_{\kappa=1} > \left[\frac{\partial E_d[u(\mathbf{y})|p]}{\partial \kappa}\right]_{\kappa=1}$$

then the inequality also holds for all allowable κ. Take the $\kappa_e \in (\kappa_{min}, \kappa_{max})$ such that $E_d[\mathbf{x}|p, \kappa_e] = E_d[\mathbf{y}|p, \kappa_e]$. If this does not exist, then $\mathbf{y}$ is never the best choice which yields the result. If it does exist, then it must mean that

$$\left[\frac{\partial E_d[u(\mathbf{x})|p]}{\partial \kappa}\right]_{\kappa=1} > \left[\frac{\partial E_d[u(\mathbf{y})|p]}{\partial \kappa}\right]_{\kappa=1}$$

But then for all $\kappa \in [\kappa_{min}, \kappa_e)$, the inequality also holds and therefore

$$E[u(x)|p, \kappa_{min}] < E[u(y)|p, \kappa_{min}]$$

as desired. □

3.10 Appendix B: Blackwell '51

This section provides a general introduction to Blackwell, going beyond the 2-stage, 2-signal case, and providing clear definitions of value functions. In Blackwell's comparison of experiments, an agent compares two statistical information experiments, P and Q, which generate a signal about the states of the world $\Omega = \{\omega_1, ..., \omega_n\}$. As before, an information experiment is defined by a $n \times n_K$ Markov matrix K, where element K_{ij} refers to the probability of receiving a signal $s = j$ in state ω_i. We denote the set of possible signals from K by $S^K = \{1, 2, \ldots, N_K\}$.

The agent is endowed with utility function $u(x, \theta_i)$, defined over $A \times \Omega$. After receiving a signal s^K, he chooses an action x from his finite action set X. His prior is denoted by the vector $\mu_0 = \{\mu(1), ..., \mu(n)\}$, whereas his posterior after receiving signal s from experiment K is denoted by $\mu(s^K)$. The agent maximizes expected utility: $\max_{x \in X} \sum_i \mu(i|s^K) u(x|\omega_i)$. Denote

the solution to this problem by $x^*(\mu(s^K))$ and the corresponding value function $V(\mu(s^K))$.

The agent's probability of receiving an s^P signal is $\pi(s^P) = \sum_{i \in n} P_{is^P}\mu_0(i)$. and expected utility from observing experiment P is $V(P) = \sum_{s^P \in S^P} \pi(s^P)V(\mu(S^P))$.

Definition 4: *Let P and Q be two experiments. We say that experiment P is more informative than Q, denoted by $P \supset Q$, if there is a $n_P \times n_Q$ Markov matrix M with $PM = Q$.*

This definition highlights how the information experiment P is garbled into Q by the matrix M; as if noise was added to create a less informative experiment. From this statistical notion, Blackwell derives his famous result:

Result B5: *Let P and Q be two $n \times n_P$ and $n \times n_Q$ Markov matrices. $V(P) \geq V(Q)$ if and only if there is a $n_P \times n_Q$ Markov matrix M such that $PM = Q$.*

Simply put, a more informative experiment implies a higher utility. Surprisingly, the converse is also true: If one experiment yields more expected utility than another, it must be that it is more informative.

In Blackwell's setup the agent is rational. He fully understands the signal generating structure of the two experiments and correctly perceives each signal. The garbling matrix doesn't represent the agent's misperception but is a tool to order the experiments by informativeness. Moreover, the agent does not take any first period choice. All that occurs is a welfare comparisons of the two experiments. Needless to say, if the rational agent had to make a choice in Blackwell, he would prefer the more informative experiment.

In order to talk about a Blackwell setting when agents misperceive information, we need a new function representing the agents' expected utility from the experiments. Suppose the unbiased, true information experiment is P but the agent misperceives some signals according to matrix M. As a result, the agent faces an information experiment PM. We denote the expected utility of a naive agent unaware of his perception bias by $V_n(P, M)$, where the first argument always indicates the original signal generating process and the second argument the agent's misperception. This agent believes the information experiment is P and thus chooses his action as if he was rational facing P. If the agent is sophisticated, that is he is cognisant of his misperception M, we write his utility as: $V_s(P, M) = \sum_{s^P \in S^{PM}} \pi(s^{PM})V(\mu(S^{PM}))$.

3.11 Appendix C: Decision Problem Appendix

Result 12: *For any action $x_1 \in X_1$ and any $x_2 \in X_2$ such that for some $\mu(s) \in (0,1)$ we have $E[u(x_1, x_2)|\mu(s)] \geq E[u(x_1, y)|\mu(s)]$ $\forall y \in X_2$, there exists an interval $\mathcal{I} \subset [0,1]$ with the property that x_2 is chosen if and only if $\mu(s) \in \mathcal{I}$.*

Proof of Result 12. If for some $(x_1, x_2) \in X$ we have

$$E[u(x_1, x_2)|\mu(s)] \geq E[u(x_1, y_2)|\mu(s)] \quad \forall y_2 \in X_2$$

then continuity in $\mu(s)$ implies that this is also true for at least some $[\mu(s), \overline{\mu})$ or $(\underline{\mu}, \mu(s)]$. If this interval is $[0,1]$, the proof is complete. If not, there is some μ_j for which $x_j \in X_2$ is the optimal choice. Then either $u(x_1, x_j|A) > u(x_1, x_2|A)$ or $u(x_1, x_j|B) > u(x_1, x_2|B)$. This shows us that for either $\mu > \mu_j$ or $\mu < \mu_j$, x_j achieves a higher expected payoff than x_2. This completes the proof. □

Bibliography

Johannes Abeler, Daniele Nosenzo, and Collin Raymond. Preferences for truth-telling. *working paper*, 2016.

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. The market for "lemons": Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500, 1970.

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 and Tymofiy Mylovanov. Diverging opinions. *American Economic Journal: Microeconomics*, 4(1):209–232, 2012.

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.

Björn Bartling and Nick Netzer. An externality-robust auction: Theory and experimental evidence. *Games and Economic Behavior*, 97:186 – 204, 2016.

Matteo Bassi, Marco Pagnozzi, and Salvatore Piccolo. Optimal contracting with altruism and reciprocity. *Research in Economics*, 68(1):27–38, 2014.

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

Roland Benabou and Jean Tirole. Self- confidence and personal motivation. *Quarterly Journal of Economics*, 117(3):817–915, 2002.

Roland Bénabou and Jean Tirole. Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678, 2006.

Daniel J. Benjamin, Matthew Rabin, and Collin Raymond. A model of non-belief in the law of large numbers. 2013. working paper.

JeanPierre Benoit and Juan Dubra. Apparent overconfidence. *Econometrica*, 79(5):1591–1625, 2011.

JeanPierre Benoit and Don A. Moore. Does the better-than-average effect show that people are overconfident? two experiments. *Journal of the European Economic Association*, 13(2): 293–329, 2015.

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.

Felix Bierbrauer, Axel Ockenfels, Andreas Pollak, and Désiré Rückert. Robust mechanism design and social preferences. *Journal of Public Economics*, 149:59–80, 2017.

David Blackwell. Comparison of experiments. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 93–102. University of California Press, 1951.

David Blackwell and M. A. Girshick. *Theory of Games and Statistical Decisions*. John Wiley & Sons, New York, 1954.

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.

Gary E. Bolton and Axel Ockenfels. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments: Comment. *American Economic Review*, 96(5):1906–1911, 2006.

Jerome S. Bruner and Mary C. Potter. Interference in visual recognition. *Science*, 144(3617): 424–425, 1964.

Markus K. Brunnermeier and Jonathan A. Parker. Optimal expectations. *The American Economic Review*, 95(4):1092–1118, 2005.

Stephen V. Burks, Jeffrey P. Carpenter, Lorenz Goette, and Aldo Rustichini. Overconfidence and social signalling. *The Review of Economic Studies*, 80(3):949?–983, 2013.

Hongbin Cai and Joseph Tao-Yi Wang. Overcommunication in strategic information transmission games. *Games and Economic Behavior*, 56(1):7–36, 2006.

Juan D. Carrillo and Thomas Mariotti. Strategic ignorance as a self-disciplining device. *The Review of Economic Studies*, 67(3):529–544, 2000.

Rory Carroll. Leveling the paying field: La cafe lets patrons choose prices - and hasn't lost cash. https://www.theguardian.com/us-news/2018/jan/28/la-pay-what-you-want-metro-cafe-inequality, 2018. Accessed: 2018-04-09.

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.

Gary Charness, Aldo Rustichini, and Jeroen van de Ven. Self-confidence and strategic behavior. 2014. working paper.

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.

Matthew T. Clements. Is more information better? social learning with confirmatory bias. *working paper*, 2013.

Oliver Compte and Andrew Postlewaite. Confidence-enhanced performance. *The American Economic Review*, 94(5):1536–1557, 2004.

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.

Vincent Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 50(6): 1431–51, 1982.

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.

John M. Darley and Paget H. Gross. A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, 44(1):20–33, 1983.

Leonidas Enrique De La Rosa. Overconfidence and moral hazard. *Games and Economic Behavior*, 73:429–451, 2011.

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.

Christoph Engel and Lilia Zhurakhovska. Conditional cooperation with negative externalities - an experiment. *Journal of Economic Behavior and Organization*, 108:252–260, 2014.

Dirk Engelmann and Martin Strobel. Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, 94(4):857–869, 2004.

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 Urs Fischbacher. Third-party punishment and social norms. *Evolution and Human Behavior*, 25(2):63–87, 2004.

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 and Klaus M. Schmidt. The economics of fairness, reciprocity and altruism - experimental evidence and new theories. In Serge-Christophe Kolm and Jean Mercier Ythier, editors, *Handbook of the Economics of Giving, Altruism and Reciprocity*, volume 1, chapter 8, pages 615 – 691. 2006.

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.

Baruch Fischoff, Paul Slovic, and Sarah Lichtenstein. Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance*, 3(4):552–564, 1977.

Drew Fudenberg and Jean Tirole. *Game Theory*. The MIT Press, Cambridge, Massachusetts, 1991.

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.

Ayelet Gneezy, Uri Gneezy, Leif D. Nelson, and Amber Brown. Shared social responsibility: A field experiment in pay-what-you-want pricing and charitable giving. *Science*, 329(5989): 325–327, 2010.

Ayelet Gneezy, Uri Gneezy, Gerhard Riener, and Leif D. Nelson. Pay-what-you-want, identity, and self-signaling in markets. *Proceedings of the National Academy of Sciences of the United States of America*, 109(19):7236–7240, 2012.

Dale Griffin and Amos Tversky. The weighing of evidence and the determinants of confidence. *Cognitive Psychology*, 24:411–35, 1992.

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

Werner Güth and Steffen Huck. From ultimatum bargaining to dictatorship - an experimental study of four games varying in veto power. *Metroeconomica*, 48(3):262–299, 1997.

Werner Güth and Eric van Damme. Information, strategic behavior, and fairness in ultimatum bargaining: An experimental study. *Journal of Mathematical Psychology*, 42(2-3):227–247, 1998.

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

Chris Guthrie, Jeffrey Rachlinski, and Andrew Wistrich. Inside the judicial mind: Heuristics and biases. *Cornell Law Review*, 56:777–830, 2001.

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.

Jack Hirshleifer. The private and social value of information and the reward to inventive activity. *The American Economic Review*, 61(4):561–574, 1971.

Navin Kartik, Frances Xu Lee, and Wing Suen. A theorem on bayesian updating and applications to signaling games. *working paper*, 2016.

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

Ju-Young Kim, Martin Natter, and Martin Spann. Pay what you want: A new participative pricing mechanism. *Journal of Marketing*, 73(1):44–58, 2009.

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.

Augustin Landier and David Thesmar. Financial contracting with optimistic entrepreneurs. *Review of Financial Studies*, 22(1):177–150, 2009.

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.

Sarah Lichtenstein, Baruch Fischoff, and Phillips Lawrence. Calibration of probabilities: The state of the art to 1980. In Daniel Kahneman, Paul Slovic, and Amos Tversky, editors, *Judgement under uncertainty: Heuristics and biases*, pages 306–334. Cambridge University Press, Cambridge, 1982.

Sandra Ludwig, Philipp C. Wichardt, and Hanke Wickhorst. Overconfidence can improve an agent's relative and absolute performance in contests. *Economic Letters*, 110:193–196, 2011.

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

Ulrike Malmendier and Geoffrey Tate. CEO overconfidence and corporate investment. *Journal of Finance*, 60(6):2661–2700, 2005.

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

Jacob Marschak and Koichi Miyasawa. Economic comparability of information systems. *International Economic Review*, 9(2):137–174, 1968.

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

Ian Mcdonald, Nikos Nikiforakis, Nilss Olekalns, and Hugh Sibly. Social comparisons and reference group formation: Some experimental evidence. *Games and Economic Behavior*, 79(1), 2012.

Michael Mitzkewitz and Rosemarie Nagel. Experimental results on ultimatum games with incomplete information. *International Journal of Game Theory*, 22(2):171–198, 1993.

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

Nick Netzer and André Volk. Intentions and ex-post implementation. *working paper*, 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.

Mathew Rabin and Joel Schrag. First impressions matter: A model of confirmatory bias. *Quarterly Journal of Economics*, 114(2):37–82., 1999.

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

Amnon Rapoport, James A. Sundali, and Darryl A. Seale. Ultimatums in two-person bargaining with one-sided uncertainty: Demand games. *Journal of Economic Behavior and Organization*, 30(2):173–196, 1996.

Tobias Regner and Javier A. Barria. Do consumers pay voluntarily? the case of online music. *Journal of Economic Behavior and Organization*, 71(2):395–406, 2009.

Jr. Roland G. Fryer, Philipp Harms, and Matthew O. Jackson. Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *working paper*, 2016.

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

David Sally. Two economic applications of sympathy. *The Journal of Law, Economics, and Organization*, 18(2):455–487, 2002.

Rene Saran. Bilateral trading with naive traders. *Games and Economic Behavior*, 72(2):544–557, 2011.

Rene Saran. How naiveté improves efficiency in trading with preplay communication. *Economics Letters*, 117(1):311–314, 2012.

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

Richard Selten. Reexamination of the perfectness concept for equilibrium points in extensive games. *Int. Journal of Game Theory*, 4(1):25–55, 1975.

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

Kathleen Valley, Leigh Thompson, Robert Gibbons, and Max H. Bazerman. How communication improves efficiency in bargaining games. *Games and Economic Behavior*, 38(1):127–155, 2002.

Ferdinand A. Von Siemens. Intention-based reciprocity and the hidden costs of control. *Journal of Economic Behavior and Organization*, 92:55–65, 2013.

Amy Wallace. How panicked parents skipping shots endanger us all. `https://www.wired.com/2009/10/ff_waronscience`, 2009. Accessed: 2017-10-20.

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.

Neil D. Weinstein. Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, 39(5):806–820, 1980.

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

Daniel John Zizzo and Andrew J. Oswald. Are people willing to pay to reduce others' incomes? *Annales d'Économie et de Statistique*, (63/64):39–65, 2001.