Moral Preferences in Bargaining

Pau Juan-Bartroli
Toulouse School of Economics

Emin Karagözoğlu*
Bilkent University

06 January 2024

Abstract

We analyze the equilibrium of a bilateral bargaining game (Nash, 1953), where at least one of the individuals has a preference for morality (*homo moralis*). We show that the equilibrium set crucially depends on these moral preferences. Furthermore, our comparative static analyses provide insights into the distributional implications of individuals' moral concerns and the composition of society. A comparison of the set of equilibria in our model with those under selfish preferences, Kantian equilibrium, fairness preferences, altruistic preferences, and inequality averse preferences reveals important differences.

JEL Classification: C72, C78, D63, D64, D74.

Keywords: Bargaining, Divide-the-Dollar Game, Equal Division, Fairness, Homo moralis, Inequality Aversion, Kantian Morality, Nash Bargaining.

---

* *Corresponding Author:* Bilkent University, Department of Economics, 06800 Bilkent, Ankara, TURKEY. E-mail: karagozoglu@bilkent.edu.tr, Phone: +90-312-290 1955. Orcid Number: 0000-0003-2442-6949.

1 Introduction

Experimental work in economics has produced many robust and replicable findings. Some of these findings were hard to reconcile with the predictions of theoretical models under rationality and selfishness assumptions. An early example was the ultimatum game (Güth et al., 1982), where the responder behavior observed in the laboratory (i.e., rejection of positive but low offers) cannot be explained if one assumes (material) payoff maximizing agents with pure self-interest. In retrospect, it is hard to deny the role of the experimental evidence from the ultimatum game (Güth and Kocher, 2014) and dictator game (Engel, 2011), contradicting standard assumptions on preferences, in the development of theories of inequity aversion and social preferences (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000, among others).

Bargaining is a ubiquitous social interaction (Schelling, 1960). Whether it is a buyer and seller bargaining in a bazaar, flatmates bargaining over the division of responsibilities, political parties in a coalition bargaining over the allocation of ministries, or M&A negotiations. Everyone is involved in some sort of bargaining situation every day. Moreover, bargaining is a form of interaction where preferences for morality and fairness play an important role (Forsythe et al., 1994; Gächter and Riedl, 2005; Birkeland and Tungodden, 2014; Luhan et al., 2019). Questions such as "What if everyone acts as greedy as I do?", "What happens if my opponent strongly dislikes inequality?", "Do I want to haggle with immoral people?" or "Will my moral/fairness concerns hurt me at the bargaining table?" are naturally valid in bargaining encounters.

The implications of fairness, inequality aversion, and reference-dependent preferences in bargaining games have been extensively studied. Thanks to this body of work, we have a better understanding of opening offers (Galinsky and Müssweiler, 2001; Kimbrough et al., 2021); concession behavior (Bolton and Karagözoğlu, 2016); costly delays (Gächter and Riedl, 2005; Karagözoğlu and Riedl, 2015); wars of attrition (Embrey et al., 2015); and disagreements (Camerer et al., 2019; Karagözoğlu and Kocher, 2019) in bargaining.

In contrast, theoretical work on moral preferences in bargaining is almost completely absent. By incorporating moral preferences (à la Alger and Weibull, 2013) into a canonical model of bargaining, the current paper aims to further our understanding of bargaining outcomes. In particular, we study the equilibria of a bilateral bargaining game (Nash, 1953), where some individuals have moral concerns. The bargaining game we use is a natural framework to study Kantian morality (i.e., considering "what if everyone behaves the way I do") since it is a static game where there is joint interest in coordinating on the bargaining frontier. We tackle the following questions: (i) How does the presence (and strength) of moral preferences influence equilibrium outcomes? (ii) How does the equilibrium division depend on the composition of the society (e.g., the share of moral individuals in the society)? (iii) Who is better off in equilibrium, moral or selfish individuals? and (iv) What are the differential implications of other preferences vis-à-vis moral preferences?

Economists started modeling Kantian morality only recently, which was arguably fuelled by the efficient outcomes of various public goods and the tragedy of the commons-type experiments (Ostrom, 1990).[1] Two distinct lines of work have developed: *Kantian equilibrium* of Roemer (2010) and *homo moralis* preferences of Alger and Weibull (2013). Roemer modified the equilibrium concept, whereas Alger and Weibull opted for modifying individuals' preferences. The current paper follows Alger and Weibull (2013) in studying the equilibrium implications of moral preferences in a bargaining context.

Before describing the bargaining model and presenting an overview of our results, a brief definition of *homo moralis* is in order. In Alger and Weibull (2013), homo moralis is an individual who is concerned with the (universal) effect of his actions. Morality enters the individual's utility function through a term that factors in her utility in the hypothetical case that her action was to be selected by everyone else, and a parameter $\kappa \in [0, 1]$ that reflects the weight the individual attaches to this concern. The utility of a homo moralis is then a convex combination of the utilities from a selfish and a moral perspective. When $\kappa = 0$, the model captures *homo oeconomicus*, while when $\kappa = 1$, it captures *homo kantiensis*. Alger and Weibull (2013) show that such utility function is evolutionarily stable if preferences are private information and the matching process is assortative. This provides strong evolutionary foundations for the theory.

The simple version of the Nash Demand (ND) game, where the bargaining frontier is symmetric and linear, is called the divide-the-dollar (DD) game. For expositional simplicity and to derive several thresholds in closed-form, we focus on the bilateral DD game in the main body of the paper. We show in the online supplementary material that most results carry out to more general specifications. In the DD game, all individuals simultaneously demand a non-negative amount (bounded above by one) from a dollar. Everyone receives his demand if these demands

[1] Jean-Jacques Laffont (1975) hinted at a framework that incorporated individuals with preferences for Kantian morality. He also highlighted the lack of an explanation for the 'good outcomes' observed in the tragedy of the commons situations.

are mutually compatible (i.e., they add up to an amount less than or equal to a dollar) and zero if they are not mutually compatible. We characterize the equilibrium of this game when at least one of the individuals has moral concerns.

We have three sets of results. First, we show that the set of Nash equilibria depends on individuals' degrees of morality (Propositions 1 and 2) in Section 4. In the model, moral concerns influence individuals' decision-making in two ways. On the one hand, individuals forgo their moral payoff when choosing a demand above the egalitarian one. Then, individuals are not willing to demand above the egalitarian demand if the monetary payoff from this demand is not sufficiently high to compensate for their forgone moral payoff. On the other hand, individuals do not maximize their moral payoff when demanding below the egalitarian demand. Thus, individuals are not willing to demand below the egalitarian demand if their demand is not sufficiently high to compensate them for not maximizing their moral payoff. These constraints depend on individuals' degrees of morality and restrict the allocations that can be sustained as a Nash equilibrium. For any $\kappa > 0$, the Nash equilibrium set is smaller than the one under selfish preferences. For sufficiently strong moral preferences, the equal division is the unique Nash equilibrium.

Second, we conduct two comparative static analyses. In Section 4, we consider the bargaining interaction of a selfish and a moral individual. We study (i) how the Nash equilibrium set changes with the moral individual's degree of morality and (ii) potential equilibrium outcomes for the selfish individual when facing more and less moral individuals. For the second question, we study how the best and worst equilibrium payoffs of the selfish individual vary with his opponent's degree of morality. Our analysis shows that the best equilibrium for a selfish individual is strictly decreasing on his opponent's degree of morality. On the other hand, the worst equilibrium payoff for a selfish individual is equal to zero when his opponent has a low degree of morality, and it is equal to the equal split when his opponent has a high degree of morality. In Section 5, we consider the incomplete information case where individuals' degree of morality is their private information. Two individuals are randomly drawn from the population to play in a bilateral DD game. We study how the set of Bayesian Nash equilibria (BNE) changes with the share of individuals of each type and the types' degree of morality. We show that there is no BNE where the more moral type of individual obtains more than half of the surplus (Proposition 3). Additionally, we characterize the set of BNE, where selfish individuals obtain more than half of the surplus. We show that for these equilibria to exist, it is necessary that

(i) the moral individuals are not too moral and (ii) the share of moral individuals is sufficiently high (Proposition 4).

Third, we compare the Nash equilibrium of DD with homo moralis individuals to the Nash equilibrium with (i) selfish individuals, (ii) altruistic individuals, (iii) inequality averse individuals, and (iv) fairness-minded individuals in Section 6. Additionally, we compare it to the Kantian equilibrium with selfish individuals. We show that the Nash equilibrium set with homo moralis preferences differs from the equilibrium set derived under the alternative preferences and equilibrium concept mentioned above.

In Section 6, we argue that homo moralis preferences can be interpreted as a type of fairness preference, with the fairness ideal endogenously determined by the game structure. This reduces the degrees of freedom available to the researcher, narrowing the set of possible predictions. Van Leeuwen and Alger (2021) show empirically that despite the significant heterogeneity in individuals' preferences, most subjects are well described as having Kantian moral concerns. This emphasizes the importance of understanding the predictions of homo moralis in comparison to alternative models when interpreting the behavior observed in laboratory experiments. Our analysis in Section 6 is a step in this direction.

2 Related Literature

Alger and Weibull (2013, 2016) laid down the evolutionary foundations for homo moralis. They were followed by theoretical papers that used homo moralis preferences to explain various economic phenomena. Alger and Weibull (2017) studied the behavior of homo moralis and altruists in comparison to selfish individuals in several games. They showed that (i) homo moralis and altruists improve the efficiency of outcomes in standard public goods games, (ii) homo moralis (but not altruists) can eliminate socially inefficient equilibria in coordination games, and (iii) both altruism and homo moralis may diminish the prospects of cooperation in infinitely repeated games. Sarkisian (2021) considered optimal incentive schemes in a moral hazard framework when agents have homo moralis preferences. Alger and Laslier (2021, 2022) studied the implications of moral voters on election participation and voting outcomes. Bomze et al. (2021) provided necessary and sufficient conditions for the existence of equilibrium in two-player games where players are at least partly concerned about morality. Muñoz (2022) showed how a model with moral taxpayers can account for both non-pecuniary and material motivations to comply with taxes and characterized the solution to the government's optimal tax problem. Finally, Rivero (2023) studied a bilateral trade problem with homo moralis individuals and asymmetric information. To the best of our knowledge, the current paper is the first to theoretically investigate the implications of moral preferences in a bargaining game. As such, it contributes to the recent literature on the role of morality in strategic interactions.

Recent empirical research has shown evidence of homo moralis preferences in laboratory settings. Miettinen et al. (2020) compared six different utility functions in a sequential prisoners' dilemma game. These authors reported that homo moralis performs among the best. Van Leeuwen and Alger (2021) conducted a finite mixture model to determine the relative importance of Kantian concerns and social preferences. They found that despite the significant heterogeneity in individuals' preferences, most subjects are well described as having Kantian moral concerns. Finally, Capraro and Rodriguez-Lara (2022) reported that experimental subjects' moral preferences are positively associated with both proposers' offers and responders' minimum acceptable offers in ultimatum and impunity games.

Our paper also contributes to the body of work in behavioral game theory that studies nonstandard preferences in the context of bargaining games. Some related papers are Kohler (2013) (envy), Birkeland and Tungodden (2014) (fairness preferences), Bolton and Karagözoğlu (2016) (fairness preferences in concession bargaining), Karagözoğlu and Keskin (2018) (time-varying fairness concerns), Kohler and Schlag (2019) (inequality aversion), Shalev (2002), Driesen et al. (2012) and Kara et al. (2021) (reference-dependent preferences), Guha (2018) (malice), and Dizarlar and Karagözoğlu (2023) (Kantian equilibrium of a modified DD game).

Dizarlar and Karagözoğlu (2023) (DK (2023) in what follows) is of special interest here since it also studied Kantian morality in a bargaining game. Here are the main differences: (i) the current paper studies moral preferences using the Nash equilibrium, whereas DK (2023) studies the Kantian equilibrium using selfish preferences; (ii) the current paper studies the DD and ND games, whereas DK (2023) only studies a modified version of the DD game (Ashlagi et al., 2012); (iii) the current paper focuses on comparative statics using the morality parameter and the composition of the society, whereas DK (2023) focuses on the axiomatic properties of the division rule to be applied in the case of disagreement; and (iv) the current paper compares the equilibria under preferences for morality with those under alternative models, whereas such a comparison is not present in DK (2023). As shown in Section 6, the two approaches proposed to capture moral concerns lead to very different predictions in this bargaining context.

Finally, our paper contributes to the literature motivated by the equilibrium multiplicity problem in the ND (or DD) game (see Malueg, 2010). Various researchers aimed to modify certain elements of the game to refine the set of Nash equilibria and ultimately single out equal division in equilibrium. We refer the reader to Karagözoğlu et al. (2023) and the references therein for contributions to this literature. In our model, introducing (sufficiently strong) moral preferences singles out equal division in equilibrium without changing the rules of the game. This is in line with the observations from economic experiments with symmetric bargaining environments (Nydegger and Owen, 1974; Kroll et al., 2014; Karagözoğlu and Riedl, 2015, Engler and Page, 2022).

3 Moral Preferences and the DD Game

3.1 Homo Moralis Preferences

We start by providing a formal definition of homo moralis in two-player symmetric normal-form games (Alger and Weibull, 2013).[2] Let X be the set of pure strategies and $\pi(x_1, x_2)$ the individual 1's material payoff when he chooses strategy x_1 and the individual 2 chooses strategy x_2. The utility function of individual 1 when he has homo moralis preferences is defined as follows:

$$u_{\kappa_1}(x_1, x_2) = (1 - \kappa_1)\pi(x_1, x_2) + \kappa_1 \pi(x_1, x_1),$$

where $\pi(x_1, x_2)$ is as defined above and $\pi(x_1, x_1)$ is individual 1's material payoff in the hypothetical case individual 2 were to choose the same strategy as himself. Interpreting the second term as an application of Kant's categorical imperative (Kant, 1785) "act only on the maxim that you would at the same time will to be a universal law", we refer to $\kappa_1 \in [0, 1]$ as individual 1's degree of morality. An individual with $\kappa_1 = 0$ (*homo oeconomicus*) maximizes his material payoff, $\pi(x_1, x_2)$, whereas an individual with $\kappa_1 = 1$ (*homo kantiensis*) maximizes the payoff from doing "the right thing", $\pi(x_1, x_1)$. The larger κ_1, the larger the weight the individual 1 attaches to the Kantian moral concern $\pi(x_1, x_1)$. We show how to generalize homo moralis preferences when $N > 2$ in the online supplementary material.

[2]Homo moralis preferences can also be applied to asymmetric games by considering the ex-ante symmetric version of the game where individuals choose their strategy behind the veil of ignorance.

3.2 DD Game

Consider a DD game with $N \geq 2$ individuals who simultaneously choose a demand $x_i \in [0,1]$. If the sum of their demands is less than or equal to one, each individual receives his demand, x_i. Otherwise, all individuals receive zero. Accordingly, individual i's material payoff under the strategy profile $(x_1, x_2, ..., x_n)$ can be described as follows:

$$\pi(x_i, \boldsymbol{x}_{-i}) = \begin{cases} x_i & \text{if } \sum_{j=1}^{N} x_j \leq 1 \\ 0 & \text{if } \sum_{j=1}^{N} x_j > 1 \end{cases}$$

We define ${x_i}^* \equiv \frac{1}{N}$ to be the egalitarian demand: the strategy where individual i requests $\frac{1}{N}$, and $t_{egalitarian} \equiv \{(t_1, .., t_N) \text{ s.t. } t_i = \frac{1}{N} \; \forall i \in \{1, ..., N\}\}$ to be the egalitarian distribution of the dollar: the allocation where each individual demands and receives $\frac{1}{N}$.

4 The Equilibrium Analysis

In this section, we derive the Nash equilibrium set of the bilateral DD game when individuals have homo moralis preferences. In Section 4.1, we present Lemma 1 and Corollaries 1 and 2, which will help us characterize the equilibrium set. In Section 4.2, we characterize the equilibrium set of the DD game between two homo moralis individuals with the same degree of morality (Proposition 1). In Section 4.3, we characterize the equilibrium set of the DD game between two homo moralis individuals with different degrees of morality (Proposition 2). In Section 4.4, we study the selfish individual's best and worst equilibria depending on the degree of morality of his opponent. In the online supplementary material, we show that the main results in this section can be extended in several dimensions.

4.1 Preliminary Results

Before deriving the equilibrium set, we consider how the individual's material and moral payoffs depend on a given allocation $(x_1, x_2) \in [0,1]^2$. First, consider how individual 1's moral payoff (i.e., $\pi(x_1, x_1)$) depends on (x_1, x_2). To do so, we consider two cases depending on the value of x_1. When $x_1 \leq \frac{1}{2}$, we have that $\pi(x_1, x_1) = x_1$. Intuitively, conditional on $x_1 \leq \frac{1}{2}$, the lower x_1, the larger the amount left on the table in the hypothetical case where individual 1's demand was to be chosen by individual 2. When $x_1 > \frac{1}{2}$, we have that $\pi(x_1, x_1) = 0$, as the sum of individuals' demands in the hypothetical case where both individuals were to demand $x_1 > \frac{1}{2}$ is above 1. Importantly, individuals' moral payoffs are not affected by others' demands. Second, consider how individual 1's material payoff (i.e., $\pi(x_1, x_2)$) depends on (x_1, x_2). Note that for any $x_2 \in [0, 1)$, individual 1 maximizes his material payoff by choosing $x_1 = 1 - x_2$ as any larger demand implies $x_1 + x_2 > 1$ (giving individual 1 a material payoff of zero), while any lower demand leaves money on the table.

Lemma 1 characterizes individual i's best response correspondence given his degree of morality. In what follows, we distinguish between three cases depending on whether individual i is selfish (i.e., $\kappa_i = 0$), has a low degree of morality (i.e., $\kappa_i \in (0, 0.5]$), or a high degree of morality (i.e., $\kappa_i \in (0.5, 1]$).

Lemma 1: Let $\kappa_i \in [0, 1]$ denote individual i's degree of morality, and let $x_j \in [0, 1]$ be individual j's demand. Let $BR_{\kappa_i}(x_j)$ denote individual i's best response to x_j given κ_i. Then, there exists $\underline{x_j} \equiv \frac{1-2\kappa_i}{2(1-\kappa_i)}$ and $\overline{x_j} \equiv 1 - \frac{\kappa_i}{2}$ such that:

- **Case 1**: Individual i is selfish.

$$BR_{\kappa_i=0}(x_j) = \begin{cases} 1 - x_j & \text{if } x_j \in [0, 1) \\ [0, 1] & \text{if } x_j = 1 \end{cases}$$

- **Case 2**: Individual i has a *low* degree of morality.

$$BR_{\kappa_i \in (0,0.5]}(x_j) = \begin{cases} 1 - x_j & \text{if } x_j \in [0, \underline{x_j}) \text{ and } \kappa_i \neq 0.5 \\ \{\frac{1}{2}, 1 - x_j\} & \text{if } x_j = \underline{x_j} \\ \frac{1}{2} & \text{if } x_j \in (\underline{x_j}, \frac{1}{2}] \\ 1 - x_j & \text{if } x_j \in (\frac{1}{2}, \overline{x_j}) \\ \{1 - x_j, \frac{1}{2}\} & \text{if } x_j = \overline{x_j} \\ \frac{1}{2} & \text{if } x_j \in (\overline{x_j}, 1] \end{cases}$$

- **Case 3**: Individual i has a *high* degree of morality.

$$BR_{\kappa_i \in (0.5,1]}(x_j) = \begin{cases} \frac{1}{2} & \text{if } x_j \in [0, \frac{1}{2}] \\ 1 - x_j & \text{if } x_j \in (\frac{1}{2}, \overline{x_j}) \\ \{1 - x_j, \frac{1}{2}\} & \text{if } x_j = \overline{x_j} \\ \frac{1}{2} & \text{if } x_j \in (\overline{x_j}, 1] \end{cases}$$

Proof. We distinguish between three cases depending on the value of x_j: (I) $x_j > \frac{1}{2}$, (II) $x_j < \frac{1}{2}$ and (III) $x_j = \frac{1}{2}$. For (I), individual i's utility is given by:

$$u(x_i, x_j) = \begin{cases} x_i & \text{if } x_i \in [0, 1 - x_j] \\ \kappa_i x_i & \text{if } x_i \in (1 - x_j, \frac{1}{2}] \\ 0 & \text{if } x_i \in (\frac{1}{2}, 1] \end{cases}$$

Note that $u(x_i, x_j)$ is positive and increasing in x_i in the first two intervals. We distinguish between two cases: (I-i) $\kappa_i = 0$ and (I-ii) $\kappa_i \in (0, 1]$. In (I-i), individual i's best response to $x_j > \frac{1}{2}$ is $1 - x_j$ when $x_j \in (\frac{1}{2}, 1)$ and [0, 1] when $x_j = 1$.[3] In (I-ii), there exists a threshold $\overline{x_j} = 1 - \frac{\kappa_i}{2} \in [\frac{1}{2}, 1]$ such that individual i's best response to $x_j > \frac{1}{2}$ is $x_i = 1 - x_j$ when $x_j \in (\frac{1}{2}, \overline{x_j})$, $x_i = \frac{1}{2}$ when $x_j \in (\overline{x_j}, 1]$ and $x_i = \{1 - x_j, \frac{1}{2}\}$ when $x_j = \overline{x_j}$.

For (II), individual i's utility is given by:

$$u(x_i, x_j) = \begin{cases} x_i & \text{if } x_i \in [0, \frac{1}{2}] \\ (1 - \kappa_i) x_i & \text{if } x_i \in (\frac{1}{2}, 1 - x_j] \\ 0 & \text{if } x_i \in (1 - x_j, 1] \end{cases}$$

In this case, we distinguish between three cases: (II-i) $\kappa_i \in [0, 0.5)$, (II-ii) $\kappa_i = 0.5$, and (II-iii) $\kappa_i \in (0.5, 1]$.[4] In (II-i), there exists a threshold $\underline{x_j} \equiv \frac{1-2\kappa_i}{2(1-\kappa_i)} \in [0, \frac{1}{2})$ such that individual i's best response to $x_j < \frac{1}{2}$ is $x_i = 1 - x_j$ when $x_j \in [0, \underline{x_j})$, $x_i = \frac{1}{2}$ when $x_j \in (\underline{x_j}, \frac{1}{2})$ and $x_i = \{\frac{1}{2}, 1 - x_j\}$ when $x_j = \underline{x_j}$. In (II-ii), individual i's best response to $x_j < \frac{1}{2}$ is $x_i = \frac{1}{2}$ for any $x_j \in (0, \frac{1}{2})$

[3] Note that when the individual j demands $x_j = 1$ and $\kappa_i = 0$, individual i's utility is 0 for any $x_i \in [0, 1]$. This is not the case when $\kappa_i > 0$, as moral individuals can get a positive utility by demanding $x_i \in (0, \frac{1}{2}]$.

[4] This distinction is necessary as in (II-iii) $\underline{x_j} = \frac{1-2\kappa_i}{2(1-\kappa_i)} < 0$.

and $\{\frac{1}{2}, 1\}$ when $x_j = 0$. In (II-iii), individual i's best response to $x_j < \frac{1}{2}$ is $x_i = \frac{1}{2}$ for any $x_j \in [0, \frac{1}{2})$.

Finally, for (III), note that conditional on $x_j = \frac{1}{2}$, individual i's utility is maximized at $x_i = \frac{1}{2}$ for any $\kappa_i \in [0, 1]$. □

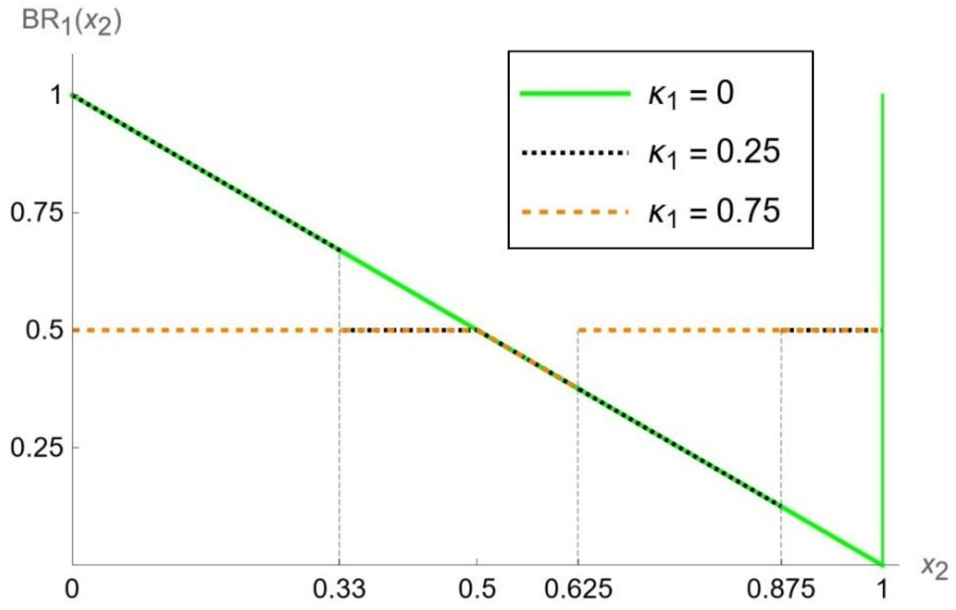

Figure 1: Individual 1's best response given $x_2 \in [0, 1]$ and (i) $\kappa_1 = 0$ in green, (ii) $\kappa_1 = 0.25$ in black or (iii) $\kappa_1 = 0.75$ in orange.

Figure 1 displays individual 1's best response given $x_2 \in [0, 1]$ and $\kappa_1 = 0$ (in green), $\kappa_1 = 0.25$ (in dotted black), or $\kappa_1 = 0.75$ (in dashed orange). Three remarks are important to emphasize. First, given x_j, the best response of a moral individual can only take the values of $\frac{1}{2}$ and $1 - x_j$. In the former, individuals maximize their moral payoff, while in the latter, they maximize their material payoff. Second, both $\underline{x_j}$ and $\overline{x_j}$ depend negatively on κ_i: the larger (resp. lower) is κ_i, the larger (resp. lower) is the range of x_j such that individual i's best response to x_j is $\frac{1}{2}$.[5] Finally, individuals' best response correspondences do not depend on the other individual's degree of morality.

We illustrate the intuition behind individuals' best response correspondence with the case with $\kappa_1 = 0.25$ (dotted black line in Figure 1). In that case, individual 1's best response is divided into four regions. First, when $x_2 \in [0, \frac{1}{3})$, individual 1's best response is $x_1 = 1 - x_2$, since $1 - x_2$ is sufficiently high, so that the material payoff he receives by demanding above $\frac{1}{2}$

[5]As it will be shown later, this has important implications for the shape of the Nash equilibrium set.

compensates him for not receiving the moral payoff. Second, when $x_2 \in (\frac{1}{3}, \frac{1}{2}]$, individual 1's best response is $\frac{1}{2}$, since $1 - x_2$ is not sufficiently high to compensate individual 1 for forgoing the moral payoff. Third, when $x_2 \in (\frac{1}{2}, 0.875)$, individual 1's best response is $1-x_2$, since $1-x_2$ is not sufficiently low for individual 1 to forgo the material payoff and demand $\frac{1}{2}$. Finally, when $x_2 \in (0.875, 1]$, $1 - x_2$ is too low for individual 1, so he prefers to deviate to demand $\frac{1}{2}$ and forgo the material payoff.[6]

Corollaries 1 and 2 follow from Lemma 1. They will later be used to characterize the equilibrium set. Corollary 1 shows that regardless of individuals' degrees of morality, the egalitarian outcome is always a Nash equilibrium. Corollary 2 shows that if at least one of the individuals has a positive degree of morality, then any Nash equilibrium must be on the bargaining frontier.

Corollary 1: Let $\kappa_1 \in [0,1]$ and $\kappa_2 \in [0,1]$ be the degrees of morality of individual 1 and individual 2, respectively. Then, ${x_1}^* = {x_2}^* = \frac{1}{2}$ is a Nash equilibrium of the bilateral DD game.

Proof. Follows from Lemma 1. For any individual $i \in \{1,2\}$ and $\kappa_i \in [0,1]$, the individual i's best response to $x_j = \frac{1}{2}$ is $x_i = \frac{1}{2}$. □

Corollary 2: Let $({x_1}^*, {x_2}^*)$ be a Nash equilibrium of the bilateral DD game where at least one individual has a strictly positive degree of morality. Then, ${x_1}^* + {x_2}^* = 1$.

Proof. A demand profile $({x_1}^*, {x_2}^*)$ is a Nash equilibrium if ${x_1}^* \in BR_{\kappa_1}({x_2}^*)$ and ${x_2}^* \in BR_{\kappa_2}({x_1}^*)$. Suppose that $\kappa_1 > 0$ and that there exists a Nash equilibrium $({x_1}^*, {x_2}^*)$ with ${x_1}^* + {x_2}^* \neq 1$. By Lemma 1, it must be the case that ${x_1}^* \in BR_{\kappa_1}({x_2}^*) \subset \{1 - {x_2}^*, \frac{1}{2}\}$. By the assumption ${x_1}^* + {x_2}^* \neq 1$, we must then have ${x_1}^* = \frac{1}{2}$ (as otherwise ${x_1}^* + {x_2}^* = 1$). However, then ${x_2}^* = BR_{\kappa_2}(\frac{1}{2}) = \frac{1}{2}$, which implies ${x_1}^* + {x_2}^* = 1$. □

4.2 Nash Equilibrium Set with Two Homogeneous Individuals

In this section, we characterize the Nash equilibrium set in a DD game with two individuals who have the same degree of morality (i.e., $\kappa_1 = \kappa_2 = \kappa \in [0,1]$). We start considering the two polar cases: i) $\kappa = 0$ and ii) $\kappa = 1$.

[6] The same reasoning follows when considering the best response of individuals with different morality levels. The only difference is in the thresholds that define the different regions.

When $\kappa = 0$, we have the case with two selfish individuals. It is well known that any strategy profile $({x_1}^*, {x_2}^*)$ satisfying ${x_1}^* + {x_2}^* = 1$ is a Nash equilibrium. Additionally, ${x_1}^* = {x_2}^* = 1$ is also a Nash equilibrium.

When $\kappa = 1$, individuals are solely concerned with maximizing $\pi(x, x)$. Thus, individuals only consider what their material payoff would be in the hypothetical case the other individual were to choose the same demand as themselves. In that case, the unique Nash equilibrium is ${x_1}^* = {x_2}^* = \frac{1}{2}$.

We now consider the case where $\kappa \in (0, 1)$, where individuals have both material and moral concerns. From Corollaries 1 and 2, we know that ${x_1}^* = {x_2}^* = \frac{1}{2}$ is the unique symmetric Nash equilibrium (as otherwise we would have ${x_1}^* + {x_2}^* \neq 1$). From Corollary 2, we know that we can focus on strategy profiles $({x_1}^*, {x_2}^*)$ with ${x_1}^* + {x_2}^* = 1$. Proposition 1 characterizes the set of Nash equilibria in the DD game with two individuals who have the same degree of morality $\kappa \in (0, 1]$.

Proposition 1: The set of Nash equilibria of the DD game between two individuals with the same degree of morality $\kappa \in (0, 1]$ consists of all pairs $({x_1}^*, {x_2}^*) \in [0, 1]^2$ such that ${x_1}^* + {x_2}^* = 1$ and either ${x_1}^* = {x_2}^* = \frac{1}{2}$ or $\frac{1}{2}\kappa \leq {x_i}^* \leq \frac{1-2\kappa}{2-2\kappa}$ for $i = 1$ or $i = 2$.

Proof. Note that for any $\kappa \in (0, 1]$, we have that ${x_1}^* = {x_2}^* = \frac{1}{2}$ is a Nash equilibrium (Corollary 1) and that any Nash equilibrium must satisfy ${x_1}^* + {x_2}^* = 1$ (Corollary 2). Finally, $({x_1}^*, {x_2}^*)$ with ${x_2}^* > \frac{1}{2}$ is a Nash equilibrium when (i) ${x_1}^* \geq \frac{1}{2}\kappa$ and (ii) $(1-\kappa){x_2}^* \geq \frac{1}{2}$.[7] As ${x_1}^* + {x_2}^* = 1$, we have that (ii) is equivalent to ${x_1}^* \leq \frac{1-2\kappa}{2-2\kappa}$, which gives the result. The analogous conditions are found when considering $({x_1}^*, {x_2}^*)$ with ${x_2}^* < \frac{1}{2}$. $\square$

[7] Where (i) and (ii) represent the conditions such that individual 1 and individual 2 do not have incentives to deviate to $\frac{1}{2}$, respectively. By Lemma 1, these deviations are the only ones we have to consider.

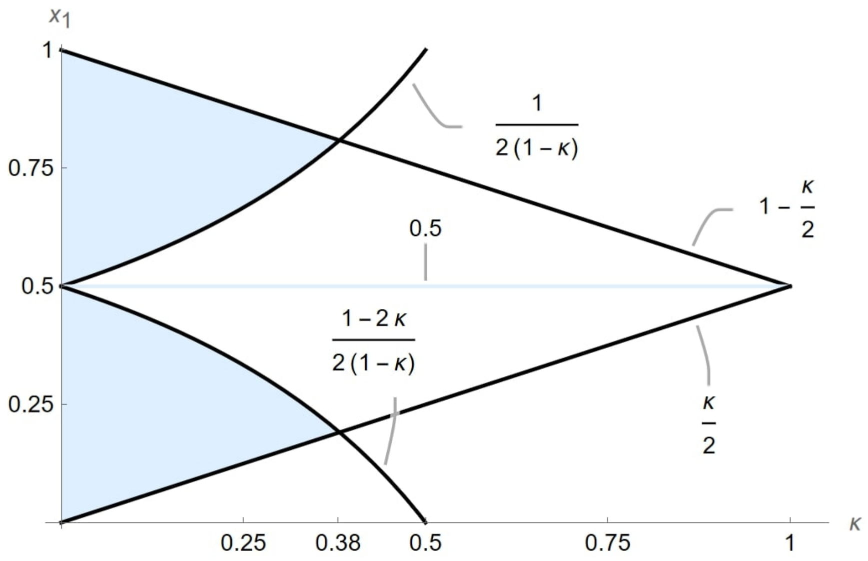

Figure 2: Nash equilibrium set $(x_1{}^*, 1-x_1{}^*)$ (in blue) in a DD game with two individuals with a common degree of morality $\kappa \in (0,1]$.

In Figure 2, we display the demands $x_1{}^*$ that belong to the Nash equilibrium set $(x_1{}^*, 1-x_1{}^*)$ in a DD game with two individuals and a common degree of morality $\kappa \in (0,1]$.

Three points are worth emphasizing. First, as shown in Corollary 1, $x_1{}^* = x_2{}^* = \frac{1}{2}$ is a Nash equilibrium regardless of κ. Second, the larger κ, the smaller the set of allocations that can be sustained as a Nash equilibrium. The shape of the Nash equilibrium set may be surprising, as for low κ, some allocations with high inequality are a Nash equilibrium while others with low inequality are not. The intuition behind this result is the following: Consider allocations (0.51, 0.49) and (0.75, 0.25). In the first allocation, individual 1 wants to deviate for any $\kappa \geq 0.02$, as the extra material payoff he receives above $\frac{1}{2}$ does not compensate him for not receiving the moral payoff. On the other hand, in the second allocation, individual 1 wants to deviate for any $\kappa \geq \frac{1}{3}$, as in this case when the individual is not too moral, this extra material payoff compensates him for not receiving the moral payoff.[8] Note that this implies that the more moral the individual, the larger the lowest demand above the equal split, such that he does not have any incentive to deviate.

Finally, there exists a unique cut-off $\underline{\kappa} \approx 0.382$ such that for any $\kappa > \underline{\kappa}$, the egalitarian division of the surplus is the unique Nash equilibrium. This $\underline{\kappa}$ is the smallest κ such that the

[8]Note that individual 2 wants to deviate from (0.51, 0.49) (resp. (0.75, 0.25)) when $\kappa \geq 0.98$ (resp. $\kappa \geq 0.50$). In both cases, the relevant constraints are the ones limiting individual 1's degree of morality.

interval $\frac{1}{2}\kappa \leq {x_i}^* \leq \frac{1-2\kappa}{2-2\kappa}$ derived in Proposition 1 collapses. Note that this cut-off is higher than what appears to be empirically plausible given the estimates of κ in Miettinen et al. (2020) and van Leeuwen and Alger (2021).[9]

4.3 Nash Equilibrium Set with Two Heterogeneous Individuals

In this section, we generalize Proposition 1 by considering a DD game with two individuals with heterogeneous degrees of morality. Proposition 2 characterizes the Nash equilibrium set for a DD game with two individuals with degrees of morality of $\kappa_1 \in [0,1]$ and $\kappa_2 \in (0,1]$, respectively.

Proposition 2: The set of Nash equilibria of the DD game between two individuals with degrees of morality $\kappa_1 \in [0,1]$ and $\kappa_2 \in (0,1]$, consists of all pairs $({x_1}^*, {x_2}^*) \in [0,1]^2$ such that ${x_1}^* + {x_2}^* = 1$ and either ${x_1}^* = {x_2}^* = \frac{1}{2}$ or $\frac{1}{2}\kappa_i \leq {x_i}^* \leq \frac{1-2\kappa_j}{2-2\kappa_j}$ for $i = 1$ and $j = 2$, or for $i = 2$ and $j = 1$.

Proof. For any $\kappa_1 \in [0,1]$ and $\kappa_2 \in (0,1]$ we have ${x_1}^* = {x_2}^* = \frac{1}{2}$ is a Nash equilibrium (Corollary 1), and that any Nash equilibrium must satisfy ${x_1}^* + {x_2}^* = 1$ (Corollary 2). Finally, $({x_1}^*, {x_2}^*)$ with ${x_2}^* > \frac{1}{2}$ is a Nash equilibrium if (i) ${x_1}^* \geq \kappa_1 \frac{1}{2}$ and (ii) $(1-\kappa_2){x_2}^* \geq \frac{1}{2}$. These two constraints represent the conditions under which individual 1 and 2 do not have incentives to deviate to $\frac{1}{2}$, respectively. By Lemma 1, these deviations are the only deviations we have to consider. As ${x_1}^* + {x_2}^* = 1$ (Corollary 2), we have that (ii) is equivalent to ${x_1}^* \leq \frac{1-2\kappa_2}{2-2\kappa_2}$, which gives the desired result. The analogous conditions can be found by considering $({x_1}^*, {x_2}^*)$ with ${x_2}^* < \frac{1}{2}$. □

Figure 3 displays the demands ${x_1}^*$ that belong to the Nash equilibrium set $({x_1}^*, 1-{x_1}^*)$ in a DD game where individual 1 is selfish (i.e., $\kappa_1 = 0$) and individual 2 is moral (i.e., $\kappa_2 \in (0,1]$). This shows us how the Nash equilibrium set changes with κ_2.

[9]Miettinen et al. (2020) and van Leeuwen and Alger (2021) use random utility models to estimate a $\kappa \approx 0$ and a κ between 0.1 and 0.2, respectively.

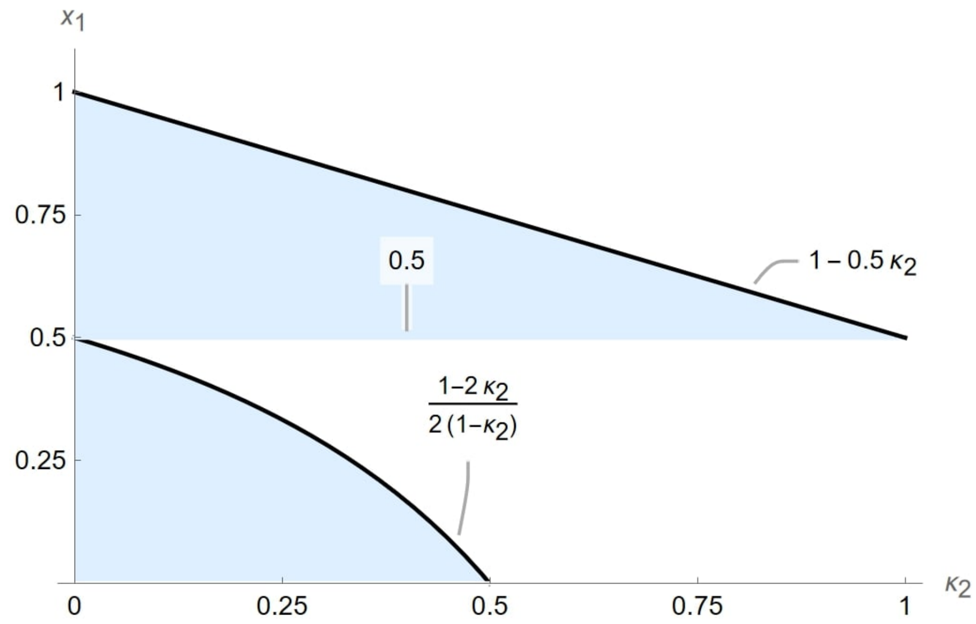

Figure 3: Nash equilibrium Set $(x_1{}^*, 1 - x_1{}^*)$ (in blue) in a DD game with a selfish and a moral individual with $\kappa_2 \in (0, 1]$.

As before, the larger κ_2, the smaller the set of allocations that can be sustained as a Nash equilibrium. Additionally, note that if $\kappa_2 > \frac{1}{2}$, then any $(x_1{}^*, x_2{}^*)$ satisfies $x_1{}^* \geq x_2{}^*$, which implies that individual 1 receives at least half of the surplus.

4.4 Selfish vs Moral Individual: Potential Equilibrium Outcomes

In Section 4.3, we characterized the equilibrium set for any profile of individuals' degrees of morality. In this section, we explore more deeply the bargaining interaction between a selfish and a moral individual by studying the potential equilibrium outcomes of the selfish individual when bargaining with a moral individual. We consider a DD game where individual 1 has $\kappa_1 = 0$ and individual 2 has $\kappa_2 \in (0, 1]$. Figure 3 displays the demands $x_1{}^*$ that belong to the Nash equilibrium set $(x_1{}^*, 1 - x_1{}^*)$ of this interaction.

Since the game admits multiple equilibria, we study how the best and worst equilibrium of the selfish individual changes with κ_2. Note that this characterizes the bounds of the selfish individual's utility when bargaining with a moral individual. In the presence of multiplicity, focusing on extreme equilibria (i.e., best and worst) is a standard approach (see Kim, 2003; Rocheteau and Wright, 2005; Jackson and Pernoud, 2020; and Houba et al., 2022, among others).

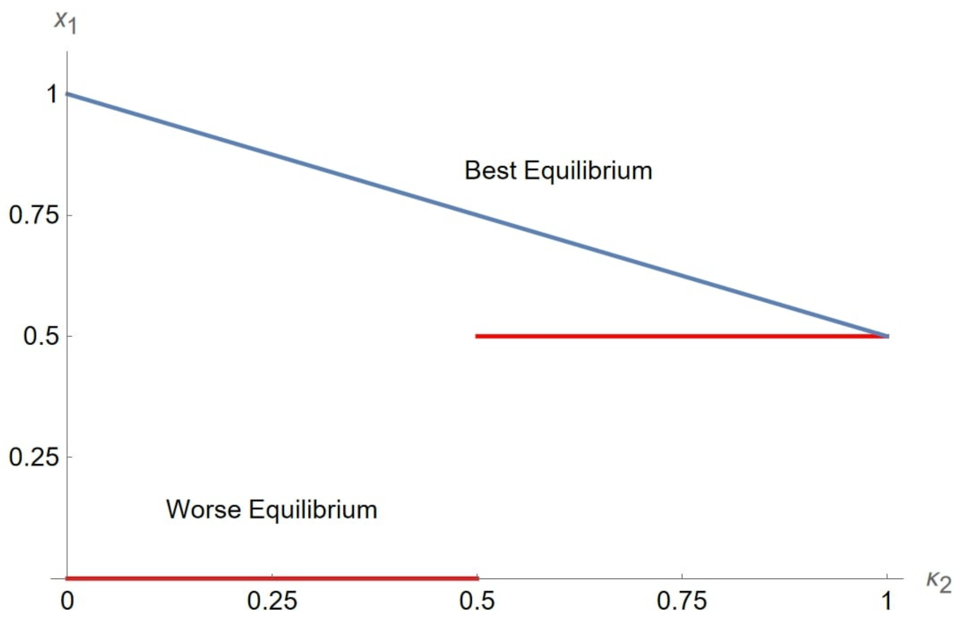

Figure 4: Best (in dark blue) and worst (in red) equilibrium of the individual 1 when bargaining with a moral individual with a degree of morality κ_2.

Figure 4 characterizes the best and worst equilibrium for the selfish individual when bargaining against a moral individual with a degree of morality κ_2. Two remarks are important to emphasize. First, the best equilibrium of the selfish individual is strictly decreasing in κ_2. Thus, the largest utility that a selfish individual can obtain when bargaining with a moral individual is negatively related to his opponent's degree of morality. Second, the worst equilibrium of a selfish individual is equal to zero when $\kappa_2 < \frac{1}{2}$ and equal to the equal split when $\kappa_2 > \frac{1}{2}$.

To sum up, we show that κ_2 significantly affects the best and worst equilibrium outcomes of the selfish individual. More precisely, facing a more moral individual is worse for the selfish individual if we focus on the best equilibrium outcome, whereas it is better if we focus on the worst equilibrium outcome. It is interesting to see that an increase in the morality of his opponent has two opposing effects on the well-being of the selfish individual: a limitation on the best equilibrium outcome on one hand and an improvement on the worst equilibrium outcome on the other. From a theoretical perspective, it is challenging to determine whether selfish individuals would prefer to bargain with selfish or moral partners, as this would require making additional assumptions on the equilibrium selection. We view this as a promising avenue for future empirical research. Questions examining whether individuals' degrees of morality influence their bargaining ability, choice of preferred bargaining partner, or willingness to enter bargaining scenarios carry significant potential for exploration.

5 Incomplete Information: Society's Composition and Equilibrium Outcomes

In this section, we study how the composition of society affects the equilibrium distribution of the surplus. We consider the following setting: the society is composed of an infinite number of individuals who are divided into two different types, only differing in their degree of morality. More concretely, a share $p \in (0,1)$ of the individuals belongs to the first type (with $\kappa = \underline{\kappa} \in [0,1)$), and the remainder $1-p$ belongs to the second type (with $\kappa = \overline{\kappa} > \underline{\kappa}$). The values of p, $\underline{\kappa}$, and $\overline{\kappa}$ are common knowledge. Two individuals are randomly drawn from the population and matched with each other to play in a DD game. An individual's degree of morality is his private information. Thus, when individuals are facing their opponent in the DD game, they know their type but do not know their opponent's type.

This defines a Bayesian game, $G = (N, (S_i)_{i\in N}, (\Theta_i)_{i\in N}, (\mu_i)_{i\in N}, (u_i)_{i\in N})$, where $N = \{1,2\}$ is the set of players, $S_i = [0,1] \times [0,1]$ is i's strategy set, $\Theta_i = \{\underline{\kappa}, \overline{\kappa}\}$ is the set of types, and $\mu_i \in \Delta(\Theta_{-i})$, where $\mu_i(\underline{\kappa}) = p$ and $\mu_i(\overline{\kappa}) = 1-p$, represents i's probabilities of facing each type, for all $i \in N$. We study how the set of Bayesian Nash equilibria (BNE) depends on p, $\underline{\kappa}$, and $\overline{\kappa}$. For simplicity, we restrict attention to equilibria in symmetric, pure strategies, where all individuals with the same type choose the same strategy. A strategy for individual i is a mapping $s_i : \Theta_i \to S_i$, prescribing an action for each possible type of individual i. Thus, for any $i = \{1,2\}$, let $s_i = (s_{i,\underline{\kappa}}, s_{i,\overline{\kappa}}) \in [0,1]^2$ denote individual i's strategy, where s_{i,θ_i} is the action that individual i of type θ_i plays. Accordingly, $s = (s_1, s_2) \in [0,1]^2 \times [0,1]^2$ denotes the strategy profile. Finally, the utility function of individual i, $u_i : S \times \Theta \to \mathbb{R}$, is defined as $u_i(s_i, s_j, \theta_i, \theta_j) = (1-\theta_i)\pi(s_{i,\theta_i}, s_{j,\theta_j}) + \theta_i \pi(s_{i,\theta_i}, s_{i,\theta_i})$, for $i \neq j$ and θ_i, $\theta_j \in \{\underline{\kappa}, \overline{\kappa}\}$. Definition 1 presents the equilibrium concept.

Definition 1: $s^* = ({s_{\underline{\kappa}}}^*, {s_{\overline{\kappa}}}^*) \in [0,1]^2$ is a (symmetric) Bayesian Nash equilibrium strategy if the following conditions are satisfied:

$$
\begin{aligned}
{s_{\underline{\kappa}}}^* &\in \mathrm{argmax}_{x\in[0,1]}(1-\underline{\kappa})(p\pi(x, {s_{\underline{\kappa}}}^*) + (1-p)\pi(x, {s_{\overline{\kappa}}}^*)) + \underline{\kappa}\pi(x,x), \\
{s_{\overline{\kappa}}}^* &\in \mathrm{argmax}_{x\in[0,1]}(1-\overline{\kappa})(p\pi(x, {s_{\underline{\kappa}}}^*) + (1-p)\pi(x, {s_{\overline{\kappa}}}^*)) + \overline{\kappa}\pi(x,x)
\end{aligned}
$$

Note that since we focus on symmetric equilibria, we dropped the i (or j) subscript in the definition in favor of a simpler notation. Furthermore, as the Kantian counterfactual does not depend on the opponent's type, it simply appears as $\theta_i \pi(x, x)$, with $\theta_i \in \{\underline{\kappa}, \overline{\kappa}\}$.

Lemmas 2 and 3 below will be useful in proving Propositions 3 and 4. In particular, Lemma 2 guarantees the existence of an equilibrium, while Lemma 3 allows us to restrict the strategy profiles to consider when deriving the equilibrium set.

Lemma 2: The strategy $s^*_{\underline{\kappa}} = s^*_{\overline{\kappa}} = \frac{1}{2}$ is a (symmetric) Bayesian Nash equilibrium strategy of the Bayesian game G for any $\underline{\kappa} \in [0, 1)$, $\overline{\kappa} > \underline{\kappa}$, and $p \in (0, 1)$.

Lemma 3: Let $(s^*_{\underline{\kappa}}, s^*_{\overline{\kappa}}) \in [0, 1]^2$ be a (symmetric) Bayesian Nash equilibrium strategy of the Bayesian game G. Then, $s^*_{\underline{\kappa}} + s^*_{\overline{\kappa}} = 1$.

The proofs of Lemmas 2 and 3 can be found in the Appendix. Proposition 3 shows that it is not possible to have a (symmetric) BNE strategy where the more moral individuals obtain a higher surplus than the less moral ones.

Proposition 3: The set of (symmetric) Bayesian Nash equilibrium strategies of the Bayesian game G consists of strategies $({s_{\underline{\kappa}}}^*, {s_{\overline{\kappa}}}^*) \in [0, 1]^2$ such that ${s_{\underline{\kappa}}}^* + {s_{\overline{\kappa}}}^* = 1$ and either ${s_{\underline{\kappa}}}^* = {s_{\overline{\kappa}}}^* = \frac{1}{2}$ or ${s_{\underline{\kappa}}}^* > {s_{\overline{\kappa}}}^*$.

Proof. By Lemma 3, we restrict attention to strategies $({s_{\underline{\kappa}}}^*, {s_{\overline{\kappa}}}^*)$ with ${s_{\underline{\kappa}}}^* + {s_{\overline{\kappa}}}^* = 1$. By contradiction, suppose that there exists $({s_{\underline{\kappa}}}^*, {s_{\overline{\kappa}}}^*)$ with ${s_{\underline{\kappa}}}^* + {s_{\overline{\kappa}}}^* = 1$ and ${s_{\overline{\kappa}}}^* > {s_{\underline{\kappa}}}^*$. Then, ${s_{\overline{\kappa}}}^* > \frac{1}{2}$ and ${s_{\underline{\kappa}}}^* < \frac{1}{2}$. When both individuals choose this strategy, an individual with $\kappa = \underline{\kappa}$ obtains an expected utility of $1 - {s_{\overline{\kappa}}}^*$, while an individual with $\kappa = \overline{\kappa}$ receives an expected utility of $p(1-\overline{\kappa}){s_{\overline{\kappa}}}^*$. For $({s_{\underline{\kappa}}}^*, {s_{\overline{\kappa}}}^*)$ to be a (symmetric) BNE strategy, we need to determine the conditions such that the individuals do not have incentives to deviate. An individual with $\kappa = \underline{\kappa}$ may deviate to $({s_{\overline{\kappa}}}^*, {s_{\overline{\kappa}}}^*)$, which would give him an expected utility of $p(1 - \underline{\kappa}){s_{\overline{\kappa}}}^*$. Therefore, a necessary condition for $({s_{\underline{\kappa}}}^*, {s_{\overline{\kappa}}}^*)$ to be a (symmetric) BNE strategy is ${s_{\overline{\kappa}}}^* \leq \frac{1}{1+p(1-\underline{\kappa})}$. On the other hand, an individual with $\kappa = \overline{\kappa}$ may deviate to $({s_{\underline{\kappa}}}^*, {s_{\underline{\kappa}}}^*)$ which would give him an expected utility of ${s_{\underline{\kappa}}}^*$. Therefore, a necessary condition for $({s_{\underline{\kappa}}}^*, {s_{\overline{\kappa}}}^*)$ to be a (symmetric) BNE strategy is ${s_{\overline{\kappa}}}^* \geq \frac{1}{1+p(1-\overline{\kappa})}$. However, note that $\overline{\kappa} > \underline{\kappa}$ implies that $\frac{1}{1+p(1-\overline{\kappa})} > \frac{1}{1+p(1-\underline{\kappa})}$ for any $p > 0$. Therefore, when $\overline{\kappa} > \underline{\kappa}$, the two previous conditions are mutually exclusive, and therefore, it is not possible to have a (symmetric) BNE strategy with ${s_{\overline{\kappa}}}^* > {s_{\underline{\kappa}}}^*$. $\square$

Note that Proposition 3 implies that regardless of $\overline{\kappa}$ and p, there is no equilibrium where selfish individuals obtain less than half of the surplus. However, is it possible to have a BNE where they obtain more than half of the surplus? If so, how does this depend on p and $\overline{\kappa}$? Proposition 4 characterizes the set of (symmetric) BNE strategies when $\underline{\kappa} = 0$. Note that in this case, we distinguish two cases depending on the values of $\overline{\kappa}$ and p.

Proposition 4: The set of (symmetric) Bayesian Nash equilibrium strategies of the Bayesian game G, with $\Theta_1 = \Theta_2 = \{0, \overline{\kappa}\}$, consists of all strategies $({s_0}^*, {s_{\overline{\kappa}}}^*) \in [0,1]^2$ with ${s_0}^* + {s_{\overline{\kappa}}}^* = 1$ such that:

- **Case 1**: If $\overline{\kappa} > \frac{1}{2}$, then ${s_0}^* = {s_{\overline{\kappa}}}^* = \frac{1}{2}$.
- **Case 2**: If $\overline{\kappa} < \frac{1}{2}$ and $p \in (\frac{1-2\overline{\kappa}}{1-\overline{\kappa}}, 1]$, then ${s_0}^* = {s_{\overline{\kappa}}}^* = \frac{1}{2}$.
- **Case 3**: If $\overline{\kappa} \leq \frac{1}{2}$ and $p \in (0, \frac{1-2\overline{\kappa}}{1-\overline{\kappa}}]$, then either ${s_0}^* = {s_{\overline{\kappa}}}^* = \frac{1}{2}$ or ${s_0}^* \in [\frac{1}{2-p}, \frac{1}{2} + \frac{p}{2}(1-\overline{\kappa})]$.

Proof. From Proposition 3, we know that we can focus on $({s_0}^*, {s_{\overline{\kappa}}}^*)$ with ${s_0}^* \geq {s_{\overline{\kappa}}}^*$. By Lemma 2, we know that ${s_0}^* = {s_{\overline{\kappa}}}^* = \frac{1}{2}$ is a (symmetric) BNE strategy for any $\underline{\kappa}$, $\overline{\kappa}$ and p. Additionally, by Lemma 3, we can focus on (symmetric) strategies with ${s_0}^* + {s_{\overline{\kappa}}}^* = 1$. Hence, any (symmetric) BNE strategy with ${s_0}^* > {s_{\overline{\kappa}}}^*$ must satisfy ${s_0}^* > \frac{1}{2}$ and ${s_{\overline{\kappa}}}^* < \frac{1}{2}$. When both individuals select strategy $({s_0}^*, {s_{\overline{\kappa}}}^*)$, an individual with $\kappa = 0$ gets an expected utility of $(1-p){s_0}^*$, while an individual with $\kappa = \overline{\kappa}$ gets an expected utility of $1 - {s_0}^*$. For $({s_0}^*, {s_{\overline{\kappa}}}^*)$ to be a (symmetric) BNE strategy, we need to determine the conditions such that individuals do not have any incentive to deviate.

The best deviation of an individual with $\kappa = 0$ is $({s_{\overline{\kappa}}}^*, {s_{\overline{\kappa}}}^*)$, which would give him an expected utility of $1 - {s_0}^*$.[10] Then, a necessary condition for $({s_0}^*, {s_{\overline{\kappa}}}^*)$ to be a (symmetric) BNE strategy is that ${s_0}^* \geq \frac{1}{2-p}$. On the other hand, the best deviation of an individual with $\kappa = \overline{\kappa}$ is $({s_0}^*, \frac{1}{2})$, which would give him an expected utility of $(1-\overline{\kappa})(1-p)\frac{1}{2} + \frac{1}{2}\overline{\kappa}$. Then, a necessary condition for $({s_0}^*, {s_{\overline{\kappa}}}^*)$ to be a (symmetric) BNE strategy is that ${s_0}^* \leq \frac{1}{2} + \frac{1}{2}p(1-\overline{\kappa})$. Therefore, for $({s_0}^*, {s_{\overline{\kappa}}}^*)$ to be a (symmetric) BNE strategy we need ${s_0}^* \geq \frac{1}{2-p}$ and ${s_0}^* \leq \frac{1}{2} + \frac{1}{2}p(1-\overline{\kappa})$.

[10] It is never profitable for a selfish individual to deviate to $x \in ({s_{\overline{\kappa}}}^*, {s_0}^*)$, as this does not bring him an additional payoff when bargaining with a selfish individual (as $x + {s_0}^* > 1$) and decreases his payoff when bargaining with a moral individual (as $x < {s_0}^*$).

Then, to determine which strategies are not a (symmetric) BNE strategy, we need to determine the values of $\overline{\kappa}$ and p such that $\frac{1}{2-p} > \frac{1}{2} + \frac{1}{2}p(1-\overline{\kappa})$. Note that when $\overline{\kappa} > \frac{1}{2}$ we have that $\frac{1}{2-p} > \frac{1}{2} + \frac{1}{2}p(1-\overline{\kappa})$ for any $p \in (0,1)$. On the other hand, when $\overline{\kappa} \leq \frac{1}{2}$, then $\frac{1}{2-p} > \frac{1}{2} + \frac{1}{2}p(1-\overline{\kappa})$ when $p > \frac{1-2\overline{\kappa}}{1-\overline{\kappa}}$. In sum, (i) $\overline{\kappa} > \frac{1}{2}$ and (ii) $\overline{\kappa} < \frac{1}{2}$ and $p > \frac{1-2\overline{\kappa}}{1-\overline{\kappa}}$ are the conditions such that there is no (symmetric) BNE strategy with ${s_0}^* > {s_{\overline{\kappa}}}^*$. By Lemma 2 and Proposition 3, ${s_0}^* = {s_{\overline{\kappa}}}^* = \frac{1}{2}$ is the unique (symmetric) BNE strategy.

On the other hand, when $\overline{\kappa} \leq \frac{1}{2}$ and $p \in (0, \frac{1-2\overline{\kappa}}{1-\overline{\kappa}}]$, then $\frac{1}{2-p} < \frac{1}{2} + \frac{1}{2}p(1-\overline{\kappa})$ for any $\overline{\kappa}$ and p, implying that any strategy profile $({s_0}^*, {s_{\overline{\kappa}}}^*)$ with ${s_0}^* \in [\frac{1}{2-p}, \frac{1}{2} + \frac{1}{2}p(1-\overline{\kappa})]$ and ${s_{\overline{\kappa}}}^* = 1 - {s_0}^*$ is a (symmetric) BNE strategy. $\square$

Note that p and $\overline{\kappa}$ make it difficult for an equilibrium where selfish individuals receive more than half of the surplus to exist. On the one hand, in an equilibrium with ${s_0}^* > {s_{\overline{\kappa}}}^*$, the selfish individual gets a material payoff of zero with probability p (i.e., the likelihood that he is facing another selfish individual). Then, the larger p, the more likely selfish individuals are to get a material payoff of zero, and therefore, the more incentives for them to deviate. On the other hand, as in the complete information case, the larger $\overline{\kappa}$, the larger is the lowest ${s_{\overline{\kappa}}}^*$ such that individuals with $\kappa = \overline{\kappa}$ do not have incentives to deviate to demand $\frac{1}{2}$.

Proposition 4 shows two things: First, if individuals with $\kappa = \overline{\kappa}$ are sufficiently moral or if their share in the population is sufficiently low, then the equal division of the surplus is the unique possible equilibrium outcome. Second, when individuals with $\kappa = \overline{\kappa}$ are not sufficiently moral and their share in the population is sufficiently high, then it is possible for selfish individuals to obtain more than half of the surplus.

6 Comparisons with Alternative Behavioral Models

In this section, we derive the equilibrium set under alternative preferences. We consider inequity aversion (Fehr and Schmidt, 1999), (unconditional) altruism (Becker, 1953), and fairness preferences (Birkeland and Tungodden, 2014). Additionally, we consider the predictions of the Kantian equilibrium (Roemer, 2010). In comparison to the previously mentioned models, homo moralis is unique in satisfying simultaneously the two following properties: (i) all the equilibrium outcomes are on the bargaining frontier, and (ii) changes in the preference parameters refine the equilibrium set.

6.1 Inequity Averse Preferences

We now characterize the equilibrium set when individuals have inequity averse preferences (Fehr and Schmidt, 1999) and compare it with the equilibrium set derived with homo moralis. The main difference is that with inequity averse preferences, there may exist equilibrium outcomes that are not on the bargaining frontier (i.e., ${x_1}^* + {x_2}^* > 1$). We consider the $N = 2$ case:

$$u_i(x_i, x_j) = \pi(x_i, x_j) - \alpha_i max[\pi(x_j, x_i) - \pi(x_i, x_j), 0] - \beta_i max[\pi(x_i, x_j) - \pi(x_j, x_i), 0],$$

where $\beta_i \in [0, 1)$ represents individual i's degree of *aheadness aversion*, $\alpha_i \geq \beta_i$ represents individual i's degree of *behindness aversion*, and $\pi(x_i, x_j)$ (resp. $\pi(x_j, x_i)$) represents individual i's (resp. j's) material payoff under strategy profile (x_i, x_j).

We first show that, with inequity averse preferences, it is not possible to restrict attention to allocations on the bargaining frontier. To do so, we characterize the Nash equilibrium set satisfying ${x_1}^* + {x_2}^* > 1$ when $\alpha_i = \alpha_j = \alpha$ and $\beta_i = \beta_j = \beta$.

Proposition 5: Let $\beta \in [0, 1)$ and $\alpha \geq \beta$ be individuals' degrees of aheadness and behindness aversion, respectively. Then, any strategy profile $({x_1}^*, {x_2}^*)$ with ${x_i}^* \in [\frac{1+\alpha}{1+2\alpha}, 1]$ for any $i \in \{1, 2\}$ is a Nash equilibrium.

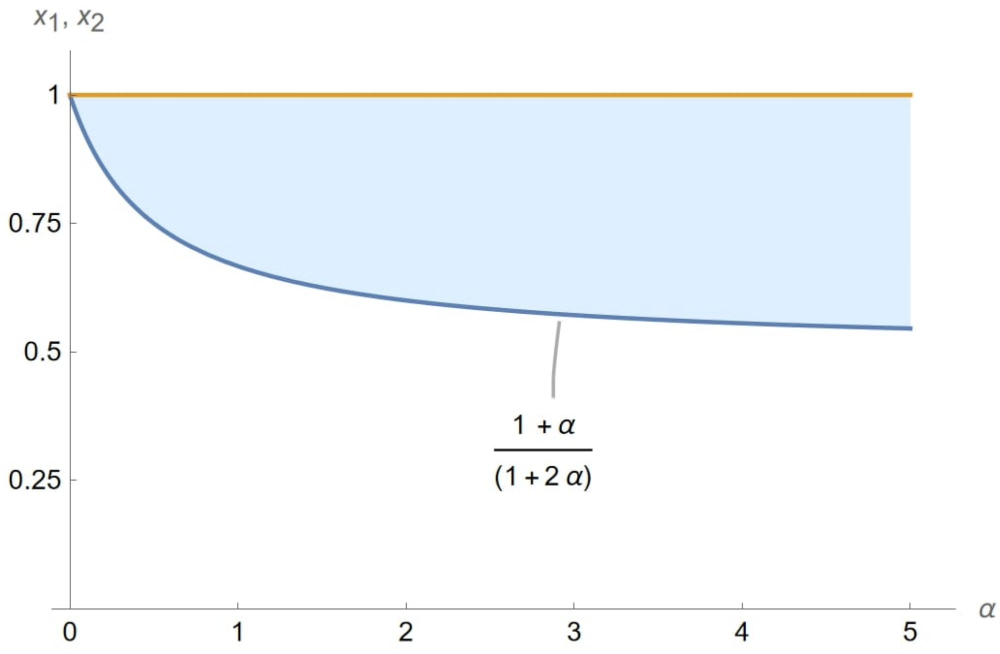

Figure 5: Nash equilibrium set with $x_1 + x_2 > 1$, $\beta = 0$ and $\alpha \geq 0$.

All proofs of this section are in the Appendix. Intuitively, disagreement equilibria (i.e., equilibria with ${x_1}^* + {x_2}^* > 1$) occur when individuals are not willing to choose a lower demand as the associated material gain is lower than the disutility of falling behind. Figure 5 displays the Nash equilibrium set $(x_1{}^*, x_2{}^*)$ satisfying $x_1{}^* + x_2{}^* > 1$ for $\beta = 0$ and $\alpha \geq 0$.

We now derive the Nash equilibrium set satisfying $x_1{}^* + x_2{}^* = 1$. We first note that $x_1{}^* = x_2{}^* = \frac{1}{2}$ is a Nash equilibrium for any (α, β).[11] Proposition 6 characterizes the equilibrium set of a DD game with two inequity averse individuals.

Proposition 6: The set of "agreement" Nash equilibria of the DD game between two individuals with the degrees of aheadness aversion $\beta_1 \in [0, 1)$ and $\beta_2 \in [0, 1)$, and degrees of behindness aversion $\alpha_1 > \beta_1$ and $\alpha_2 > \beta_2$, consists of all pairs $(x_1{}^*, x_2{}^*) \in [0, 1]^2$ such that $x_1{}^* + x_2{}^* = 1$ and either $x_1{}^* = x_2{}^* = \frac{1}{2}$ or the two following conditions hold:

- $x_i{}^* - \alpha_i \max[x_j{}^* - x_i{}^*, 0] \geq 0$
- $x_j{}^* - \beta_j \max[x_j{}^* - x_i{}^*, 0] \geq 0$

for $i = 1$ and $j = 2$, or for $i = 2$ and $j = 1$.

Figure 6 displays the Nash equilibrium allocations $(x_1{}^*, 1 - x_1{}^*)$ when the two individuals have $\alpha = \beta \in [0, 1)$.

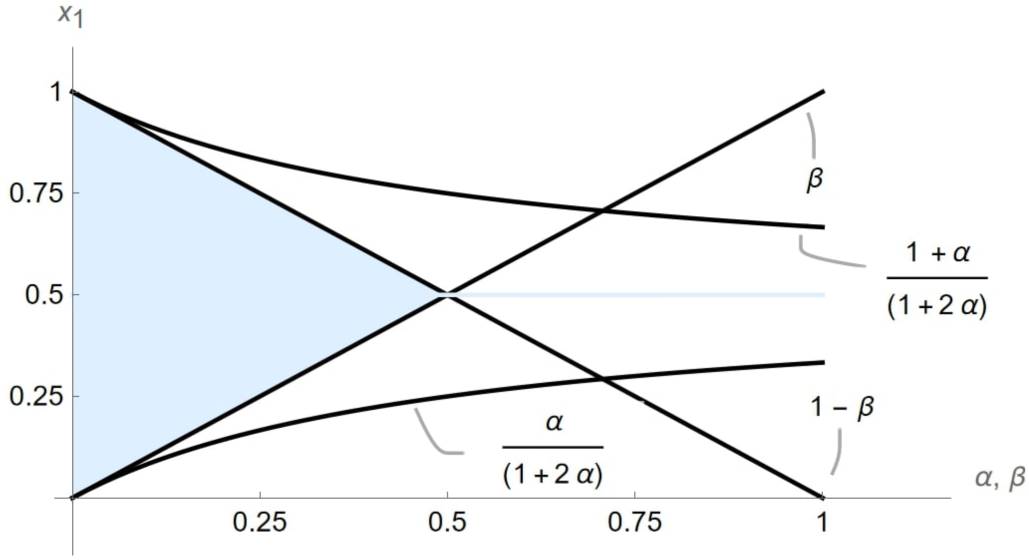

Figure 6: Equilibrium set (in blue) with $x_1{}^* + x_2{}^* = 1$ and $\alpha = \beta \in [0, 1)$.

Note that when $\alpha = \beta \in [0, 1)$, the aheadness aversion constraints are the ones defining the equilibrium set. In this specification, when both individuals have $\beta > \frac{1}{2}$, the egalitarian

[11]When individual j chooses $x_j = \frac{1}{2}$, individual i maximizes all three terms of his utility function by choosing $x_i = \frac{1}{2}$.

allocation is the unique Nash equilibrium. In the online supplementary material, we display the Nash equilibrium set when individuals are only behindness averse (i.e., $\beta = 0$ and $\alpha > 0$), which displays similar qualitative patterns.

6.2 Altruistic Preferences

We now characterize the Nash equilibrium set when individuals have altruistic preferences and compare it with the equilibrium set derived with homo moralis. The main difference is that with altruistic preferences, the degree of altruism does not refine the Nash equilibrium set. For simplicity, we consider the $N = 2$ case where individuals have linear (unconditional) altruistic preferences:

$$u_i(x_i, x_j) = \pi(x_i, x_j) + \alpha_i \pi(x_j, x_i),$$

where $\alpha_i \in [0, 1]$ represents individual i's degree of altruism, the weight individual i attaches to individual j's material payoff (i.e., $\pi(x_j, x_i)$).

Proposition 7: The set of Nash equilibria of the DD game between two individuals with degrees of altruism $\alpha_1 \in (0, 1]$ and $\alpha_2 \in (0, 1]$, consists of all pairs $({x_1}^*, {x_2}^*) \in [0, 1]^2$ such that ${x_1}^* + {x_2}^* = 1$.

Proposition 7 characterizes the equilibrium set of a DD game with two altruistic individuals. With altruistic preferences, any Nash equilibrium is on the bargaining frontier, but the equilibrium set does not change with individuals' degrees of altruism.

6.3 Fairness Preferences

We now discuss the differences between the predictions of a model with fairness concerns and homo moralis preferences. To do so, we consider the functional form introduced in Birkeland and Tungodden (2014):

$$u_i(x_i, x_j, s_i) = \begin{cases} x_i - \beta_i (x_i - s_i)^2 & \text{if } x_i + x_j \leq 1 \\ 0 & \text{if } x_i + x_j > 1 \end{cases}$$

where s_i represents individual i's fairness ideal and $\beta_i > 0$ represents the (relative) weight individual i attaches to fairness. An individual with fairness concerns (i.e., $\beta_i > 0$) suffers a disutility when, conditional on $x_i + x_j \leq 1$, the $|x_i - s_i|$ increases.

Homo moralis and fairness preferences have a similar structure. In both cases, individuals maximize their material payoff but also attach a weight to doing the right thing and fairness concerns. However, while homo moralis has a unique (exogenous) parameter (i.e., κ_i), fairness preferences have two (i.e., β_i and s_i). Birkeland and Tungodden (2014) show that when two individuals hold fairness views that are incompatible (i.e., $s_1+s_2 > 1$), they may not necessarily reach an agreement. This is never the case with homo moralis preferences.

We argue that homo moralis can be interpreted as a type of fairness preference, with the fairness ideal being endogenously determined based on the structure of the interaction. More precisely, the *Kantian fairness ideal* can be defined as $s_{HM} \equiv \text{argmax}_{x\in[0,1]}\, \pi(x, x)$, which leads to $s_{HM} = \frac{1}{N}$ in the N player DD game. Accordingly, *Kantian fairness* preferences can be defined simply by replacing s_i with s_{HM} in the utility function above.

This has the main advantage of reducing the model's degrees of freedom and, therefore, disciplining the model's predictions. This is exemplified in the online supplementary material, where we derive the equilibrium set for the ND game. While with fairness preferences, it would be necessary to (exogenously) specify a fairness ideal for each individual, with homo moralis preferences, the shape of the bargaining frontier endogenously determines a shared fairness ideal.

6.4 Selfish Preferences under Nash and Kantian Equilibrium

Finally, we consider the equilibrium set when individuals have selfish preferences under (i) Nash equilibrium and (ii) Kantian Equilibrium.[12] In both cases, the equilibrium set is the same: any strategy profile where the sum of demands is equal to one is a Nash (Kantian) equilibrium. Moreover, ${x_1}^* = {x_2}^* = 1$ is also a Nash (Kantian) equilibrium. Thus, the set of Kantian equilibria does not refine the set of Nash equilibria.

6.5 Summary Comparison

We have shown that the predictions from homo moralis preferences differ from the predictions of the alternative preferences and equilibrium concepts above. The set of Nash equilibria with homo moralis preferences has the following properties: (i) any Nash equilibrium must be on the

[12]A strategy profile is a Kantian equilibrium if no individual would prefer everybody to change their demand by the same non-negative factor (Roemer, 2010).

bargaining frontier, (ii) morality refines the Nash equilibria set, and (iii) the egalitarian outcome is the unique Nash equilibrium when individuals are sufficiently moral. This differs from the predictions of the alternative models in the following ways: When individuals are inequity averse there may exist Nash equilibria not in the bargaining frontier. When individuals are altruistic, altruism does not refine the set of Nash equilibria. When considering the Kantian equilibrium, the equilibrium set is identical to the one with selfish preferences. Finally, the model with fairness preferences admits disagreement equilibria.[13] Additionally, we argue that homo moralis preferences endogenize the fairness ideal of a general fairness preference, reducing the model's number of free parameters.

We have conducted a comparison of various behavioral models on the basis of their theoretical predictions. Another natural promising route is to test them against each other using controlled experiments. Such an experiment should elicit participants' moral, fairness, and inequity aversion preferences using incentivized tasks and/or questionnaires. One could then use these measures to estimate relevant parameter values, which would be finally used to explain their behavior in bargaining interactions. Such experiments can also be used to gain insights into the stability of moral preferences across various contexts and types of strategic interactions. Bolton and Ockenfels (2005), Miettinen et al. (2020), van Leeuwen and Alger (2021), and Capraro and Rodriguez-Lara (2022) are some examples that followed the aforementioned route.

7 Concluding Remarks

We studied the Nash equilibrium set of the DD game when some individuals have moral preferences. To the best of our knowledge, this is the first paper to incorporate moral preferences into

[13] It is fair to say that predicting disagreement for some parameter combinations is reasonable. For instance, Birkeland and Tungodden (2014) predicted and reported that disagreements are more likely when fairness preferences are not mutually compatible. Similarly, Gächter and Riedl (2005), Karagözoğlu and Riedl (2015), and Embrey, Hyndman, and Riedl (2021) reported a positive correlation between the frequency of disagreements and conflict between individuals' fairness judgments. Depending on the particular experimental design, disagreement rates in unstructured bargaining experiments where data on fairness preferences/judgments were collected vary between 3 percent and 31 percent. The existence of disagreement outcomes in experiments likely imply that finite mixture models will be useful in capturing individuals' preferences.

a bargaining game. We provided a set of variations, comparative static analyses, and comparisons with alternative models with various behavioral assumptions on individual preferences.

We believe that our paper lays the groundwork for the incorporation of moral preferences in bargaining games. It will be a useful resource for experimental researchers who aim to understand bargaining behavior and identify the types of individual preferences at work in such interactions. Along these lines, we think that a natural future research agenda is conducting experiments in a bargaining context. On the theoretical side, we believe that evolutionary bargaining models with homo oeconomicus and homo kantiensis types (similar to works by Young, 1993; Skyrms, 1996; Robles, 2008; Andreozzi, 2010) will be natural venues for future research. Finally, the incorporation of moral preferences into dynamic bargaining interactions is a question that theoretical researchers will possibly address in the coming years.

Acknowledgments: *The authors would like to thank the editor and three anonymous referees for helpful comments and suggestions. They would also like to thank Ingela Alger, Kerim Keskin, Esteban Muñoz, and Enrico-Mattia Salonia for fruitful discussions and comments, and Tara Sylvestre for language editing. Pau Juan-Bartroli acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 789111 - ERC EvolvingEconomics). The usual disclaimers apply.*

References

[1] Alger I, Laslier J-F (2021). Homo moralis goes to the voting booth: a new theory of voter participation. *TSE Working Paper*, **21-1193**.

[2] Alger I, Laslier J-F (2022). Homo moralis goes to the voting booth: coordination and information aggregation. *Journal of Theoretical Politics*, **34**, 280-312.

[3] Alger I, Weibull JW (2013). Homo moralis - Preference evolution under incomplete information and assortative matching. *Econometrica*, **81**, 2269-2302.

[4] Alger I, Weibull JW (2016). Evolution and Kantian morality. *Games and Economic Behavior*, **98**, 56-67.

[5] Alger I, Weibull JW (2017). Strategic behavior of moralists and altruists. *Games*, **8**, 38.

[6] Andreozzi L (2010). An evolutionary theory of social justice: Choosing the right game. *European Journal of Political Economy*, **26**, 320-329.

[7] Ashlagi I, Karagözoğlu E, Klaus B-E (2012). A non-cooperative support for equal division in estate division problems. *Mathematical Social Sciences*, **63**, 228-233.

[8] Birkeland S, Tungodden B (2014). Fairness motivation in bargaining: a matter of principle. *Theory and Decision*, **77**, 125-151.

[9] Bolton GE, Karagözoğlu E (2016). On the influence of hard leverage in a soft leverage bargaining game: the importance of credible claims. *Games and Economic Behavior*, **99**, 164–179.

[10] Bolton GE, Ockenfels A (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, **90**, 166-193.

[11] Bolton GE, Ockenfels A (2005). A stress test of fairness measures in models of social utility. *Economic Theory*, **25**, 957-982.

[12] Bomze I, Schachinger W, Weibull J (2021). Does moral play equilibriate? *Economic Theory*, **71**, 305-315.

[13] Camerer CF, Nave G, Smith A (2019). Dynamic unstructured bargaining with private information: Theory, experiment, and outcome prediction via machine learning. *Management Science*, **65**, 1867-1890.

[14] Capraro V, Rodriguez-Lara I (2022). Moral preferences in ultimatum and impunity games. *Working Paper.*

[15] Dizarlar A, Karagözoğlu E (2023). Kantian equilibria of a class of Nash bargaining games. *Journal of Public Economic Theory*, **25**, 867-891.

[16] Driesen B, Perea A, Peters H (2012). Alternating offers bargaining with loss aversion. *Mathematical Social Sciences*, **64**, 103–118.

[17] Embrey M, Frechette GR, Lehrer SF (2015). Bargaining and reputation: An experiment on bargaining in the presence of behavioural types. *Review of Economic Studies*, 82, 608-631.

[18] Embrey M, Hyndman K, Riedl A (2021). Bargaining with a residual claimant: An experimental study. *Games and Economic Behavior*, **126**, 335-354.

[19] Engel C (2011). Dictator games: a meta study. *Experimental Economics*, **14**, 583-610.

[20] Engler Y, Page L (2022). Driving a hard bargain is a balancing act: how social preferences constrain the negotiation process. *Theory and Decision*, **93**, 7-36.

[21] Fehr E, Schmidt K (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, **114**, 817-868.

[22] Galinsky AD, Müssweiler T (2001). First offers as anchors: the role of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology*, **87**, 657-669.

[23] Gächter S, Riedl A (2005). Moral property rights in bargaining with infeasible claims. *Management Science*, **51**, 249–263.

[24] Guha B (2018). Malice in the Rubinstein bargaining game. *Mathematical Social Sciences*, **94**, 82-86.

[25] Güth W, Schmittberger R, Schwarze B (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, **3**, 367-388.

[26] Güth W, Kocher MG (2014). More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. *Journal of Economic Behavior and Organization*, **108**, 396-409.

[27] Houba H, Li D, Wen Q (2022). Bargaining with costly competition for the right to propose. *Journal of Mathematical Economics*, **98**, 102558.

[28] Jackson MO, Pernoud A (2020). Credit freezes, equilibrium multiplicity, and optimal bailouts in financial networks. *arXiv preprint arXiv*, **2012.12861**.

[29] Kant I 1996 (1781). *Critique of Pure Reason*, eds. JW Ellington, P Kitcher and WS Pluhar. Hackett Publishing.

[30] Kara T, Karagözoğlu E, Özcan-Tok E (2021). Bargaining, reference points, and limited influence. *Dynamic Games and Applications*, **11**, 326-362.

[31] Karagözoğlu E, Keskin K (2018). Time-varying fairness concerns, delay, and disagreement in bargaining. *Journal of Economic Behavior and Organization*, **147**, 115-128.

[32] Karagözoğlu E, Keskin K, Sağlam Ç (2023). (In)efficiency and equitability of equilibrium outcomes in a family of bargaining games. *International Journal of Game Theory*, **52**, 175-193.

[33] Karagözoğlu E, Riedl A (2015). Performance information, production uncertainty, and subjective entitlements in bargaining. *Management Science*, **61**, 2611–2626.

[34] Kim Y (2003). Income distribution and equilibrium multiplicity in a stigma-based model of tax evasion. *Journal of Public Economics*, **87**, 1591-1616.

[35] Kimbrough E, Porter D, Schneider M (2021). Reference dependent prices in bargaining: An experimental examination of precise first offers. *Journal of Economic Psychology*, **86**, 102406.

[36] Kohler S (2013). Envy can promote more equal division in alternating-offer bargaining. *Journal of Neuroscience, Psychology, and Economics*, **6**, 31–41.

[37] Kohler S, Schlag KH (2019). Inequality Aversion Causes Equal or Unequal Division in Alternating-Offer Bargaining. *Bulletin of Economic Research*, **71**, 47-57.

[38] Kroll EB, Morgenstern R, Neumann T, Schosser S, Vogt B (2014). Bargaining power does not matter when sharing losses – Experimental evidence of equal split in the Nash bargaining game. *Journal of Economic Behavior and Organization*, **108**, 261-272.

[39] Laffont J-J (1975). Macroeconomic constraints, economic efficiency, and ethics: An introduction to Kantian economics. *Economica*, **42**, 430-437.

[40] Luhan W, Poulsen O, Roos M (2019). Money or morality: fairness ideals in unstructured bargaining. *Social Choice and Welfare*, **53**, 655-675.

[41] Malueg D (2010). Mixed-strategy equilibria in the Nash Demand Game. *Economic Theory*, **44**, 243-270.

[42] Miettinen T, Kosfeld M, Fehr E, Weibull J (2020). Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions. *Journal of Economic Behavior and Organization*, **173**, 1-25.

[43] Muñoz E (2022). Taxing moral agents. *Working Paper.*

[44] Nash JF (1953). Two person cooperative games. *Econometrica*, **21**, 128–140.

[45] Nydegger RV, Owen G (1975). Two person bargaining: an experimental test of Nash axioms. *International Journal of Game Theory*, **3**, 239–249.

[46] Ostrom E (1990). *Governing the Commons: the Evolution of Institutions for Collective Action.* Cambridge: Cambridge University Press.

[47] Robles J (2008). Evolution, bargaining, and time preferences. *Economic Theory*, **35**, 19-36.

[48] Rocheteau G, Wright R (2005). Money in search equilibrium, in competitive equilibrium, and in competitive search equilibrium. *Econometrica*, **73**, 175-202.

[49] Roemer JE (2010). Kantian equilibrium. *Scandinavian Journal of Economics*, **112**, 1–24.

[50] Rivero JI (2023). Trade among moral agents with information asymmetries. *THEMA Working Paper*, **no: 2023-10**.

[51] Shalev J (2002). Loss aversion and bargaining. *Theory and Decision*, **52**, 201–232.

[52] Schelling TC (1960). *The Strategy of Conflict.* Cambridge, MA: Harvard University Press.

[53] Skyrms B (1996). *Evolution of the Social Contract.* Cambridge: Cambridge University Press.

[54] Van Leeuwen B, Alger I (2021). Estimating social preferences and Kantian morality in strategic interactions. *Working Paper.*

[55] Young PH (1993). An evolutionary model of bargaining. *Journal of Economic Theory*, **59**, 145–168.

8 Appendix

8.1 Omitted Proofs Section 5

Proof. Lemma 2: By Definition 1, we know that the strategy $s = (\frac{1}{2}, \frac{1}{2})$ is a (symmetric) Bayesian Nash equilibrium strategy if the following conditions are satisfied:

$$\tfrac{1}{2} \in \text{argmax}_{x\in[0,1]}(1-\underline{\kappa})(p\pi(x,\tfrac{1}{2}) + (1-p)\pi(x,\tfrac{1}{2})) + \underline{\kappa}\pi(x,x),$$

$$\tfrac{1}{2} \in \text{argmax}_{x\in[0,1]}(1-\overline{\kappa})(p\pi(x,\tfrac{1}{2}) + (1-p)\pi(x,\tfrac{1}{2})) + \overline{\kappa}\pi(x,x)$$

Note that this is equivalent to

$$\tfrac{1}{2} \in \text{argmax}_{x\in[0,1]}(1-\underline{\kappa})\pi(x,\tfrac{1}{2}) + \underline{\kappa}\pi(x,x),$$

$$\tfrac{1}{2} \in \text{argmax}_{x\in[0,1]}(1-\overline{\kappa})\pi(x,\tfrac{1}{2}) + \overline{\kappa}\pi(x,x),$$

which is satisfied for any p, $\underline{\kappa}$ and $\overline{\kappa}$, as both $\pi(x,\frac{1}{2})$ and $\pi(x,x)$ are maximized at $x = \frac{1}{2}$. □

Proof. Lemma 3: By contradiction, suppose that there exists $(s^*_{\underline{\kappa}}, s^*_{\overline{\kappa}})$ with $s^*_{\underline{\kappa}} + s^*_{\overline{\kappa}} < 1$ and $s^*_{\underline{\kappa}} < \frac{1}{2}$ (with analogous arguments for the case $s^*_{\overline{\kappa}} < \frac{1}{2}$). By Definition 1, we know that $s^*_{\underline{\kappa}}$ must satisfy:

$$s^*_{\underline{\kappa}} \in \text{argmax}_{x\in[0,1]}(1-\underline{\kappa})(p\pi(x,\mathrm{s}^*_{\underline{\kappa}}) + (1-p)\pi(x,\mathrm{s}^*_{\overline{\kappa}})) + \underline{\kappa}\pi(x,x)$$

However, $s^*_{\underline{\kappa}} < \frac{1}{2}$ implies that $s^*_{\underline{\kappa}} + s^*_{\underline{\kappa}} < 1$, which implies that a strategy $(x, s^*_{\overline{\kappa}})$ with x slightly above $s^*_{\underline{\kappa}}$ is a profitable deviation, as this increases the payoff $\pi(x, s^*_{\underline{\kappa}})$, $\pi(x, s^*_{\overline{\kappa}})$, and $\pi(x,x)$. Thus, there cannot exist a (symmetric) equilibrium BNE strategy with $s^*_{\underline{\kappa}} + s^*_{\underline{\kappa}} < 1$.

By contradiction, suppose that there exists $(s^*_{\underline{\kappa}}, s^*_{\overline{\kappa}})$ with $s^*_{\underline{\kappa}} + s^*_{\overline{\kappa}} > 1$ and $s^*_{\underline{\kappa}} > \frac{1}{2}$ (with analogous arguments for the case $s^*_{\overline{\kappa}} > \frac{1}{2}$). The expected utility of an individual with $\kappa = \underline{\kappa}$ is equal to zero. But then, an strategy $(1 - s^*_{\overline{\kappa}}, s^*_{\overline{\kappa}})$ is a profitable deviation if $s^*_{\overline{\kappa}} < 1$. When $s^*_{\overline{\kappa}} = 1$, then an individual with $\kappa = \overline{\kappa}$ gets a utility of zero. An strategy $(s^*_{\underline{\kappa}}, x)$ with $x \leq \frac{1}{2}$ is a profitable deviation for any $s^*_{\underline{\kappa}}$. Thus, there cannot exist a (symmetric) equilibrium BNE strategy with $s^*_{\underline{\kappa}} + s^*_{\underline{\kappa}} > 1$.

□

8.2 Proofs of Section 6

Proof. Proposition 5: Note that any Nash equilibrium characterized by the Proposition 5 satisfies ${x_1}^* + {x_2}^* > 1$, and therefore it gives both individuals a utility of zero. Therefore, to check that individuals do not have any incentive to deviate, we need to consider the following constraint:

$$0 \geq (1-x) - \alpha(x - (1-x)) = (1-x) - \alpha(2x-1).$$

Intuitively, the unique deviation that we need to consider is the one where the individual moves to the highest demand that satisfies ${x_1}^* + {x_2}^* = 1$. Then, the previous constraint can be rewritten as $x \geq \frac{1+\alpha}{1+2\alpha}$. Therefore, any allocation where both individuals have $x_i \geq \frac{1+\alpha}{1+2\alpha}$ is a Nash equilibrium. □

Proof. Proposition 6: To characterize the set of asymmetric Nash equilibria with ${x_1}^* + {x_2}^* = 1$, we start by considering the case where ${x_1}^* < {x_2}^*$. Any Nash equilibrium $({x_1}^*, {x_2}^*)$ with ${x_1}^* < {x_2}^*$ must satisfy the following constraints:

$$\begin{aligned} {x_1}^* - \alpha_1({x_2}^* - {x_1}^*) &\geq 0 \\ {x_2}^* - \beta_2({x_2}^* - {x_1}^*) &\geq 0 \end{aligned}$$

where the left-hand side represents individuals' utilities under the allocation $({x_1}^*, {x_2}^*)$, and the right-hand side is their utility when they deviate by increasing their demand.[14] The same arguments follow for the case with ${x_1}^* > {x_2}^*$. □

Proof. Proposition 7: To show this result suppose that the individual 2 demands ${x_2}^* \in [0,1]$. In this case, the unique best response of the individual 1 is to demand ${x_1}^* = (1 - {x_2}^*) \in [0,1]$ which gives him a strictly positive utility. If the individual 1 chooses a demand $x_1 > {x_1}^*$ he obtains a utility of zero, and therefore he decreases his utility. On the other hand, if the individual 1 chooses a demand $x_1 < {x_1}^*$ he also decreases his utility as he decreases his material payoff without increasing the material payoff of his pair. □

[14]Note that individual 1 does not have incentives to decrease his demand, as this decreases his material payoff while increasing the inequality between the players. Individual 2 does not have incentives to decrease their demand, as decreasing his demand by $\Delta {x_2}^*$ decreases individual 2's utility by $(1-\beta)\Delta {x_2}^* < \Delta {x_2}^*$.