Selective Beliefs over Preferences[*]

Luke A. Stewart

University of California, Berkeley

April 29, 2009

Abstract

People tend to believe what they want to believe. In this paper, I examine how people interpret others' behavior to form beliefs about others' preferences. I develop a "selective beliefs" model of social preferences in which people have intrinsic preferences over outcomes and also hold prior beliefs about the preferences of others. I propose that people attempt to selectively perceive and interpret others' behavior such that their posterior beliefs about others' preferences minimally interfere with their own self-interest. From a general model, I derive two specific models designed for the regression analysis of a semi-transparent simplified ultimatum game experiment, where I assume that the second-mover forms posterior beliefs about first-mover's preferences over the second-mover's two possible outcome choices. The inequality aversion model is designed to measure the second-mover's behavioral response to an available signal indicating that the first-mover has some degree of aversion to inequality. The income concern model measures the second-mover's behavioral response to an available signal indicating that the first-mover is either concerned or unconcerned about income.

[*] I am ever grateful to my thesis adviser, Professor Botond Kőszegi, for his invaluable input and for coolly fielding an extraordinary number of questions on everything from psychology and economics to word-processing. I also thank Professor Roger Craine for his generous advice on the research process.

1 Introduction

In the absence of meaningful communication, social interaction is governed by personal preferences, social convention, and the assumptions people make about the preferences of others based on incomplete information. In this paper, I study the latter phenomenon. I develop a general model for examining how people interpret others' behavior to form beliefs about others' preferences, and then apply the model to a simplified ultimatum game where I assume that the second-mover forms beliefs about the first-mover's degree of aversion to inequality or concern for income by observing the first-mover's prior behavior. I extend the model to incorporate an important qualification: people selectively perceive and interpret others' behavior such that their posterior beliefs about others' preferences minimally interfere with their own self-interest. In a sense, people trick themselves through their beliefs in order to "morally wriggle" around their concern for others' welfare.

The selective beliefs model assumes that people have intrinsic preferences over outcomes and are also concerned about the preferences of others. Intrinsic preferences are of two sorts: personal preferences and social convention. Personal preferences are independent of the outside world and account for self-interest and our internal moral compass. Social convention is conditioned by the outside world—like an external moral compass.[1] The selective beliefs model assumes that personal preferences and social convention are inextricable—they both come under the umbrella of intrinsic preferences. On the other hand, peoples' concern for others' preferences is dictated by their *beliefs* about others' preferences. These beliefs are subject to a *weight of concern*, which people adjust to different situations depending on how much they care about the utility of others. The weight of concern captures traditional reciprocity: people increase or decrease their concern for others, conditional on signals in others' behavior that indicate others' kindness or meanness.[2] Importantly, others' behavior may also signal their preferences. Signals that others' prefer a particular property

[1] Li (2008) develops a model of social convention that can explain results from several previous experiments, including Charness and Rabin (2002).

[2] The weight of concern also captures concern that is situation-dependent. For example, in competitive markets, peoples' weight of concern for others' preferences may be null.

of outcomes cause a player to update her beliefs about others' preferences by increasing the weight she assigns to that specific property of her beliefs. Likewise, signals that others are averse to a particular property cause a player to decrease the weight she assigns to that property of her beliefs. These specific weights differ from the player's overall weight of concern: they assign a degree of importance to properties of outcomes as individual components of the player's beliefs about others' preferences; the weight of concern applies equally to all components of the player's beliefs.

In the total absence of information about the preferences of others, people have *prior beliefs* about others' preferences. This extreme no-information case rests upon the assumption that the only information people have about others is that they are fellow human beings, and thus their prior beliefs about the preferences of others likely comprise of their beliefs about the preferences of the "average" person. Admittedly, this case is a bit hypothetical and may not ever apply exactly to the real world.[3] Nevertheless, the no-information case provides a baseline from which we can examine the role of intrinsic preferences and prior beliefs in decision-making. For example, take the classic "battle of the sexes" game—you play the woman, I'll play the man. To set up our personal preferences, suppose I prefer football and you prefer opera, but we both prefer to go to any event together. (For simplicity, assume that the only properties of outcomes we consider are each other's direct utility payoffs from attending each event.) Unfortunately, the football game and the opera both begin at six o'clock tonight, so we can only attend one event together. To illustrate our prior beliefs in an approximate no-information case, suppose this is a blind first date, and thus the only information we have about each other is our genders. I know that the average woman prefers opera, and you know that the average man prefers football, but we are not sure who is willing to acquiesce to the preferences of the other such that we attend one event together and achieve a pareto-dominant equilibrium.[4] In this case, successful coordination may arise through

[3] The corresponding laboratory experiment is a well-implemented anonymous dictator game, although even here, subjects may garner information about the possible preferences of others by judging unavoidable clues such as the subject recruitment process.

[4] The reader may wonder why we do not simply communicate our preferences. Fair point, but the absence of communication is not quite as far-fetched as it may at first seem: communication would interfere with the ability to pass judgment on the other's intrinsic character. Furthermore, this *is* just a hypothetical example.

social convention: perhaps it is customary for the man to submit to the woman's outing preference on the first date. Hence, we attend the opera.

At the other extreme is the complete information case, which is best exemplified by meaningful communication. Although it is impossible to communicate preferences without some ambiguity, meaningful communication is still effective at promoting cooperation because it provides decision-makers with near-complete information about others' preferences.[5] Complete information ensures that peoples' *posterior beliefs*—their beliefs given a signal indicating others' preferences—accurately reflect others' *actual preferences* or "type," thus others are likely to be more satisfied with the outcome. In the battle of the sexes example, suppose that our first date went well, and now we are husband and wife. Additionally, our marriage is quite healthy, so we are good communicators. I tell you that I want to attend the football game, so you increase the individual weight you assign to your belief about my preference for the football game. Then, you remind me that I forgot our anniversary two months ago, and that you already told me about the opera three weeks ago and I agreed to go with you. My guilt about missing our anniversary causes me to increase my weight of concern for your preferences. Furthermore, I increase the individual weight I assign to my belief about your preference for opera, since you had been planning our attendance for three weeks. In equilibrium, opera it is.

In between these two information extremes lies the case of incomplete information. In many situations where meaningful communication is absent, people do not have access to complete information about others' preferences, yet they can update their beliefs using perceived signals contained in myriad alternate sources of information. These sources include others' behavior, appearance, peer groups, and even seemingly meaningless blabber. In this paper, I focus on signals contained solely in others' behavior. There is an important caveat here: unlike the case of complete information, the signals that people perceive as corresponding to others' actual preferences under incomplete information may not correspond to their actual preferences at all. Depending on the situation, dif-

[5] For example, Cooper et al. (1992) find that "cheap talk" among players before a coordination game significantly increases the play of the pareto-dominant equilibrium.

ferent preferences may manifest in the same observed behaviors, and thus the backward inference of others' preferences from behavior is subject to *inference error*. To highlight the fact that signals may not accurately reflect the sender's type due to inference error, I will use the term *perceived signals* (as opposed to "received" signals) to denote information gathered in the context of incomplete information. Other sorts of errors can further confuse the correlation between peoples' beliefs and others' actual preferences. Pinkley, Griffith and Northcraft (1995) describe two kinds of *information processing errors*: perception errors occur when people ignore available signals; interpretation errors occur when, given a perceived signal, people distort its meaning.[6] The selective beliefs model assumes that information processing errors occur in the context of peoples' intrinsic preferences, or self-interest. *Selective perception* occurs when people only perceive signals that would cause their posterior beliefs to complement their self-interest, and ignore available signals that would cause their posterior beliefs to conflict with their self-interest. *Selective interpretation* occurs when people overweight the importance of signals that cause their posterior beliefs to complement their self-interest, and underweight the importance of signals that cause their posterior beliefs to conflict with their self-interest. Selective perception and selective interpretation may occur simultaneously. They differ primarily in that, with selective perception, people either ignore an available signal or they perceive it in full, whereas selective interpretation is a matter of degree.

The primary implication of selective perception and selective interpretation is the same: if, under incomplete information, peoples' concern for others' preferences conflicts with their self-interest, they may be able to circumvent this concern through their beliefs about others' preferences. Dana, Weber and Kuang (2007) label a similar phenomenon "moral wriggling." Moral wriggling with respect to beliefs may explain the lack of evidence for positive reciprocity in empirical studies. To illustrate this, suppose the original battle of the sexes example was instead a sequential game:

[6] I assume that both inference errors and information processing errors occur without the conscious knowledge of the person making the error. In other words, people always presume that their conscious beliefs about the preferences of others are as accurate as possible given their interpretation of perceived information. Perception and interpretation may be subject to conscious internal negotiation, but people believe that the end result is a good-faith approximation of the truth.

instead of a blind first date, suppose you had asked me out on our first date and kindly decided to take me to the football game; now it's time for our second date, and it is my turn to choose the location. Again, we can attend the football game or the opera, but we cannot attend both at the same time. For our second date, however, I do not need to resort to social convention: I perceive the fact that you acquiesced to my desire to attend the football game on our first date as a signal about your preferences, so I can take some action based upon my posterior beliefs about your preferences. Traditional reciprocity models would have me perceive your behavior as kind: you sacrificed to increase my utility last time, so I will sacrifice to increase your utility this time—hence, in equilibrium, we attend the opera. But the selective beliefs model allows me to perceive your behavior in a different way: since you acquiesced to my desire to attend the football game last time, obviously opera is not that important to you. I may increase my weight of concern due to positive reciprocity, but I also decrease the individual weight I assign to my belief about your preference for opera. Or, conversely, I may perceive that you really prefer football and increase the weight I assign to your preference for football. Either way, in selective-beliefs equilibrium, we attend the football game again (or at least I do—you might not show up). The selective beliefs model allows me to morally wriggle around my concern for your preferences such that I can guiltlessly pursue my own self-interest.

We should not overlook the other fundamental assumption of the selective beliefs model, however: that people do indeed respond to perceived signals indicating others' preferences—moral wriggling with respect to beliefs is merely an extension of this assumption. In many situations, there is no incentive for moral wriggling, and people are happy to oblige their beliefs about others' preferences. After a summary of related literature in Section 2, I describe the general selective beliefs model with slightly more rigor in Section 3. Then, I apply the model to the perspective of the second-mover in a simplified ultimatum game with respect to two specific properties of outcomes: payoff equality, and the payoff of the first-mover. In Section 4, I explain how to test both fundamental assumptions of the selective beliefs model using a simplified ultimatum game experiment akin to that of Charness and Rabin (2002), where the treatment game hides relevant information from which the second-mover could otherwise form posterior beliefs about the first-mover's preferences.

2 Related Literature

To explain other-regarding behavior, economists have developed models of *social preferences*, which generally assume people are self-interested but are also concerned about others. Theories of social preferences are divided into two broad classes: distributional preferences and reciprocal preferences. *Distributional preferences* are unconditional on the behavior and disposition of others, and depend on material outcomes only. For example, in a dictator game, a "dictator" decides how to split an endowment between herself and a "receiver." Studies show that a significant proportion of dictators sacrifice a portion of their payoff to increase the payoff of the receiver.[7] The dictator has no information about the receiver, yet chooses to depart from self-interest. Therefore, distributional models assume that the dictator is intrinsically concerned about the utility of the receiver.

Models of distributional preferences are further classified by their assumptions.[8] *Inequality aversion* models assume that players are averse to differences in payoffs. Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) are perhaps the archetypal models of this class. In Fehr and Schmidt (1999), players dislike *any* differences between their own payoff and the payoffs of other players. In Bolton and Ockenfels (2000), players compare their payoff to only the average payoff of other players, and thus the distribution of payoffs among other individual players is irrelevant. In Section 3.2, I develop a selective beliefs model of inequality aversion, which is derived from that of Fehr and Schmidt (1999). *Altruism* models such as Andreoni (1990) and Andreoni and Miller (2002) assume players' utility is an increasing function of others' payoffs.[9] In Section 3.3,

[7] For example, in two separate dictator games in Charness and Rabin (2002), 48% and 50% of proposers sacrifice 25 for an allocation of (375,750) verses an even split of (400,400). See also Andreoni and Miller (2002).

[8] For a good summary of the advantages and limitations of each model, see Sobel (2005) and Engelmann and Strobel (2007).

[9] Cox and Sadiraj (2007) present a model of "egocentric altruism," which differs in that players with lower payoffs always prefer to swap payoffs with players with higher payoffs.

I develop an income concern model, which is derived from Andreoni and Miller (2002). *Quasi-maximin* models assume that players like to make the lowest payoff as high as possible, yet are themselves willing to sacrifice for the sake of efficiency. Employing a dictator game and a simplified ultimatum game, Charness and Rabin (2002) find quasi-maximin preferences superior to inequality aversion models at explaining subject behavior. Various games in Engelmann and Strobel (2004) reveal similar evidence (although Engelmann and Strobel 2007 find that predictive power of quasi-maximin decreases in multiplayer dictator games). I do not develop a model where players hold beliefs about others' quasi-maximin preferences; instead, I assume that players' beliefs about others' adherence to quasi-maximin norms would affect their weight of concern for others' preferences.

Distributional models fail to fully explain behavior in more complex bargaining games such as the ultimatum game. Early studies have demonstrated that situational context significantly influences peoples' bargaining decisions. For example, Kahneman, Knetsch and Thaler (1986) find that customers consider it acceptable for a firm to raise prices when the firm is losing money, but consider raising prices (or, conversely, lowering wages) in response to demand shifts unfair. Akerlof and Yellen (1990) propose a model whereby employee effort depends upon the relationship between what the employee considers the “fair” wage and the actual wage offered by the employer. In the laboratory, Brandts and Solà (2001) and Falk, Fehr and Fischbacher (2003) examine the ultimatum game and find that the second-mover's decision to accept or reject an offered outcome depends on properties of the outcome that the first-mover did not offer. In models of social preferences, context-dependent effects manifest in the form of assumptions about player reciprocity.

In general, *reciprocal preferences* seek to explain behavior that is conditional on information or beliefs about others' behavior or character. *Intentions-based reciprocity* models assume that players' preferences over outcomes are motivated by the perceived intentions of other players. In turn, players' perceptions of others' intentions depend upon the context of the game. Outcomes may be more or less preferable depending on the player's strategy set as a whole. In a seminal article, Rabin (1993) proposed a model in which players hold beliefs about the strategies that others intend to pursue, and beliefs about the strategy that others believe they themselves will pursue.[10] Players then judge the kindness or unkindness of others' strategies by comparing the payoff implied by their beliefs to an "equitable payoff" (which is somewhat arbitrarily defined as the average of the highest and lowest possible payoffs). If a player perceives another's intended strategy as unkind, then "negative reciprocation" prescribes that the player will prefer to choose an unkind strategy herself. Likewise, if a player perceives another's strategy as kind, then "positive reciprocation" prescribes that the player will prefer to behave kindly as well.

The *signal-based reciprocity* model proposed by Levine (1998) is also context-dependent, though in less direct manner. Each player has a privately-known altruism coefficient and is uncertain about the altruism coefficient of others. Players may signal the value of their coefficient, or type, through their actions. Players reward others whom they perceive to be altruistic and punish those whom they perceive to be spiteful. Players' preferences over outcomes do not depend on the context of their strategy sets per se. Rather, preferences for differing discrete strategies in dissimilar strategy sets may alter the signal sent to other players, whom in turn alter their preferences to reward altruistic players and punish spiteful ones. At first glance, the selective beliefs model seems more akin to a signal-based reciprocity model, in that the weight of concern is affected by signals of others' altruism or spitefulness; however, these signals are in fact *perceived* signals, subject to inference error, and may not accurately reflects others' types. Therefore, the selective beliefs model is really a hybrid intentions-based and signal-based model: kind or mean intentions matter, but only as perceived signals about others' altruism or spitefulness.

In empirical research, Charness and Rabin (2002), Engelmann and Strobel (2004), and Falk, Fehr and Fischbacher (2008) find that models that incorporate both distributional and reciprocal preferences are significantly better predictors of subject behavior than either distributional or reciprocal models alone. Nevertheless, the evidence for reciprocal behavior is not entirely straightforward. Bolton, Brandts and Ockenfels (1998), Charness and Rabin (2002), Offerman (2002), and

[10] In equilibrium, all beliefs are assumed to be correct. Dufwenberg and Kirchsteiger (2004) modify the Rabin (1993) model to include sequential play and learning.

Charness (2004) find that positive reciprocation is either insignificant or much weaker than negative reciprocation. The selective beliefs model may explain this finding: although it allows players to increase their concern for others' welfare in response to kind behavior, it also allows players to circumvent this concern through their beliefs when positive reciprocation would conflict with self-interest. On the other hand, Falk, Fehr and Fischbacher (2008) find that both positive and negative reciprocation is rather weak, and Charness and Rabin (2002) and Charness and Rabin (2005) find that negative reciprocation takes the form of *concern withdrawal*, where rather than choosing pareto-damaging outcomes as punishment, players merely disregard the payoffs of others and pursue their own self-interest in response to mean treatment. In the selective beliefs model, this finding implies that players would decrease their weight of concern to null in response to mean treatment, thus withdrawing their concern for the preferences of players they perceive as spiteful, but that the weight of concern would not assume negative values.

Researchers have also found that reciprocal preferences are more pronounced when people can attribute some volition to the actions of others. This phenomenon is known as *causal attribution.* Blount (1995) uses an ultimatum game to show that responders accept lower payoffs when the offer is generated randomly by nature as opposed to a human proposer. Charness (2004) uses a gift-exchange game in which a wage is chosen either voluntarily by another player—an "employer"—or involuntarily by nature. "Employees" are told whether their wage was assigned voluntarily or involuntarily, and then asked to return a portion of their wage as "effort." Charness finds that when employees believe that a high wage was voluntary or a low wage was involuntary, they will contribute more effort to the employer. With these findings in mind, I control for casual attribution in my experiment by employing an ultimatum game in both the baseline and the treatment. This ensures that the degree to which the second-mover holds the first-mover accountable for her decision is consistent.

Charness and Rabin (2005) explore the role of expressed preferences in dictator and simplified ultimatum games. Receivers in dictator games and proposers in ultimatum games were allowed to express a preference for the deciding player's behavior. Requests for favorable treatment ("help me") or unfavorable treatment ("don't help me") had a large and significant effect on the deciding player's behavior, except when, in the ultimatum games, the proposers had violated the norms of quasi-maximin preferences—the responders exhibited concern withdrawal. Charness and Dufwenberg (2006) construct a modified ultimatum game in which the second-mover can promise to choose a particular strategy if the first-mover does not exit the game. They find that an expressed preference from the second-mover significantly enhances cooperation among players.[11] They develop the concept of *guilt aversion*, whereby the first-mover is averse to the guilt she experiences if she believes that she has let the second-mover down by exiting the game. The stronger a person believes that another person prefers a particular outcome, the greater the potential guilt from ignoring the preference, and thus the higher the propensity to cooperate.

Guilt is a potentially important concept in the selective beliefs model. For one, it may act as the enforcement mechanism behind peoples' concern for others. Furthermore, it may explain the apparent constraint on selective interpretation. That is, the fact that the degree to which people underweight or overweight perceived signals is limited implies that there is some moral cost or constraint associated with selective interpretation. Otherwise, people would always minimize the importance of signals that conflicted with their self-interest to null, and maximize signals that complemented their self-interest such that their importance outweighed any other consideration. I speculate that, if people are too liberal with their interpretation, they realize that they are lying to themselves and thus experience guilt—their presumption that their beliefs are a good-faith approximation of the truth is broken. With selective perception, guilt is probably irrelevant: people cannot feel guilty about ignoring an available signal if they have no conscious knowledge about what it is they are ignoring.

Other than employing selective perception, players can potentially avoid guilt by avoiding information about the outcomes for other players. In other words, if you don't know that you've

[11] The expressed preferences from the second-mover in Charness and Dufwenberg (2006) take the form of detailed and occasionally hilarious messages.

screwed another person, you can't feel guilty about it. Dana, Weber and Kuang (2007) employ a dictator game to study information-avoidance and its effects on decision-making. As a baseline, they use the standard dictator game. As a treatment, they reduce the transparency of the dictator game such that the receiver's two possible payoffs are hidden from the dictator. The dictator can, however, choose to reveal the receiver's payoffs before making a decision. They find that slightly less than 50% of dictators choose not to reveal information, although those that choose not to reveal the receiver's payoffs always pursue their own self-interest. Overall, the proportion of dictators that pursue self-interest at the expense of the receiver in the hidden-information treatment game is greater than the proportion that do so in the transparent baseline game.[12] Dana, Weber and Kuang propose that information-avoidance allows people to exploit "moral wiggle room," whereby people are able to circumvent their concern for others' preferences by willfully avoiding information about the payoffs of others.[13] In this case, guilt—the concern enforcement mechanism—is effectively neutralized.

3 Models of Selective Beliefs

In this section, I present three models of selective beliefs. The first is a general model of selective beliefs for multiple properties of outcomes, which, with some tinkering, can be applied to a range of sequential two-player games. The second and third models are derived from the general model and are designed specifically for the regression analysis of selective beliefs in my simplified ultimatum game experiment. The inequality aversion model is designed to measure the second-mover's behavioral response to an available signal indicating that the first-mover has some degree of aversion to inequality. The income concern model measures the second-mover's behavioral response

[12] Similarly, Munyan (2005) finds that players who make equitable decisions in transparent games tend to choose to reveal information (although less so if they know there is a high probability that they will not hurt the other person by following self-interest), and that those who hurt another person by acting self-interestedly in transparent games tend not to reveal information. See also Larson (2005).

[13] Rabin (1995) proposes several models where morals act as constraints on behavior, which people seek to circumvent.

to an available signal indicating that the first-mover is either concerned or unconcerned about income. Both the inequality aversion model and the income concern model are presented from the perspective of the second-mover, since the first-mover cannot form posterior beliefs about the second-mover's preferences before she takes an action. The models could potentially be applied to the first-mover's behavior, though they would likely function as considerably less robust versions of Fehr and Schmidt (1999) and Andreoni and Miller (2002), respectively. All three models are linear with respect to players' intrinsic utility functions and players' beliefs about others' utility functions. Readers seeking to skip the details and derivation of the specific models can find the complete inequality aversion model in Section 3.2.2 and the complete income concern model in Section 3.3.2.

3.1 The General Model

I begin with a simple and general model of other-regarding preferences for any two-player game Γ. I specify player i's *concerned utility* for any outcome (π_i, π_j) as $V_i(U_i, \hat{U}_j)$. Let $Q(\pi_i, \pi_j) = (Q_1(\pi_i, \pi_j), Q_2(\pi_i, \pi_j), ..., Q_R(\pi_i, \pi_j))$ be various outcome properties other than player i's own payoff. Player i's concerned utility has two components: $U_i(\pi_i, Q(\pi_i, \pi_j))$ represents player i's *intrinsic utility* derived from personal preferences and social convention; $\hat{U}_j(Q(\pi_j, \pi_i))$ represents player i's *beliefs about player j's utility function*, where α_i is player i's *weight of concern* for j's preferences.

$$V_i(U_i, \hat{U}_j) = U_i(\pi_i, Q(\pi_i, \pi_j)) + \alpha_i \hat{U}_j(Q(\pi_j, \pi_i)) \qquad (1)$$

I exclude player i's own payoff π_i from her beliefs about j's utility function to simplify the model. Although i may have beliefs about the altruism of j, she would not withdrawal concern for the altruism of j. Instead, i's beliefs about the altruism (or spitefulness) of j will affect her weight of concern for j.

To introduce the role of perceived signals and beliefs in the model, I will use the simplified ultimatum game of my experiment as an example. In this game, first-movers either choose a certain outcome, or they forgo this certain outcome and allow the second-mover to choose between two different outcomes. I assume that the second-mover forms posterior beliefs about first-mover's preferences over her two possible outcomes, based on the properties of the "forgone" outcome relative to the properties of the second-movers' two possible outcomes. (Her posterior beliefs are subject to *inference error*, however, in that her perception of the first-movers' behavior may not accurately reflect the first-movers' actual preferences or type.) Information on the properties of the forgone outcome takes the form of *negative perceived signals*, in that the second-mover perceives that the first-mover derives negative utility from those forgone properties. Information on the properties of the second-mover's two possible outcomes takes the form of *positive perceived signals*, in that the second-mover perceives that the first-mover derives positive utility from the properties associated with one or both possible outcomes. Then, the second-mover's posterior beliefs about the first-mover's preferences is a function of her mental reconciliation of all component positive and negative signals, subject to selective perception and selective interpretation.[14] For simplicity, I use the term *perceived signal* as an amalgamation of all relevant *component signals*, which is then subject to selective processing.

To look at the role of perceived signals more closely, I first expand equation (1), such that

$$V_i(\pi_i, Q(\pi_i, \pi_j)) = \pi_i + \beta_i \cdot Q(\pi_i, \pi_j) + \alpha_i[\,\hat{\beta}_j \cdot Q(\pi_j, \pi_i)] \tag{2}$$

where $\beta_i = (\beta_{i1}, \beta_{i2}, ..., \beta_{iR})$ are player i's intrinsic-preference weights for outcome properties and $\hat{\beta}_j = (\hat{\beta}_{j1}, \hat{\beta}_{j2}, ..., \hat{\beta}_{jR})$ are player i's beliefs about j's weights for outcome properties. Player i may perceive signals in player j's behavior, which indicate j's actual weights for various outcome properties. If i perceives a signal that the actual weight j assigns to a particular outcome property differs from that of i's prior beliefs, then i will adjust the weight she assigns to that particular

[14] Positive and negative signals are not necessarily "two sides of the same coin." For example, in my simplified ultimatum game, if the second-mover could not see the forgone outcome, then she would only know that the first-mover liked her two possible outcomes—in this case, there is only a positive perceived signal. Or, if the second-mover knows that the first-mover cannot see her two possible outcomes, then there is only a negative perceived signal.

property in her beliefs accordingly. Also, player i may perceive a signal indicating that player j is either kind or mean, and will thus adjust her weight of concern for j's preferences. Incorporating perceived signals, player i's utility is:

$$V_i(\pi_i, Q(\pi_i, \pi_j), s_\alpha, s_{\hat{\beta}}) = \pi_i + \beta_i \cdot Q(\pi_i, \pi_j) + (\alpha_i | s_\alpha)[(\hat{\beta}_j | s_{\hat{\beta}}) \cdot Q(\pi_j, \pi_i)] \tag{3}$$

where $s_{\hat{\beta}} = (s_{\hat{\beta}1}, s_{\hat{\beta}2}, ..., s_{\hat{\beta}R})$ are perceived signals indicating j's actual weights for various outcome properties, and s_α is a perceived signal that j is either kind or mean. Player i's intrinsic-preference weights are not conditional on perceived signals, because i's personal and social convention preferences do not depend on her posterior beliefs about j. If $s_{\hat{\beta}r} = 0$, then no signal was perceived, and the weight player i assigns to the r^{th} outcome property in her *posterior beliefs* will equal that of her *prior beliefs*. Likewise, if $s_\alpha = 0$, then player i will not know if player j is kind or mean, thus her weight of concern for j will remain unchanged. The exact relationship between a perceived signal and the weight i assigns to an outcome property in her posterior beliefs depends upon the nature of the outcome property in question and i's interpretation of the perceived signal. I will not specify this relationship in the general model; however, I will later specify this relationship for the property of outcome equality in my inequality aversion model, and for j's payoff in my income concern model.

In the selective beliefs model, player i's interpretation of a perceived signal partly depends upon whether the perceived signal would cause her posterior beliefs about j's preferences to complement her own intrinsic preferences or to conflict with her own intrinsic preferences, relative to her prior beliefs. Player i may selectively perceive or selectively interpret signals in order to "morally wriggle" around her concern for j's preferences and instead pursue her own intrinsic preferences. *Selective perception* occurs when player i either totally ignores a signal or perceives it in full. *Selective interpretation* occurs when player i either underweights or overweights the magnitude of a perceived signal. Selective interpretation is not without bounds: if player i overweights or underweights a perceived signal "too much," then I assume she experiences guilt, to which she is averse. (I do not define these bounds, however. I will leave this up to the estimation of the specific models.)

To explain selective perception and selective interpretation, I assume that player i's weight of concern for j's preferences is fixed, thus $\alpha_i = 1$, and rewrite equation (3) in terms of i's intrinsic utility function and i's beliefs about j's utility function:

$$V_i(U_i, \hat{U}_j, s_{\hat{\beta}}) = U_i(\pi_i, Q(\pi_i, \pi_j)) + \hat{U}_j(Q(\pi_j, \pi_i), s_{\hat{\beta}}) \tag{4}$$

Suppose player i intrinsically prefers an outcome X over an outcome Y, such that $U_i(X) > U_i(Y)$. Player i's concerned preferences over outcomes given her prior beliefs about j's preferences may be ordered similarly, such that $V_i(U_i(X), \hat{U}_j(X), s_{\hat{\beta}} = 0) > V_i(U_i(Y), \hat{U}_j(Y), s_{\hat{\beta}} = 0)$, or the order may be reversed, such that $V_i(U_i(X), \hat{U}_j(X), s_{\hat{\beta}} = 0) < V_i(U_i(Y), \hat{U}_j(Y), s_{\hat{\beta}} = 0)$.[15] Take the first case, where player i's prior concerned preferences are ordered similarly to her intrinsic preferences. Suppose now that player i perceives some signal indicating j's actual preference for a particular outcome property, thus $s_{\hat{\beta}} \neq 0$. If the perceived signal would cause player i's posterior concerned preferences to maintain a similar order to that of her prior concerned preferences, such that $V_i(U_i(X), \hat{U}_j(X), s_{\hat{\beta}} \neq 0) > V_i(U_i(Y), \hat{U}_j(Y), s_{\hat{\beta}} \neq 0)$, then i is indifferent about her perception or interpretation of the signal—she is able to pursue her intrinsic preferences regardless of her beliefs. On the other hand, if the perceived signal would cause player i's posterior concerned preferences to have the reverse order of her prior concerned preferences, such that $V_i(U_i(X), \hat{U}_j(X), s_{\hat{\beta}} \neq 0) < V_i(U_i(Y), \hat{U}_j(Y), s_{\hat{\beta}} \neq 0)$, then i will have an incentive to ignore the signal or underweight its meaning and pursue her prior concerned preferences instead.

Now, take the second case of prior concerned preferences, where the preference order is the reverse of that of i's intrinsic preferences. Considering *only* selective perception, if the perceived signal would cause player i's posterior beliefs to have the reverse order of her prior concerned preferences, and thus a similar order to her intrinsic preferences, such that $V_i(U_i(X), \hat{U}_j(X), s_{\hat{\beta}} \neq 0) > V_i(U_i(Y), \hat{U}_j(Y), s_{\hat{\beta}} \neq 0)$, then i will have an incentive to perceive the signal in full. However,

[15] For simplicity, I am ignoring the indifference case, where $V_i(U_i(X), \hat{U}_j(X), s_{\hat{\beta}} = 0) = V_i(U_i(Y), \hat{U}_j(Y), s_{\hat{\beta}} = 0)$

there is a difference with selective interpretation here. In the previous case, player i was happy about her prior concerned preferences because they complemented her intrinsic preferences, so there was an incentive to ignore or underweight only if a signal would cause her posterior concerned preferences to interfere with her prior concerned preferences. In this case, player i is unhappy about her prior concerned preferences because they conflict with her intrinsic preferences, thus she is searching for an excuse to circumvent them. Depending on the relationship between the perceived signal and its respective property-weight in i's beliefs, there is a chance that a fully-perceived *and* overweighted signal would ensure that i's posterior concerned preferences have a similar order to her intrinsic preferences, whereas a signal that was fully-perceived but not overweighted would not. Therefore, if the *overweighted* perceived signal would cause player i's posterior beliefs to have the reverse order of her prior concerned preferences, and thus a similar order to her intrinsic preferences, such that $V_i(U_i(X), \hat{U}_j(X), s_{\hat{\beta}} \neq 0) > V_i(U_i(Y), \hat{U}_j(Y), s_{\hat{\beta}} \neq 0)$, then i will indeed overweight the perceived signal. On the other hand, if the perceived signal would cause player i's posterior concerned preferences to maintain a similar order to that of her prior concerned preferences despite any attempt at selective interpretation, such that $V_i(U_i(X), \hat{U}_j(X), s_{\hat{\beta}} \neq 0) < V_i(U_i(Y), \hat{U}_j(Y), s_{\hat{\beta}} \neq 0)$, then i is indifferent about her perception or interpretation of the signal—there is no way for i to circumvent her concern for j through her beliefs—she is just going to have to play nice this time.

I will not attempt to model this thought process. Instead, I will simply define the magnitude of the incentive to morally wriggle for any given outcome as the difference between the intrinsic utility derived from that outcome and the intrinsic utility derived from an intrinsically-preferred outcome. Let the *intrinsically-preferred outcome* be the possible outcome for which player i's intrinsic utility is highest. Let $\bar{U}_i$ be the intrinsic utility that i derives from the intrinsically-preferred outcome. Hence, the *incentive to morally wriggle* is:

$$m(U_i, \bar{U}_i) = \bar{U}_i - U_i(\pi_i, Q(\pi_i, \pi_j)) \tag{5}$$

The greater the magnitude of m, the stronger the incentive for i to attempt to morally wriggle with respect to her beliefs if, without moral wriggling, her beliefs about j's preferences would conflict with her intrinsic preferences.

Returning to the model, the weights that player i assigns to outcome properties in her beliefs are now conditional on the both the perceived signal and the incentive to morally wriggle. I relax the assumption that i's weight of concern for j is constant and rewrite equation (4) to include moral wriggling:

$$V_i(U_i, \hat{U}_j, \bar{U}_i, s_\alpha, s_{\hat{\beta}}) = U_i(\pi_i, Q(\pi_i, \pi_j)) + (\alpha_i | s_\alpha)\, \hat{U}_j(Q(\pi_j, \pi_i), s_{\hat{\beta}}, m(U_i, \bar{U}_i)) \tag{6}$$

To look under the hood (although not very closely), I expand equation (6):

$$V_i(\pi_i, Q(\pi_i, \pi_j), \bar{U}_i, s_\alpha, s_{\hat{\beta}}) = \pi_i + \beta_i \cdot Q(\pi_i, \pi_j) + (\alpha_i | s_\alpha)[(\hat{\beta}_j |\, s_{\hat{\beta}},\ m(\pi_i, Q(\pi_i, \pi_j), \bar{U}_i)) \cdot Q(\pi_j, \pi_i)] \tag{7}$$

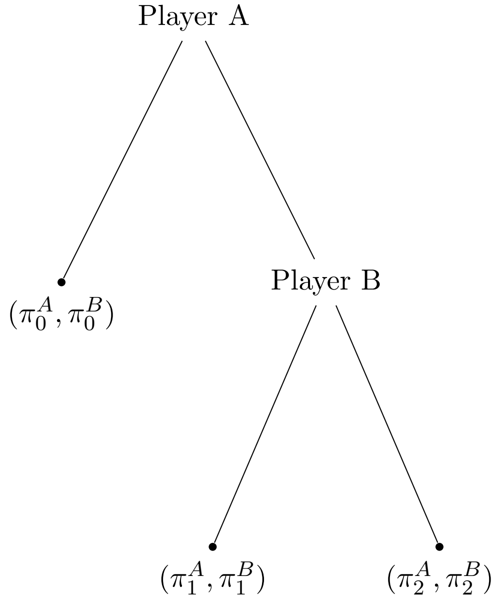
With a some adjustments, the general model of selective beliefs is now ready for application to specific games and properties of outcomes. I begin with the property of equality in a simplified ultimatum game.

3.2 The Inequality Aversion Model

In the inequality aversion model, players are intrinsically self-interested with respect to their own payoffs, but are also intrinsically averse to inequality in payoffs. Additionally, players hold prior beliefs about other players' aversion to inequality and may update these beliefs conditional on perceived signals. The stronger a perceived signal that another player is inequality averse, the higher the propensity of a player to choose a more equal outcome. In other words, players are concerned about others' aversion to inequality, but are not concerned about others' preferences for any other property of outcomes. Nevertheless, players may selectively interpret and selectively perceive signals that do not coincide with their self interest, thus a player's propensity to choose a more equal outcome due to a perceived signal will be lower when there is an incentive to morally wriggle.

3.2.1 Derivation

I construct the inequality aversion model for the specific purpose of estimating the posterior-belief weights of the second-mover in a simplified ultimatum game experiment of the form:



where player A is the first-mover and player B is the second-mover. First, player A decides either to "exit" the game such that she earns a payoff of π_0^A and player B earns a payoff of π_0^B, or to "enter" the game and pass the decision on to player B. If player A decides to enter, then player B decides between a payoff of π_1^B for herself and π_1^A for player A, or a payoff of π_2^B for herself and π_2^A for player A.

In a game where all payoffs are public both players, player B may perceive a signal indicating player A's degree of aversion to inequality. If player A decides to enter, then player B compares the distribution of payoffs in the forgone outcome (π_0^A, π_0^B) to the distribution of payoffs in the two possible outcomes, (π_1^A, π_1^B) and (π_2^A, π_2^B). If the distribution of the forgone outcome is more unequal than one or both of the possible outcomes, then player B perceives that player A is averse to inequality. In other words, player B perceives that player A rejected the forgone outcome because it was less equal than one or both of the other possible outcomes, and thus player B perceives that player A wants to "play fair." The strength of the perceived signal depends upon the measure of inequality of the forgone outcome (the component negative perceived signal) relative to the measure of inequality of each possible outcome (the component positive perceived signal), and on whether one or both possible outcomes are less unequal than the forgone outcome. I define the *inequality measure* of any outcome as the magnitude of the difference in payoffs:

$$f(\pi^A, \pi^B) = |\pi^A - \pi^B| \tag{8}$$

For simplicity, this measure ignores the exact distribution of payoffs among players—that is, whether the payoff inequality is in A or B's favor—although, admittedly, the exact distribution may influence subject behavior somewhat in experiments.

In a game where all payoffs are *not* public to both players, a signal indicating player A's degree of inequality aversion may be weak or not available to player B at all. For example, if one or both of the payoffs in the forgone outcome (π_0^A, π_0^B) are hidden from player B, then player B will not be able to determine its measure of inequality. Nevertheless, player B may still be able to infer A's preference for inequality from the positive perceived signal contained in her two possible outcomes. If player B knows that player A could see her two possible outcomes, and if one or both of those two possible outcomes seem rather equal to player B (depending on her personal reference point for equality), then she may perceive a signal—albeit a weak one—that A prefers an equal payoff. Hence, to ensure that player B cannot possibly update her beliefs about A's degree of aversion to inequality, at least one of the payoffs in the forgone outcome must be hidden from player B, and at least one of the payoffs in each of B's possible outcomes must be hidden from player A. I employ this method of hiding relevant payoffs in my "opaque" treatment game.

For any simplified ultimatum game Γ, I define the *perceived signal indicating player A's degree of aversion to inequality* as:

$$s_f(\Gamma) = T * [max(f(\pi_0^A, \pi_0^B) - f(\pi_1^A, \pi_1^B),\ 0) + max(f(\pi_0^A, \pi_0^B) - f(\pi_2^A, \pi_2^B),\ 0)] \tag{9}$$

where, by assumption, $T = 1$ in a transparent game where all payoffs are public to both players, and $T = 0$ in an opaque game where the relevant payoffs are hidden—the signal is not available in opaque games. If player A enters the game, then the strength of the signal indicating player A's degree of inequality aversion will determine the weight player B assigns to A's inequality aversion in her posterior beliefs: the stronger the perceived signal, the larger the weight. If A's forgone outcome is more unequal than *only one* of B's possible outcomes, then the magnitude of the signal will equal the difference of the inequality measures. If A's forgone outcome is more unequal than *both* of B's possible outcomes, then the magnitude of the signal will equal the sum of the both differences. If A's forgone outcome is *less unequal* than both of B's possible outcomes, then there is no perceived signal.

In my experimental games, player A always substantially increases the average of player B's possible payoffs by entering, thus player B always perceives player A as kind. Therefore, in my specific models, I assume that player B's weight of concern for player A's preferences is constant.[16] Excluding moral wriggling for now, I derive a simple inequality aversion model from the general model of selective beliefs. I rearrange equation (3) from the general model for a single property of outcomes, where player i's weight of concern $\alpha_i = 1$:[17]

$$V_i(\pi_i, Q(\pi_i, \pi_j), s_{\hat{\beta}}) = \pi_i + [\beta_i + (\hat{\beta}_j | s_{\hat{\beta}})]Q(\pi_i, \pi_j) \tag{10}$$

[16] Although player B's weight of concern may increase depending on how much player A increases B's payoffs by entering, this effect is consistent between the baseline and the treatment, and thus does not present a confound.

[17] Note that the inequality aversion model can also be derived from equation (7) of the general model, including the incentive to morally wriggle. I only exclude moral wriggling for simplicity. In Section 3.3, I derive the income concern model from equation (7).

Next, I rewrite equation (10) for the inequality aversion model. Since I am interested in the beliefs of the second-mover, player i becomes player B. The property of outcomes $Q(\pi_i, \pi_j)$ becomes the inequality measure $f(\pi_k^A, \pi_k^B)$ for player B's possible outcomes, where $k \in \{1, 2\}$. Furthermore, now the perceived signal is endogenous:

$$V_f^B(\pi_k^A, \pi_k^B) = \pi_k^B - [\lambda_B + (\hat{\lambda}_A | s_f(\Gamma))] f(\pi_k^A, \pi_k^B) \tag{11}$$

In this simple model of inequality aversion, λ_B is player B's intrinsic-preference weight of inequality aversion, and $\hat{\lambda}_A$ is the weight player B assigns to player A's inequality aversion in her beliefs. The (-1) is an artifact of construction: I assume players prefer *not* to have unequal outcomes—that is, they are inequality averse. I do not expect players to ever be "inequality seeking." Players may prefer an unequal outcome, but this would be due to self-interest with respect to their own payoff, which the model already captures. Inequality is not preferred for its own sake; equality is.

For the purpose of estimation, I make an important adjustment to the model of equation (11). I am interested only in player B's *posterior* beliefs given a perceived signal indicating player B's aversion to inequality—that is, the change in B's beliefs in response to a perceived signal. Therefore, I must separate player B's posterior beliefs from her prior beliefs. Since I assume that player B's weight of concern for player A's preferences is fixed, this is easily accomplished by incorporating both player B's intrinsic preferences and prior beliefs in the first parameter λ_B, instead of the second parameter.[18] Another advantage to including prior beliefs in the first parameter is that we can then define the relationship between the perceived signal $s_f(\Gamma)$ and the weight on B's posterior beliefs $\hat{\lambda}_A$ as a simple multiplicative relationship. Incorporating these changes, the model of inequality

[18] This is only possible because we assume that $\alpha_B = 1$: in the general model, we assume that any change in B's concern for A's preferences will affect the weight she assigns both posterior *and* prior beliefs about A's preferences, but this cannot occur if we assume prior beliefs are incorporated in the first parameter. However, since we now assume that B's weight of concern is constant, incorporating her prior beliefs in the first parameter is consistent with the general model.

aversion from equation (11) becomes:

$$V_f^B(\pi_k^A, \pi_k^B) = \pi_k^B - [\tilde{\lambda}_B + \hat{\lambda}_A s_f(\Gamma)] f(\pi_k^A, \pi_k^B) \tag{12}$$

where the tilde over the first parameter $\tilde{\lambda}_B$ indicates that it now represents the weight of B's inequality aversion due to a mix of her intrinsic preferences and her prior beliefs about A's degree of inequality aversion.

I define player B's moral wriggling function in terms of her own payoff only. Although player B's intrinsic preferences include an aversion to inequality, a perceived signal that player A is inequality averse could not possibly conflict with an intrinsic aversion to inequality. I assume that player B's intrinsically-preferred outcome is the outcome with the highest payoff, so player B will have an incentive to morally wriggle with respect to her beliefs if the outcome under evaluation is not the intrinsically preferred outcome. Therefore, player B's *incentive to morally wriggle* is the indicator function:

$$m(\pi_k^B, \pi_{3-k}^B) = \begin{cases} 1 & \text{if } \pi_k^B < \pi_{3-k}^B \\ 0 & \text{if } \pi_k^B \geq \pi_{3-k}^B \end{cases} \tag{13}$$

where π_k^B is player B's payoff in the outcome under evaluation, and π_{3-k}^B is player B's payoff in the other possible outcome.[19]

3.2.2 Complete Model

For any simplified ultimatum game Γ, player B's possible outcomes are (π_k^A, π_k^B) where $k \in \{1, 2\}$, and $(\pi_{3-k}^A, \pi_{3-k}^B)$ is the possible outcome not under evaluation, and (π_0^A, π_0^B) is Player A's forgone outcome. Incorporating moral wriggling, the complete inequality aversion model is:

$$V_f^B(\pi_k^A, \pi_k^B) = \pi_k^B - [\tilde{\lambda}_B + \hat{\lambda}_A^m s_f(\Gamma) m(\pi_k^B, \pi_{3-k}^B) + \hat{\lambda}_A^n s_f(\Gamma)(1 - m(\pi_k^B, \pi_{3-k}^B))] f(\pi_k^A, \pi_k^B) \tag{14}$$

[19] Player B's incentive to morally wriggle is an indicator function for simplicity. In my experiment, I am not interested in player B's response to varying magnitudes of moral-wriggling incentives; rather, I am interested in player B's response to varying magnitudes of perceived signals.

where $f(\pi_k^A, \pi_k^B)$ is the inequality measure of the outcome under evaluation, $s_f(\Gamma)$ is a perceived signal indicating player A's degree of inequality aversion, and $m(\pi_k^B, \pi_{3-k}^B) = 1$ if, for the outcome under evaluation, player B has an incentive to morally wriggle with respect to her posterior beliefs about player A's aversion to inequality. $\tilde{\lambda}_B$ is the weight of player B's inequality aversion due to her intrinsic preferences and her prior beliefs about player A's degree of aversion to inequality. The two subsequent parameters are "adjusted" weights, indicating that their values are not directly comparable to that of $\tilde{\lambda}_B$, due to the presence of the additional signal variable. They are also "additive," in that they reflect the adjusted weight that player B's posterior beliefs *add* to the weight already captured in her prior beliefs, not her posterior beliefs in full. $\hat{\lambda}_A^m$ is the additive adjusted weight of player B's inequality aversion due to her posterior belief about player A's degree of inequality aversion where, for the outcome under evaluation, there is an incentive for player B to morally wriggle with respect to her posterior beliefs. $\hat{\lambda}_A^n$ is the additive adjusted weight of player B's inequality aversion due to her posterior belief about player A's degree of inequality aversion where, for the outcome under evaluation, there is no incentive for player B to morally wriggle.

3.2.3 Predictions

The selective beliefs hypothesis implies that $0 \leq \hat{\lambda}_A^m < \hat{\lambda}_A^n$. If player B perceives a signal that player A is inequality averse and has no incentive to morally wriggle around this perception, then her propensity to choose more equal outcomes will be greater than her intrinsic and prior propensity to choose more equal outcomes without a perceived signal. However, if player B does have an incentive to morally wriggle, then her propensity to choose more equal outcomes will be less than if she did not have an incentive to morally wriggle, and perhaps no greater than her intrinsic and prior propensity to choose more equal outcomes. Player B may or may not have an intrinsic propensity to choose more equal outcomes due to an intrinsic aversion to inequality or her prior beliefs about player A's aversion to inequality.

3.3 The Income Concern Model

In the income concern model, players are intrinsically self-interested with respect to their own payoffs, but are also intrinsically altruistic with respect to others players' payoffs. Additionally, players hold prior beliefs about other players' concern for income—that is, how much other players need or desire income—and may update these beliefs conditional on perceived signals. Players may perceive that others are unconcerned about income or that they have some degree of concern for income. Furthermore, players may selectively interpret and selectively perceive signals that others are concerned about income, but do not selectively interpret or selectively perceive signals that others are unconcerned about income. Therefore, a signal that another player is unconcerned about income will likely decrease a player's propensity for altruistic behavior, while a signal that another player is concerned about income will increase a player's propensity for altruistic behavior, so long as there is no incentive for the player to morally wriggle.

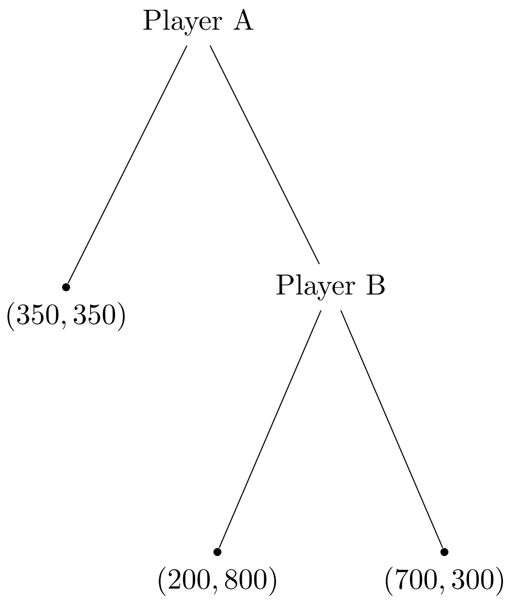
3.3.1 Derivation

Like the inequality aversion model, I construct the income concern model for the specific purpose of estimating the posterior-belief weights of the second-mover in a simplified ultimatum game (see Section 3.2.1 for a brief description). In a game where all payoffs are public to both players, player B may perceive a signal indicating player A's degree of concern for income. If player A's payoff in the forgone outcome (π_0^A, the component negative perceived signal) is less than an average of player A's payoffs in player B's two possible outcomes (π_1^A and π_2^A, the component positive perceived signals) weighted by B's respective payoffs, then player B perceives that player A is concerned about income. In other words, player B perceives that player A rejected the forgone outcome because she was unsatisfied with her payoff in the forgone outcome, and thus player B perceives that player A has a need or desire for more income than the forgone payoff would provide. On the other hand, if player A's payoff in the forgone outcome is greater than the weighted average of her payoffs in player B's two possible outcomes, then player B perceives that player A is unconcerned about income.

In a game where all payoffs are *not* public to both players, a signal indicating player A's concern for income may be weak or not available to player B at all. For example, if player A's payoff in the forgone outcome (π_0^A, π_0^B) is hidden from player B, then player B will have no tangible reference point (the negative perceived signal) with which to compare player A's payoffs in her two possible outcomes, (π_1^A, π_1^B) and (π_2^A, π_2^B). Nevertheless, player B may still be able to infer A's degree of concern for income using the positive perceived signal in her two possible outcomes. If player B knows that player A could see A's payoffs in B's two possible outcomes, and if the weighted average of A's payoffs seems rather high to player B (depending on her personal reference point for income), then she may perceive a weak signal that player A is concerned about income. Conversely, if those payoffs seem rather low, then she may perceive that player A is unconcerned about income. Hence, to ensure that player B cannot possible update her beliefs about A's concern for income, player A's payoff in the forgone outcome must be hidden from player B, and player A's payoffs in player B's two possible outcomes must be hidden from player A. This is the exact method I employ in my opaque treatment game.[20]

I define the perceived signal of income concern using a *weighted average* of player A's payoffs in player B's two possible outcomes on the assumption that player A takes into account what player B is likely to do if A enters the game. Take the following transparent game:

[20] While it is true that player A would have little incentive to "enter" if she could not see her possible payoffs from entering, in my experiment I employ the "strategy method" of elicitation, where player B chooses a contingent outcome before she learns of player A's decision, thus player's A actual decision to enter or not is almost irrelevant.



For the sake of argument, suppose player B infers player A's degree of income concern by comparing A's payoff in the forgone outcome to a simple *arithmetic mean* of A's payoffs in the two possible outcomes. In this game, if player A decides to enter, then player B would perceive that A is concerned about income (the mean of 200 and 700 being greater than 350). But the presupposition here is that player B assumes that player A presumes that B would be willing to sacrifice 500 to provide A with the higher payoff of 700. This hardly seems credible: relative to player B's highest possible payoff, 500 is a large sacrifice to make—surely, if player A was indeed rationally seeking to maximize her income, she would notice this conflict in their self-interests before taking action. More importantly, I assume that player B would presume that player A surely notices this conflict.[21] Hence, since player B presumes that player A knows that the likelihood of B sacrificing 500 for A's own welfare is low, player B would in fact perceive that player A is unconcerned about income. That is, player B's reference point for A's degree of income concern depends not only on A's payoffs, but also on their relationship to B's possible payoffs. I incorporate this into player B's *reference point for income concern* by weighting the average of A's payoffs in B's two possible

[21] It is possible that player B's presumption about A's motives could also be subject to selective processing; however, I overlook this potential confound for simplicity.

outcomes by B's payoffs in her two possible outcomes:

$$\omega_y^A \equiv \frac{\pi_1^B \pi_1^A + \pi_2^B \pi_2^A}{\pi_1^B + \pi_2^B} \tag{15}$$

In the game above, this weighted average of A's payoffs in the two possible outcomes is less than A's forgone payoff of 350. Using this weighted average as player B's reference point for income concern, player B would perceive the signal that, by entering, player A is obviously unconcerned about her income.[22]

I assume that the perceived signal indicating player A's degree of income concern is asymmetric: player B either perceives that player A derives increasing utility from income, or she perceives that player A derives no utility from income; she does not perceive that player A has decreasing utility of income. In other words, A's *concern* for income is a matter of degree: depending on the measure of the difference between player B's reference point for income concern and A's payoff in the forgone outcome, player B could perceive that player A is somewhat concerned for income, or greatly concerned for income, and so on. On the other hand, a perceived signal that player A is somewhat *unconcerned* about income is ostensibly indistinguishable from a perceived signal that A is greatly unconcerned about income—in both cases, player B simply updates her beliefs to reflect A's lack of concern, and then pursues her own intrinsic preferences. Player B would not ever perceive that player A dislikes income, as this would imply that out of concern for A's utility, player B would be willing to choose a pareto-inferior outcome—that is, she would sacrifice her own payoff in order to decrease A's payoff as well (and do it as a favor to A).

To ensure accurate estimates of player B's posterior-belief weights, I separate the perceived signal variable by its asymmetrical halfs. For any simplified ultimatum game Γ, the *perceived*

[22] I call this the "What are you, crazy?" reference point—that being the voice in B's head rejecting the notion that A could rationally presume such a large sacrifice from her.

signal indicating player A's degree of concern for income is:

$$s_y^c(\Gamma) = T * max(\omega_y^A - \pi_0^A,\ 0) \tag{16}$$

where, by assumption, $T = 1$ in a transparent game where all payoffs are public to both players, and $T = 0$ in an opaque game where player A's payoff in the forgone outcome is hidden from player B, and player A's payoffs in player B's two possible outcomes are hidden from player A—the signal is not available in opaque games. If player A enters the game, then the strength of the signal indicating player A's degree of concern for income will determine the weight player B assigns to A's payoff in her posterior beliefs: the stronger the signal, the larger the weight. If A's payoff in the forgone outcome is less than player B's reference point for income concern, then the measure of the signal will equal their difference. Otherwise, player B does not perceive that player A is concerned about income. Instead, she will likely perceive that player A is unconcerned about income, where the *perceived signal indicating that player A is unconcerned about income* is:

$$s_y^u(\Gamma) = \begin{cases} T & \text{if } \ \omega_y^A < \pi_0^A \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

If player A's payoff in the forgone outcome is greater than player B's reference point for income concern, then, by assumption, player B will zero the weight she assigns to A's payoff in her posterior beliefs. I do not expect player B to decrease this weight any further than zero, because this would imply that player B perceives that player A is averse to income, and thus B's propensity to choose pareto-inferior outcomes would increase.[23]

Like the inequality aversion model, I define player B's moral wriggling function in terms of her

[23] As in the inequality aversion model, the complete income concern model incorporates player B's prior beliefs in the first parameter. Although I expect player B to decrease her posterior belief weight to zero, the unconcerned-parameter is an "additive" (or perhaps more accurately, a "subtractive") weight, and thus would have a negative value to offset the weight that player B assigns to A's payoff in her prior beliefs.

own payoff only. Player B's *incentive to morally wriggle* is the indicator function:

$$m(\pi_k^B, \pi_{3-k}^B) = \begin{cases} 1 & \text{if } \pi_k^B < \pi_{3-k}^B \\ 0 & \text{if } \pi_k^B \geq \pi_{3-k}^B \end{cases} \tag{18}$$

where π_k^B is player B's payoff in the outcome under evaluation, and π_{3-k}^B is player B's payoff in the other possible outcome. I assume that player B's intrinsically-preferred outcome is the outcome with the highest payoff. Although player B may have an intrinsic preference for altruism, a perceived signal indicating that player A is *concerned* about income could not possibly conflict with this intrinsic preference, thus player B will only try to morally wriggle around a perceived "concerned" signal if her posterior beliefs would conflict with her own-payoff preference. On the other hand, a perceived signal that player A is *unconcerned* about income could not possibly conflict with player B's own payoff preference. Furthermore, an unconcerned-signal could not possibly conflict with any intrinsic preference player B might have for altruism either: if player B perceives that player A is unconcerned about income, she merely reduces the weight she assigns to A's payoff *in her beliefs* to zero; she does not need to sacrifice her intrinsic preference for altruism. Therefore, in the income concern model, the incentive to morally wriggle applies only to the perceived signal that player A is concerned about income, and not to the perceived signal that player A is unconcerned about income.

Incorporating moral wriggling, I derive the complete income concern model from the general model of selective beliefs. I rearrange equation (7) from the general model for a single property of outcomes, where $\alpha_i = 1$, and $s_{\hat{\beta}} = (s_{\hat{\beta}1}, s_{\hat{\beta}2})$ are the asymmetrical halfs of the income concern signal and $\hat{\beta}_j = (\hat{\beta}_{j1}, \hat{\beta}_{j2})$ their respective weights:[24]

$$V_i(\pi_i, Q(\pi_i, \pi_j), \bar{U}_i, s_{\hat{\beta}}) = \pi_i + [\beta_i + (\hat{\beta}_j \,|\, s_{\hat{\beta}}, m(\pi_i, Q(\pi_i, \pi_j), \bar{U}_i))]\, Q(\pi_i, \pi_j) \tag{19}$$

[24] See the discussion of equation (10) in Section 3.2.1 for an explanation of the constant weight of concern α_i assumption.

Next, I rewrite equation (19) for the income concern model: player i becomes player B, the property of outcomes $Q(\pi_i, \pi_j)$ becomes player A's payoff π_k^A, and the intrinsically-preferred outcome and each half of the perceived signal indicating player A's concern for income are now endogenous. To separate player B's posterior beliefs from her prior beliefs, I change the first parameter to incorporate a mix of player B's intrinsic preferences *and* her prior beliefs.[25] And, to measure differences in player B's posterior beliefs due to moral wriggling, I separate the parameters of the "concerned" signal by the incentive to morally wriggle.

3.3.2 Complete Model

For any simplified ultimatum game Γ, where player B's possible outcomes are (π_k^A, π_k^B) for $k \in \{1, 2\}$, and $(\pi_{3-k}^A, \pi_{3-k}^B)$ is the possible outcome not under evaluation, and (π_0^A, π_0^B) is Player A's forgone outcome, the complete income concern model is:

$$V_y^B(\pi_k^A, \pi_k^B) = \pi_k^B + [\tilde{\delta}_B + \hat{\delta}_A^u s_y^u(\Gamma) + \hat{\delta}_A^{c,m} s_y^c(\Gamma) m(\pi_k^B, \pi_{3-k}^B) + \hat{\delta}_A^{c,n} s_y^c(\Gamma)(1 - m(\pi_k^B, \pi_{3-k}^B))]\, \pi_k^A \quad (20)$$

where $s_y^u(\Gamma)$ is a perceived signal indicating that player A is unconcerned about income, $s_y^c(\Gamma)$ is a perceived signal indicating player A's degree of concern for income, and $m(\pi_k^B, \pi_{3-k}^B) = 1$ if, for the outcome under evaluation, player B has an incentive to morally wriggle with respect to her posterior beliefs about player A's concern for income. $\tilde{\delta}_B$ is the weight player B assigns to player A's payoff due to her intrinsic preferences and her prior beliefs about player A's degree of concern for income. The following parameters are "additive" (or "subtractive"), in that they reflect the adjusted weight that player B's posterior beliefs *add* to (or *subtract* from) the weight already captured in her prior beliefs, not her posterior beliefs in full. $\hat{\delta}_A^u$ is the subtractive weight that player B assigns to player A's payoff due to her posterior belief that player A is unconcerned about income. The two subsequent parameters are "adjusted" weights, indicating that their values are not directly comparable to that of $\tilde{\delta}_B$, due to the presence of the non-indicator "concerned" signal

[25] See the discussion of equation (12) in Section 3.2.1 for an explanation of this change.

variable. $\hat{\delta}_A^{c,m}$ is the additive adjusted weight that player B assigns to player A's payoff due to her posterior beliefs about player A's degree of concern for income where, for the outcome under evaluation, there is an incentive for player B to morally wriggle with respect to her posterior beliefs. $\hat{\delta}_A^{c,n}$ is the additive adjusted weight that player B assigns to player A's payoff due to her posterior beliefs about player A's degree of income concern where, for the outcome under evaluation, there is no incentive for player B to morally wriggle.

3.3.3 Predictions

The selective beliefs hypothesis implies that $\hat{\delta}_A^u \leq 0 \leq \hat{\delta}_A^{c,m} < \hat{\delta}_A^{c,n}$. If player B perceives a signal that player A is concerned about income and has no incentive to morally wriggle around this perception, then her propensity for altruistic behavior will be greater than her intrinsic and prior propensity to for altruistic behavior without a perceived signal. However, if player B does have an incentive to morally wriggle, then her propensity for altruistic behavior will be less than if she did not have an incentive to morally wriggle, and perhaps no greater than her intrinsic and prior propensity for altruistic behavior. If player B perceives a signal that player A is unconcerned about income, then her propensity for altruistic behavior will be less or equal to than her intrinsic and prior propensity for altruistic behavior, depending on her prior beliefs about player A's concern for income. The decrease in her propensity for altruistic behavior would, at most, fully offset her prior propensity for altruistic behavior due to her prior beliefs; it would not affect her propensity for altruistic behavior due to her intrinsic preferences.[26]

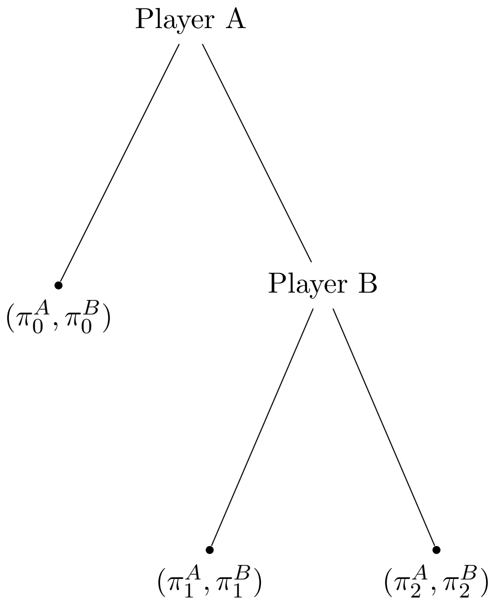
Player B may be intrinsically altruistic, or she may be intrinsically spiteful. Suppose that player B is intrinsically spiteful. If we also suppose that player B holds the prior belief that player A is unconcerned about income (and thus assigns no weight to player A's payoffs in her prior beliefs), then $\tilde{\delta}_B < 0$. However, if player B is intrinsically spiteful but also holds the prior belief that player

[26] On the assumption that a perceived signal that player A is unconcerned about income causes player B to fully withdrawal the weight she assigns to player A's payoff due to prior beliefs—that is, selective perception dominates selective interpretation—the estimated value of $\hat{\delta}_A^u$ would equal the additive inverse of player B's altruism due to prior beliefs, thus allowing us to parse intrinsic preferences from prior beliefs in the parameter $\tilde{\delta}_B$.

A is concerned about income to some degree, then $\tilde{\delta}_B \geq 0$ if and only if the weight she assigns to player A's payoff due to her prior beliefs is large enough to fully offset her spitefulness. On the other hand, if player B is intrinsically altruistic, then $\tilde{\delta}_B > 0$ necessarily, because I assume that player B never holds the prior belief that A derives decreasing utility from income—absent mean treatment, spitefulness does not arise from beliefs.

4 Experiment Design

The experiment is designed to test the two fundamental assumptions of selective beliefs in the inequality aversion and income concern models: that players respond positively to perceived signals in others' behavior indicating others' preferences, and that players ignore or underweight these perceived signals if others' indicated preferences do not complement their own self-interest. Both models apply specifically to the simplified ultimatum game of the design employed by Charness and Rabin (2002). I assume that player B (the second-mover) infers the preferences of player A (the first-mover), based on the properties of a single "forgone" outcome (π_0^A, π_0^B) relative to the properties of B's two possible outcomes, (π_1^A, π_1^B) and (π_2^A, π_2^B), where each game is structured:



First, player A decides either to "exit" the game such that she earns a payoff of π_0^A and player B earns a payoff of π_0^B, or to "enter" the game and pass the decision on to player B. If player A decides to enter, then player B decides between a payoff of π_1^B for herself and π_1^A for player A, or a payoff of π_2^B for herself and π_2^A for player A. This simple design allows for the relatively straightforward estimation of model parameters with as few relevant outcomes as possible. (See Sections 3.2.3 and 3.3.3 for the specific predictions of the selective beliefs hypothesis for each model. I provide experiment procedures and sample instructions in the appendix.)

Both models require that I control for reciprocal motivations among players. Therefore, I use the simplified ultimatum game for both the baseline and treatment, which ensures that the first-mover–second-mover relationship is maintained across experiments and players' causal attribution in each game remains constant. The baseline and the treatment also use the same menus of outcomes, which were chosen to diversify the mix of perceived signals across games. Moreover, the strength of the perceived signals for inequality aversion and income concern vary across games, as does the incentive to morally wriggle given a particular signal. Two of the outcome menus correspond to games of particular interest from Charness and Rabin (2002).[27] In all games, player A will either unambiguously increase player B's possible payoffs by entering the game, or she will substantially increase the average of player B's possible payoffs by entering. This ensures that player B does not withdrawal concern for player A's preferences in response to perceived unkind treatment. Although player A always makes her decision before player B, I employ the "strategy method" of preference elicitation, whereby player B is not informed of player A's decision before she makes a choice; instead, she chooses a "contingent" outcome that would result only if player A had decided to enter the game.[28] To facilitate a clear explanation, I present the opaque treatment first and the

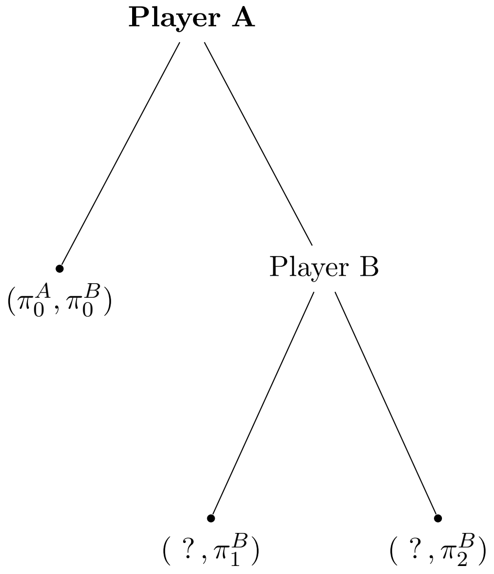
[27] Outcome menu 3 corresponds to "Barc4" and menu 12 corresponds to "Berk21" in Charness and Rabin (2002), where second-movers were less generous than in the baseline dictator game, despite the fact that the first-mover had unambiguously increased the second-mover's payoffs.

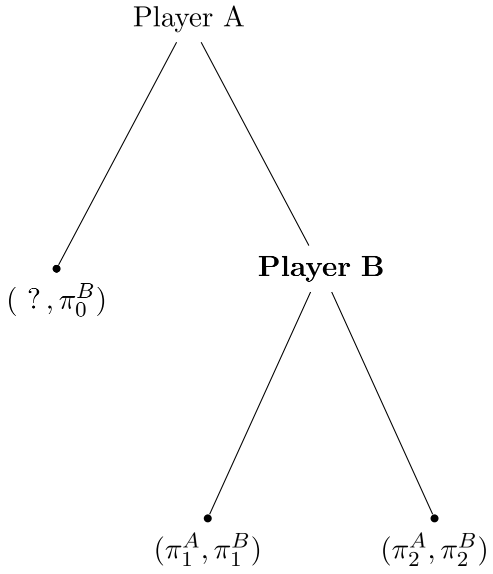
[28] Since player A's decision does not affect the decision of player B, the strategy method helps overcome a potential problem in the treatment game, where I presume few players A would wish to "enter," because they are unable to see their possible payoffs. Nevertheless, this problem is not solved entirely: player B could potentially find the possibility that player A would decide to gamble on hidden payoffs so incredulous that it would affect her choice of outcomes.

transparent baseline second.

4.1 Opaque Treatment

In the treatment game, players make decisions based on their intrinsic preferences and *prior* beliefs about others' preferences. This is accomplished by hiding relevant information—that is, the potential signals—that would allow player B to form posterior beliefs about player A's preferences. The perceived signal for inequality aversion and the perceived signal for income concern each have two component signals, which, fortunately, are closely related.[29] To block the negative component signal for both inequality aversion and income concern, player A's payoff in the forgone outcome is hidden from player B. This ensures that player B cannot infer a payoff distribution or level of income that player A dislikes or finds unsatisfactory. To block the positive component signal, player A's payoffs in player B's two possible outcomes are hidden from player A. Before making her choice, player B is informed that player A cannot see these payoffs in B's outcomes. This ensures that player B cannot infer payoff distributions or levels of income that player A does like or find satisfactory. Player A sees the game below-left, and Player B sees the game below-right:





[29] See Section 3.1 for an explanation of component signals.

Before each game, players are first shown the game with all three of player A's payoff hidden. Players are then informed of which payoffs will be revealed to which players. This ensures that both players have common knowledge of the payoffs that the other player can and cannot see.

4.2 Transparent Baseline

In the baseline game, all payoffs are public to all players, and thus players make decisions based on their intrinsic preferences and *posterior* beliefs about others' preferences given perceived signals indicating others' aversion to inequality or concern for income. If players do indeed respond positively to perceived signals of others' inequality aversion, then player B's propensity to choose a more equal outcome would increase when the inequality of the forgone outcome increases relative to B's two possible outcomes and the more equal outcome also provides player B with the highest possible payoff. If players also morally wriggle with respect to their beliefs about others' inequality aversion, then player B's propensity to choose a more equal outcome would be lower when the more equal outcome provides player B with a lower payoff. With income concern, if players do indeed respond positively to perceived signals of others' concern for income, then player B's propensity for altruism would increase when player A's payoff in the forgone outcome is less than the weighted average of player A's payoffs in B's two possible outcomes and A's highest payoff coincides with B's highest payoff. If people also morally wriggle with respect to their beliefs about others' concern for income, then player B's propensity for altruism would be lower when A's highest payoff conflicts with B's highest payoff. Finally, if people do respond positively to perceived signals that others are unconcerned about income, then player B's propensity for altruism would decrease when player A's payoff in the forgone outcome is greater than the weighted average of player A's payoffs in B's two possible outcomes (although, if player B held the prior belief that player A was unconcerned about income, there would be no change in her propensity for altruism).

The baseline game also provides a test for the dominance of either selective perception or selective interpretation. Selective perception occurs when player B either totally ignores a signal or perceives it in full. Selective interpretation occurs when player B either underweights or overweights the magnitude of a perceived signal. Therefore, if, *given the incentive to morally wriggle*, player B's posterior propensity to choose more equal outcomes or for altruistic behavior is *not* significantly greater than her prior propensity, then selective perception would dominate selective interpretation—this would imply that player B tends to ignore signals that conflict with her self-interest, rather than underweight them. On the other hand, if, given the incentive to morally wriggle, player B's posterior propensity for either behavior *is* significantly greater than her prior propensity, then selective interpretation would dominate selective perception—this would imply that player B tends to underweight signals that conflict with her self-interest, rather than ignore them. The models do not allow for the parsing of overweighted perceived signals from strictly perceived signals, however.

5 Conclusion

People tend to believe what they want to believe. If I can convince myself that you like football, then I see no reason why we shouldn't forgo the opera and attend the football game. If a player in an ultimatum game can convince herself that the other player is unconcerned about income, then why should she sacrifice her own payoff to increase his payoff? And if an employee observes that his employer seemed unconcerned about the deadline, why should he work overtime to get the job done? The selective beliefs model can apply to a wide range of social and economic situations. It may explain why people who are otherwise concerned for the welfare of others will behave selfishly in many circumstances: they selectively interpret and perceive signals about others' preferences such that they *believe* that they are acting in others' interest. This form of moral wriggling only goes so far, however. If the employer had seemed concerned about the job's completion, then perhaps the employee can find no excuse to shirk on his responsibility.

Of course, people are different. Andreoni and Miller (2002) and Engelmann and Strobel (2007) find that subjects' preferences are highly heterogeneous in experimental games. Likewise, some subjects may be more prone to signal perception and selective processing than others, and this is a significant drawback to the regression analysis of experiment results.[30] Also, subjects may respond differently to positive and negative perceived signals. The experiment can be modified to partially test for this difference by hiding *either* player A's forgone payoff from player B, or hiding A's payoffs in B's possible outcomes from player A (and ensuring common knowledge). The former would test player B's behavioral response to a positive signal only, whereas the latter would test for B's response to only a negative signal. (The positive signal is slightly ambiguous in the simplified ultimatum game, however, so this may not be the best design for the task.)

In the case of inequality aversion, the method of hiding relevant payoffs could also be employed to parse prior beliefs from intrinsic preferences using a dictator game. In a transparent baseline game, the experimenter would assure the dictator that the receiver will learn of the dictator's payoff, thus when the dictator made her decision, she would be concerned about the receiver's degree of aversion to inequality. As an opaque treatment, the experimenter would assure the dictator that the receiver will *not* learn of the dictator's income—in this case, she would be unconcerned about the receiver's aversion to equality. Therefore, the dictator would employ her prior beliefs about the receiver's aversion to inequality in the baseline, but not in the treatment game. Then, the difference in the dictator's propensity to choose equal outcomes in the baseline verses the treatment would be due to the dictator's prior beliefs about the receiver's aversion to inequality. Obviously, this method is not applicable to the dictator's prior beliefs about the receiver's concern for income, although the experimental method presented here may have further applications.[31]

The experiment and models presented in this paper serve as only starting point for the potential study of selective beliefs. Their implications occupy a middle ground between the fairness norms of

[30] Parameters should be significant so long as a substantial share of subjects exhibit behavior consistent with the models' predictions. For example, Fehr and Schmidt (1999), from whom my inequality aversion model is derived, argue that only 40 percent of subjects need exhibit inequality-aversion for results to be consistent with their model.

[31] In the simplified ultimatum game experiment, if selective perception dominates selective interpretation, then the subtractive weight that player B assigns to player A's payoff due to her posterior belief that player A is unconcerned about income allows for the parsing of player B's prior beliefs about player A's concern for income from her intrinsic preferences. See footnote 26 in Section 3.3.3.

recent social preferences models and the standard neoclassical assumption of self-interest: fairness is often subjective, not objective—and if *we believe* that our self-interested pursuits are fair, then objective norms of fairness are no obstacle. Though the selective beliefs hypothesis does little to simplify the study of social preferences, I hope that I have at least enlarged the toolkit available for future research.

Appendix: Experiment Procedures

EXPERIMENT INSTRUCTIONS

The experiment is conducted over at least two sessions. Only one game type – either the baseline or the treatment – is played per session. Subjects are guaranteed a participation payment of $5. Subjects may be paid for randomly determined outcomes only, depending on available funds.

1. Subjects randomly draw cards lettered in equal proportions either “A” or “B” to determine their role in the experiment.
2. Subjects are divided by role into two rooms, an “A” room and a “B” room.
3. Subjects, now in separate rooms, are given the Subject Instructions (not the Game Sheets, yet). The experimenter may read the instructions aloud or let the subjects read them privately. Subjects are given the opportunity to ask questions.
4. Players A randomly draw numbered cards to determine their B match for the first outcome menu (subject numbering may be determined by random seating order, or some other random assignment method).
5. **Baseline only**: Players A are given the Player A Game Sheet for outcome menu 1, which they read privately and silently. Each player A chooses either to “exit” the game for a certain outcome, or to “enter” the game and have the matched player B choose between two outcomes. Collect Player A Game Sheets.

   Treatment only: Players A are given the Player A Game Sheet for outcome menu 1, FRONT SIDE UP, which they are given thirty seconds to read privately and silently. After thirty seconds, the subjects are told to turn the game sheet over. Each player A chooses either to “exit” the game for a certain outcome, or to “enter” the game and have the matched player B choose between two outcomes. Collect Player A Game Sheets.
6. **Baseline only**: Players B are given the Player B Game Sheet for outcome menu 1, which they read privately and silently. Players B are not informed of Players A decisions, but are told that their decision will only affect the outcome if their matched player A has decided to “enter” (the strategy method of elicitation). Each player B chooses an outcome. Collect Player B Game Sheets.

   Treatment only: Players B are given the Player B Game Sheet for outcome menu 1, FRONT SIDE UP, which they are given thirty seconds to read privately and silently. After thirty seconds, the subjects are told to turn the game sheet over. Players B are not informed of Players A decisions, but are told that their decision will only affect the outcome if their matched player A has decided to “enter” (the strategy method of elicitation). Each player B chooses an outcome. Collect Player B Game Sheets.
7. Repeat steps 4-6 for each outcome menu. Subjects retain their roles (players A or B) for each outcome menu such that hidden payoffs remain consistently hidden for each player over all outcome

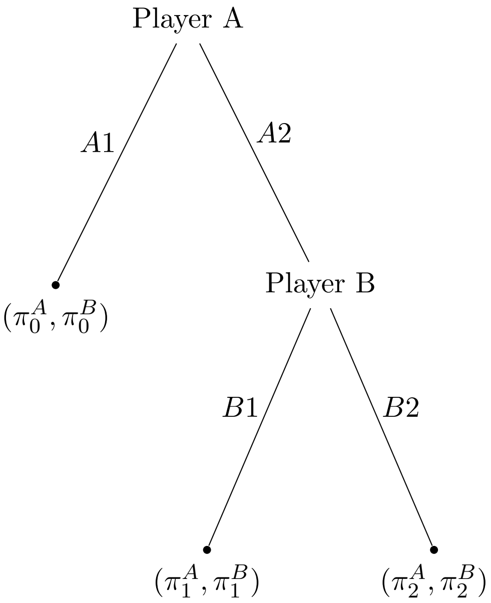
menus. This ensures that subjects do not attempt to guess the value of hidden payoffs based on instances when payoffs are transparent for a different role. (This is also the motivation behind separating game types by session.)

8. Upon completion, subjects are paid privately for game outcomes and participation.

GAME OUTCOME MENUS

Both the baseline and the treatment use these outcome menus ($100 \approx \$1$) with this game structure.

menu	(π_0^A, π_0^B)	(π_1^A, π_1^B)	(π_2^A, π_2^B)
1	(500, 300)	(400, 400)	(600, 375)
2	(350, 350)	(200, 800)	(700, 300)
3	(800, 0)	(400, 400)	(750, 375)
4	(100, 350)	(500, 450)	(550, 450)
5	(1000, 0)	(500, 500)	(0, 1000)
6	(100, 100)	(800, 200)	(200,400)
7	(400, 0)	(500, 500)	(450, 550)
8	(1000, 0)	(500, 500)	(500, 550)
9	(250, 100)	(1000, 0)	(450, 550)
10	(1000, 0)	(500, 500)	(450, 550)
11	(100, 350)	(500, 500)	(550, 450)
12	(750, 0)	(400, 400)	(750, 375)
13	(100, 500)	(500, 500)	(500, 550)
14	(450, 500)	(500, 550)	(450, 500)
15	(800, 0)	(900, 100)	(0, 1000)
16	(400, 300)	(500, 500)	(750, 550)



Games with an inequality aversion signal: 1, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 16
Incentive to morally wriggle: 5, 7, 8, 10, 13, 16
No incentive to morally wriggle: 1, 3, 4, 9, 11, 12

Games with a "concerned" income signal: 4, 6, 7, 9, 11, 13, 14, 16
Incentive to morally wriggle: 6, 7, 9, 11
No incentive to morally wriggle: 4, 13, 14, 16

Games with an "unconcerned" income signal: 1, 2, 3, 5, 8, 10, 12, 15

SUBJECT INSTRUCTIONS

Thank you for participating in this experiment. You will receive $5 for your participation, in addition to other money to be paid as a result of decisions made in the experiment.

You will make decisions in several different situations ("games"). Each decision (and outcome) is independent from each of your other decisions, so that your decisions and outcomes in one game will not affect your possible choices and outcomes in any other game.

In every case, you will be anonymously paired with one other person, so that your decision may affect the payoff of the other person, just as the decisions of the other person may affect your payoff. For every decision task, you will be paired with a different person or persons than in previous decisions.

There are two "roles" in each game – either A or B. Each game has multiple decisions, and these decisions will be made sequentially, in alphabetical order: "A" players will complete their decision sheets first and their decision sheets will then be collected. Next, "B" players complete their decision sheets and these will be collected. You will not be informed of the results of any previous game prior to making your decision.

When you have made a decision, please turn your decision sheet over, so that we will know when people have finished.

At the end of the session, you will be given a receipt form to be filled out and you will be paid individually and privately.

Please feel free to ask questions at any point if you feel you need clarification. Please do so by raising your hand. Please DO NOT attempt to communicate with any other participants in the session until the session is concluded.

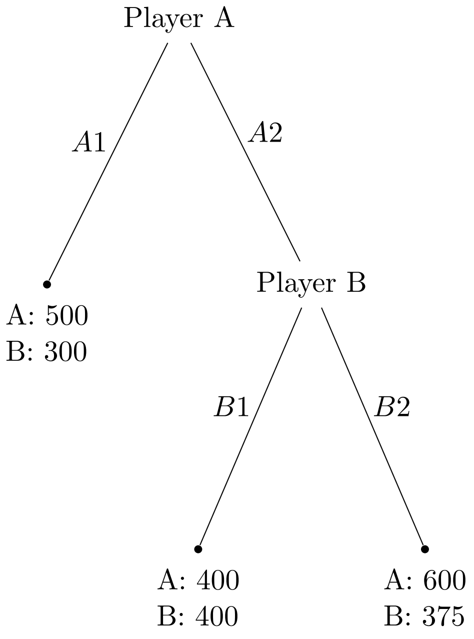
We will proceed to the decisions once the instructions are clear. Are there any questions?

PLAYER A – BASELINE, OUTCOME MENU 1

You are player A. You may choose A1 or A2.

If you choose A1, you would receive 500 and player B would receive 300. If you choose A2, then player B's choice of B1 or B2 would determine the outcome. If you choose A2 and player B chooses B1, you would each receive 400. If you choose A2 and player B chooses B2, you would receive 600 and he or she would receive 375.

Player B will make a choice without being informed of your decision. Player B knows that his or her choice only affects the outcome if you choose A2, so that he or she will choose B1 or B2 on the assumption that you have chosen A2 over A1.



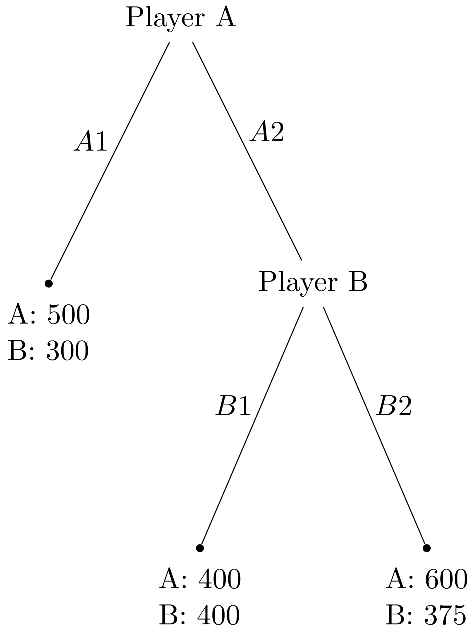
DECISION

I choose: A1____ A2____

PLAYER B – BASELINE, OUTCOME MENU 1

You are player B. You may choose B1 or B2.

Player A has already made a choice. If he or she has chosen A1, he or she would receive 500 and you would receive 300. Your decision only affects the outcome if player A has chosen A2. Thus, you should choose B1 or B2 on the assumption that player A has chosen A2 over A1. If player A has chosen A2 and you choose B1, you would each receive 400. If player A has chosen A2 and you choose B2, then player A would receive 600 and you would receive 375.



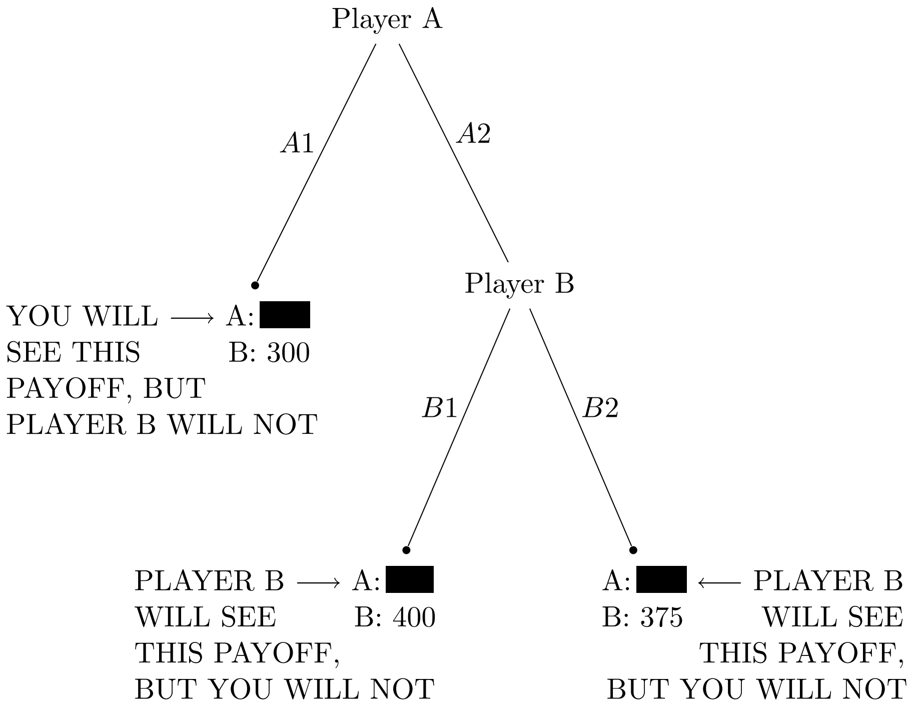
DECISION

I choose: B1____ B2____

PLAYER A – TREATMENT, OUTCOME MENU 1 – FRONT

You are player A. When you are told to do so, you may turn this sheet over and read the game instructions. Please do not turn this sheet over until you are told to do so.

Right now, your payoffs are hidden. Once you turn this sheet over, you will see your payoff below A1, but player B will not see it. Also, player B will see your payoffs below B1 and B2, but you will not see them.


Player A
A1
A2
Player B
B1
B2
YOU WILL SEE THIS PAYOFF, BUT PLAYER B WILL NOT
A: [redacted]
B: 300
PLAYER B WILL SEE THIS PAYOFF, BUT YOU WILL NOT
A: [redacted]
B: 400
A: [redacted]
B: 375
PLAYER B WILL SEE THIS PAYOFF, BUT YOU WILL NOT

PLEASE DO NOT TURN THIS SHEET OVER UNTIL YOU ARE TOLD TO DO SO

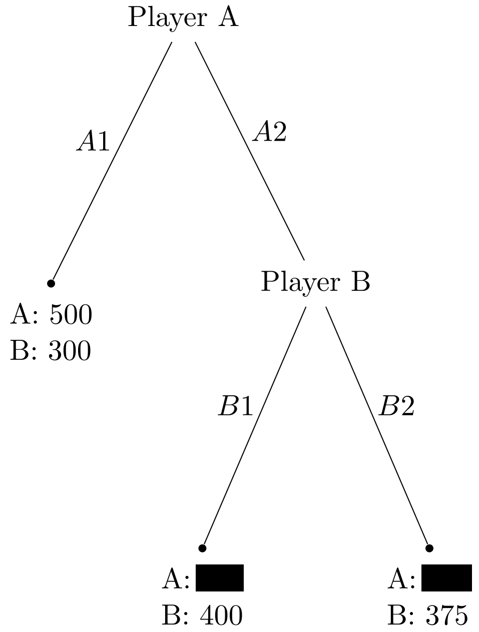
PLAYER A – TREATMENT, OUTCOME MENU 1 – BACK

You are player A. You may choose A1 or A2.

If you choose A1, you would receive a private payoff of 500 and player B would receive 300. Player B CANNOT SEE your payoff of 500 and would not know how much you received. Player B would only know his or her own payoff of 300.

If you choose A2, then player B's choice of B1 or B2 would determine the outcome. If you choose A2 and player B chooses B1, you would receive a payoff that may be higher or lower than 500, and player B will receive 400. If you choose A2 and player B chooses B2, you would receive a payoff that may be higher or lower than 500, and player B will receive 375. Player B can see both of your hidden payoffs.

Player B will make a choice without being informed of your decision. Player B knows that his or her choice only affects the outcome if you choose A2, so that he or she will choose B1 or B2 on the assumption that you have chosen A2 over A1.



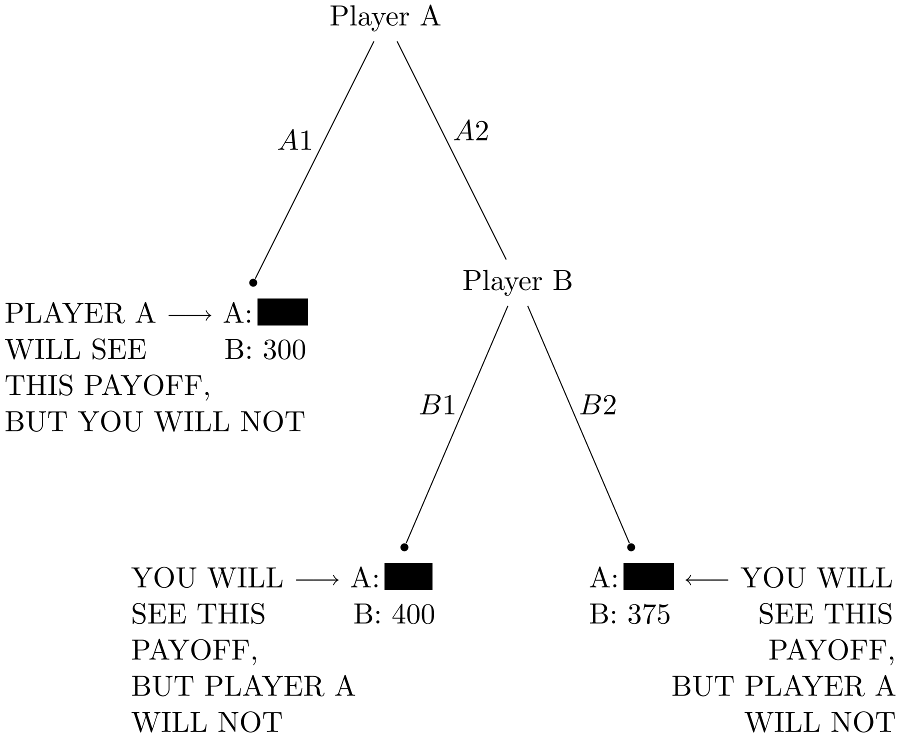
DECISION

I choose: A1____ A2____

PLAYER B – TREATMENT, OUTCOME MENU 1 – FRONT

You are player B. When you are told to do so, you may turn this sheet over and read the game instructions. Please do not turn this sheet over until you are told to do so.

Right now, player A's payoffs are hidden. Once you turn this sheet over, you will see player A's payoffs below B1 and B2, but player A will not see them. Also, player A will see his or her payoff below A1, but you will not see it.



PLEASE DO NOT TURN THIS SHEET OVER UNTIL YOU ARE TOLD TO DO SO

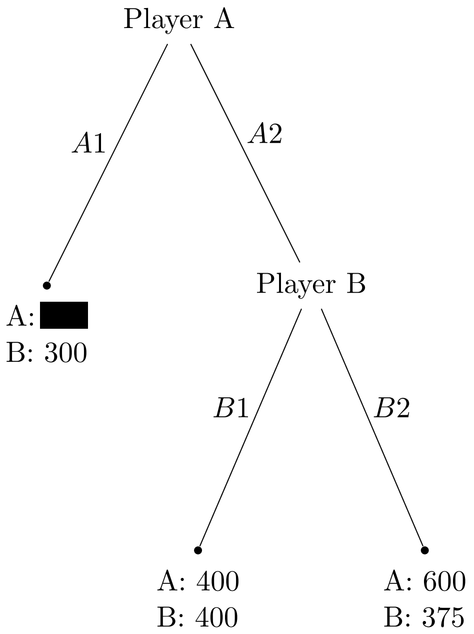
PLAYER B – TREATMENT, OUTCOME MENU 1 – BACK

You are player B. You may choose B1 or B2.

Player A has already made a choice. If player A has chosen A1, he or she would receive a private payoff and you would receive 300. Player A can see his or her private payoff.

Your decision only affects the outcome if player A has chosen A2. Thus, you should choose B1 or B2 on the assumption that player A has chosen A2 over A1. **Player A cannot see his or her own payoffs under B1 or B2**. He or she knows only that they may be higher or lower than his or her private payoff. However, player A can see your payoffs under B1 or B2.

If player A has chosen A2 and you choose B1, you would each receive 400. If player A has chosen A2 and you choose B2, then player A would receive 600 and you would receive 375.



DECISION

I choose: **B1**____ **B2**____

References

Akerlof, G. A., and J. L. Yellen (1990): "The Fair Wage-Effort Hypothesis and Unemployment," *Quarterly Journal of Economics*, 105(2), 255-283.

Andreoni, J. (1990): "Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving," *Economic Journal*, 100(401), 464-477.

Andreoni, J., and J. Miller (2002): "Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism," *Econometrica*, 70(2), 737-753.

Blount, S. (1995): "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences," *Organizational Behavior and Human Decision Processes*, 63(2), 131-144.

Bolton, G. E., and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocity and Competition," *American Economic Review*, 90(1), 166-193.

Bolton, G. E., J. Brandts, and A. Ockenfels (1998): "Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game," *Experimental Economics*, 1(3), 207-219.

Brandts, J., and C. Solà (2001): "Reference Points and Negative Reciprocity in Simple Sequential Games," *Games and Economic Behavior*, 36(2), 138-157.

Charness, G. (2004): "Attribution and Reciprocity in an Experimental Labor Market," *Journal of Labor Economics*, 22(3), 665-688.

Charness, G., and M. Dufwenberg (2006): "Promises and Partnership," *Econometrica*, 74(6), 1579-1601.

Charness, G., and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117(3), 817-869.

Charness, G., and M. Rabin (2005): "Expressed Preferences and Behavior in Experimental Games," *Games and Economic Behavior*, 53(2), 151-169.

Cooper, R., D. V. DeJong, R. Forsythe, and T. W. Ross (1992): "Communication in Coordination Games," *Quarterly Journal of Economics*, 107(2), 739-771.

Cox, J. C., and V. Sadiraj (2007): "Direct Tests of Models of Social Preferences and a New Model," working paper, Georgia State University.

Dana, J., R. A. Weber, and J. X. Kuang (2007): "Exploiting Moral Wriggle Room: Experiments Demonstrating an Illusory Preference for Fairness," *Economic Theory*, 33(1), 67-80.

Dufwenberg, M., and G. Kirchsteiger (2004): "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47(2), 268-298.

ENGELMANN, D., AND M. STROBEL (2004): "Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments," *American Economic Review*, 94(4), 857-869.

ENGELMANN, D., AND M. STROBEL (2007): "Preferences over Income Distributions: Experimental Evidence," *Public Finance Review*, 35(2), 285-310.

FALK, A., E. FEHR, AND U. FISCHBACHER (2003): "On the Nature of Fair Behavior," *Economic Inquiry*, 41(1), 20-26.

FALK, A., E. FEHR, AND U. FISCHBACHER (2008): "Testing Theories of Fairness–Intentions Matter," *Games and Economic Behavior*, 62(1), 287-303.

FEHR, E., AND K. M. SCHMIDT (1999): "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114(3), 817-868.

KAHNEMAN, D., J. L. KNETSCH, AND R. H. THALER (1986): "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review*, 76(4), 728-741.

LARSON, T. C. (2005): "The Use of Strategic Ignorance in Dictator Games when Payoffs are Not-Transparent," Department of Economics Master's Thesis, Emory University.

LEVINE, D. K. (1998): "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics*, 1(3), 593-622.

LI, J. (2008): "The Power of Conventions: A Theory of Social Preferences," *Journal of Economic Behavior and Organization*, 65(3-4), 489-505.

MUNYAN, L. (2005): "Patterns of Information Avoidance in Binary Choice Dictator Games," working paper, California Institute of Technology.

OFFERMAN, T. (2002): "Hurting Hurts more than Helping Helps," *European Economic Review*, 46(8), 1423-1437.

PINKLEY, R. L., T. L. GRIFFITH, AND G. B. NORTHCRAFT (1995): "'Fixed Pie' a la Mode: Information Availability, Information Processing, and the Negotiation of Suboptimal Agreements," *Organizational Behavior and Human Decision Processes*, 62(1), 101-112.

RABIN, M. (1993): "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 83(5), 1281-1302.

RABIN, M. (1995): "Moral Preferences, Moral Constraints, and Self-Serving Biases," unpublished manuscript.

SOBEL, J. (2005): "Interdependent Preferences and Reciprocity," *Journal of Economic Literature*, 43(2), 392-436.