ON BARGAINING NORMS

TYMON TATUR[†]

Abstract. This paper studies bargaining outcomes in economies in which agents may be able to impose outcomes that deviate from the relevant social norm, but incur costs when they decide to do so. It characterizes bargaining outcomes that are easiest to sustain as a social norm to which everybody will want to adhere.

Depending on the nature of the costs, the approach yields concepts like the Nash Bargaining solution, the Kalai-Smorodinsky solution, or – for coalitional games with transferable payoffs – refinements of the core. Set-valued solution concepts are derived that are relevant if one is unable or unwilling to make specific assumptions about the costs.

1. Introduction

A bargaining situation is a situation in which participating individuals have the opportunity to collaborate for mutual benefit and can divide the resulting surplus in more than one way. Following Nash (1950, 1953) much of the modern theory on bargaining can be divided into two branches. The axiomatic approach, born with Nash (1950) makes predictions about bargaining outcomes based on assumptions about how outcomes in different bargaining situations differ. The strategic approach, initiated in Nash (1953), considers a single bargaining situation in isolation and uses a non-cooperative game to model the strategic incentives players may face when negotiating an agreement. Once an agreement is reached, the game typically ends.

[†] Department of Economics, University of Bonn, Adenauerallee 24-42, D-53113 Bonn; Email: tatur@uni-bonn.de. I thank Philip Strack, Faruk Ghul, Stephen Morris, Antonio Penta, Wolfgang Pesendorfer, Ran Spiegler, Yuso Valimaki, Francoise Forges, Marek Pycia, Stephan Lauermann, Benny Moldovanu, Larry Samuelson, and Paul Heidhues for helpful questions and comments.

In contrast, this paper is concerned with bargaining outcomes in societies where cooperation is governed by social norms. The idea that cooperation will be often governed by social norms is not new. As a matter of fact, Kenneth Arrow (1971, p. 22) argued that a primary reason for the existence of social norms may be to facilitate cooperation by creating environments in which individuals can trust each other:

> It is a mistake to limit collective action to state action... I want to [call] attention to a less visible form of social action: norms of social behavior, including ethical, and moral ones. I suggest as one possible interpretation that they are reactions of society to compensate for market failure. It is useful for individuals to have some trust in each other's word. In the absence of trust, it would become very costly to arrange for alternative sanctions and guarantees, and many opportunities for mutual beneficial cooperation would have to be forgone.

In the above quote Arrow points out the potential benefits of norms in situations where in absence of such norms (whether internalized or enforced through sanctions) cooperation may be impossible or highly inefficient. It is clear that trust can be very important in prolonged cooperations - efficient cooperation may, for instance, be difficult if each of the participating parties has to constantly watch the other knowing that the other party will steal the entire jointly produced surplus if given the opportunity. Less obvious is perhaps that norms may also prevent inefficiencies in situations where complete contracts specifying any division of surplus are available. Consider, for instance, a buyer and a seller who can write down a legally binding contract specifying the terms of delivery and the price for the sold good, knowing that courts will if needed enforce both the delivery of the good and the payment. As was pointed out by Crawford (1982), in such situations inefficiencies will often occur if agents can imperfectly commit to bargaining positions before bargaining starts. Of course, if individuals who deviate from the existing norm are sanctioned (and sanctions are sufficiently high) than individuals will have no incentive to imperfectly commit to alternative divisions of surplus and the inefficiency disappears.

What allocation of surplus should we expect if norms are used to avoid inefficiencies like those mentioned above? To understand the basic idea of the approach proposed in this paper, consider the following extremely stylized example. Two risk-neutral agents can engage in an activity that creates a monetary surplus of \$1. Agents can agree on any allocation of surplus $(x_1, x_2) \in \{[0,1]^2 : x_1 + x_2 = 1\}$, where x_1 is the surplus received by player 1 and x_2 is the surplus received by player 2. However, in the spirit of the inefficiencies mentioned by Arrow, assume such agreements are not easily perfectly enforceable - before the cooperation is complete and the \$1 can be divided, each side will repeatedly have an opportunity to "steal" 90 cents of the produced surplus and leave while the other party is taking a break. Note that no matter what allocation of surplus of the \$1 the two agents have agreed on, at least one party will have an incentive to break the agreement if he or she can get away with 90 cents. Perhaps, this problem can be overcome, for instance if both players don't make any breaks, or, if each player hires somebody to watch the other player while he or she takes a break. However, it is not hard to imagine situations where, as Arrow eloquently put it "in the absence of trust, it would become very costly to arrange for alternative sanctions and guarantees".

Imagine now, that a social norm is in place that mandates a division of surplus x, if a player deviates from the norm and leaves with 90 cents, that player will incur a "deviation cost" of m dollars. In case of an internalized norm, m could capture a feeling of guilt or anxiety after breaking the norm. In case of a norm which is upheld by sanctions, m could represent opportunity costs the agent incurs if he is shunned by others or actual costs if, for example, individuals breaking the norm are later bullied by others. It is clear that if m is sufficiently large, in the considered example at least 90 cents, any allocation x can be sustained as a norm in the sense that neither individual will find it advantageous to break the norm and leave with 90 cents, and thus both individuals will be able to take their breaks, trusting the other agent not to commit a theft.

While any allocation of surplus can be sustained as a norm if m is sufficiently large, the minimal m needed to sustain trust will depend on the allocation. Indeed, if player i gets x_i under the allocation x then for him not to be willing to break the norm and leave with 90 cents it must be that $0.9 - m \leq x_i$. Thus, an allocation x can be sustained as a norm given deviation costs m (in the sense that neither player would want to deviate and leave with 90 cents) if and only if $m \geq 0.9 - \min(x_1, x_2)$. In other words, the minimal deviation costs needed to sustain an allocation x are given by

$$m(x) = 0.9 - \min(x_1, x_2).$$

It is trivial to see that for x_1, x_2 which add up to 1 dollar, $m(x)$ is minimized for an allocation x^* where each player gets \$0.5, i.e. where the dollar is split evenly.

In other words, in the above situation, x^* is easier to sustain that any other allocation of surplus in the sense that the range of parameters m for which x^* can be sustained is strictly larger than the range of parameters for which any other allocation can be sustained. In particular, this implies that if an allocation $x \neq x^*$ can be sustained, so can x^*. Moreover, if maintaining a social system in which deviations yielding a higher net benefit are offset by more severe punishments for deviators is more costly for a society - see Remark 1 in Section 2 for a discussion why this may be the case for the type of deviation costs considered here - then x^* will be the allocation which can be sustained at lowest cost for that society.

In this paper we only compare allocations of surplus in terms of how easy they are to sustain as part of a social norm. The question when the benefits of a norm will be sufficiently large for norms to form is not addressed in this paper. If an important role of social norms is indeed to "compensate for market failures" then one would expect that whether a social norm will form or not will crucially depend on assumptions about how costly those "market failures" are for a society and how costly it is for society to monitor and punish deviators. In contrast, in this paper we only analyse how the allocation of surplus in a norm affects incentives of agents to deviate from the norm and thus will require no assumptions about how other agents are affected

by a deviation. Thus, in some sense, our results have a similar spirit as a result describing the cheapest way in which an agent can be incentivized to incur "high effort" in principal-agent problem with moral hazard, while ignoring the question whether a principal would actually want to induce "high effort" as the answer to this question would depend on assumptions on the preferences of the principal which are orthogonal to the studied incentive problem of the agent.

The paper is organized as follows. Section 2 considers bargaining between two players. We start by considering three concrete examples. The first is similar to the one sketched above - norms are sustained because agents who deviate incur monetary costs. In the other two examples norms are maintained because an attempt to deviate from the norm may result in the cooperation being permanently abandoned. For generic preferences over risk the allocation that is easiest to sustain as a norm is different in each of the three examples: in the first it is the equal division, in the second example Kalai-Smorodinsky solution, and in the third example the Nash Bargaining solution. What can we said more generally if one is unable or unwilling to make very specific assumptions about the underlying costs? Theorem 2 in Section 2 addresses this question in the context of two-player bargaining between players who differ in their attitude towards risk.

Section 3 considers bargaining between three or more individuals. Bargaining with more than two players has an interesting aspect that is not present in two-player bargaining. If subgroups can implement certain agreements even if others are unwilling to participate, then the threat of such an agreement can affect bargaining outcomes. In Section 3 we apply our approach to coalitional games with transferable payoffs to analyze this aspect of bargaining. One nice feature of the proposed approach is that one can immediately generalize solution concepts obtained for particular monitoring technologies in Section 2 (like the Nash bargaining solution or the Kalai-Smorodinsky solution) to the coalitional bargaining problems studied in Section 3, simply by considering the same monitoring technologies. For instance, the generalization of the

Nash bargaining solution to the considered class of coalitional games yields a refinement of the core concept that is related to classical concepts like the nucleolus. Solution concepts that yield sharp predictions when the core of a game is empty are also derived.

Since this paper provides alternative foundations for concepts like the Nash bargaining solution, the Kalai-Smorodinsky solution, refinements of the core, and more, our work can be seen as part of a large body of literature discussing foundations for those and related concepts. Our approach, however, differs from typical papers using the axiomatic approach (see, for example, Nash (1950), Kalai-Smorodinsky (1975), or Rubinstein et al. (1992)) as a single type of bargaining problem is considered in isolation and no assumptions are made about how bargaining outcomes will change if some aspects of the bargaining situation (like the set of alternatives or the preferences of the players) are modified. Our approach also differs from papers using the strategic approach (see, for example, Nash (1953), Rubinstein (1982), Abreu and Gul (2000), Compte and Jehiel (2010), Perry and Reny (1994)) and, more generally, papers using non-cooperative game theory, as we do not select outcomes based on standard solution concepts used in non-cooperative game theory.

If one thinks about social norms that are internalized (i.e. part of the agents preferences) the proposed approach seems related to a literature studying the evolution of preferences in reduced models in which Nature designs preferences to avoid certain inefficiencies as in Samuelson (2004) or Samuelson and Swinkels (2006). Papers that use evolutionary game theory to select Nash equilibria in non-cooperative bargaining games (see, for instance, Young (1993)) appear less related as the methodology is again very different.

2. Bargaining between Two Individuals

Consider the problem of two agents who can engage in some activity that creates a monetary surplus - for the sake of concreteness we will assume that the surplus is equal to \$1 - and have to decide how to divide the dollar.

Let

$$\mathcal{X} = \{(x_1, x_2) \in [0,1]^2 : x_1 + x_2 = 1\}$$

be the set of possible allocations of the monetary surplus, where $(x_1, x_2) \in \mathcal{X}$ is interpreted as an allocation where player 1 receives x_1 and player 2 receives x_2. We will allow that players differ in their attitudes toward risk. More formally, if an allocation of surplus $x \in \mathcal{X}$ is implemented, players receive von Neumann-Morgenstern utilities $u_1(x_1)$ and $u_2(x_2)$ respectively, where u_i for $i = 1, 2$ are differentiable functions satisfying $u_i' > 0$ and $u_i'' \leq 0$. In the following three subsections we will assume that the utility functions have been normalized so that $u_1(0)$ and $u_2(0)$ are both zero. In the context of the general framework introduced later, Remark 2 points out that these assumptions are without loss of generality.

We start by analyzing three simple ways in which norms can be sustained. In each case, there will be a unique outcome that is easiest to sustain as a social norm. The three outcomes obtained in this way correspond to the Nash bargaining solution, the Kalai-Smorodinsky solution, and the equal division in which each player receives fifty cents.

2.1. **Example: Norms Sustained through Simple Monetary Sanctions.** Let us start by imagining that each time a player deviates from the social norm he or she incurs a fixed monetary cost of m dollars.

A fixed monetary deviation cost allows a number of different interpretations. For example, it could be that a player violates relevant norms or customs is later shunned by others and incurs opportunity costs of m dollars in some unrelated interactions. Alternatively, it could be that a players who violates the social norm is actively punished by others and incurs actual losses that correspond to m dollars. In case of internalized norms, m could represent the psychological pain or anxiety that an individual feels after deviating from a norm. Finally, m could correspond to the costs that an individual has to incur to avoid sanctions given the level of social monitoring in the particular society - such costs will be discussed in Remark 3 after the general framework has been introduced.

Assume the relevant social norm dictates an allocation $x \in \mathcal{X}$. Consider whether a player $i \in \{1, 2\}$ who expects to receive x_i under the allocation x would want to impose an alternative allocation of surplus $x' \in \mathcal{X}$ if this would result in a fixed monetary cost of $m \in [0, \infty)$.[1] If the player receives x_i his utility will be $u_i(x_i)$. If he imposes x' and pays a cost of m his utility will be $u_i(x'_i - m)$. Thus, player i would have no incentive to impose the alternative allocation x' if and only if

$$u_i(x_i) \geq u_i(x'_i - m). \tag{1}$$

We will say that an allocation $x \in \mathcal{X}$ *can be sustained for deviation costs* m if and only if, for all players $i \in \{1, 2\}$ and $x' \in \mathcal{X}$, inequality (1) holds. If an allocation x can be sustained given deviation costs m that means that if the norm specifies a division according to x and deviators face a monetary cost of m each individual can trust that the other will not impose a different allocation of surplus even if given a chance.

We will say that an allocation $x \in \mathcal{X}$ is *easier to sustain as a norm* than an allocation $x' \in \mathcal{X}$ if and only if it is the case that $\mathcal{S}(x') \subsetneq \mathcal{S}(x)$. An allocation $x \in \mathcal{X}$ is said to be *easiest to sustain as a norm* if it is easier to sustain than any other allocation $x' \neq x$.

For any $x \in \mathcal{X}$, the set of all numbers $m \in [0, \infty)$ such that x can be sustained for a given m will be denoted by $\mathcal{S}(x)$. Note that, since the functions u_i are increasing and m does not depend on x'_i, an allocation x can be sustained for a given m (i.e. $m \in \mathcal{S}(x)$) if and only if, for each $i \in \{1, 2\}$, $x_i \geq 1 - m$ or, equivalently, $m \geq 1 - x_i$. Since $x_1 + x_2 = 1$, this means that

$$\mathcal{S}(x) = [\max(x_1, 1 - x_1), \infty).$$

In other words, the minimal m needed to sustain an allocation x is given by $\max(x_1, 1 - x_1)$.

[1] In this example we consider the case where m is fixed and does not depend on x' and x - situations where m can depend on x and x' will be considered later.

Clearly, $\max(x_1, 1-x_1)$ as a function of x_1 has a unique minimum in $x_1 = \frac{1}{2}$. Thus, for any allocation $x \in \mathcal{X}$ such that $x \neq (\frac{1}{2}, \frac{1}{2})$, it is the case that $\mathcal{S}(x) \subsetneq \mathcal{S}((\frac{1}{2}, \frac{1}{2}))$ and we obtain the following result.

Proposition 1. *For the deviation costs considered in this subsection, there exists an allocation of surplus that is easier to sustain as a norm than any other allocation of surplus and that allocation is* $(\frac{1}{2}, \frac{1}{2})$*, the allocation in which each player receives 50 cents.*

The proposition immediately implies that if we can sustain some allocation $y \neq (\frac{1}{2}, \frac{1}{2})$ can be sustained using sanctions corresponding to m dollars, we can also sustain $(\frac{1}{2}, \frac{1}{2})$ with m dollars. Thus, from a social point of view, sustaining $(\frac{1}{2}, \frac{1}{2})$ as a norm never has to be more costly than sustaining y.

In addition the proposition also implies that if an allocation $y \neq (\frac{1}{2}, \frac{1}{2})$ can be sustained using sanctions corresponding to m dollars, $(\frac{1}{2}, \frac{1}{2})$ can be sustained with with sanctions that are $m' = \frac{1}{2}$ and $m' < m$. This implies that if a social system with higher sanctions m is more costly to maintain then $(\frac{1}{2}, \frac{1}{2})$ the costs needed to sustain $(\frac{1}{2}, \frac{1}{2})$ will actually be strictly lower than the costs needed to sustain $y \neq (\frac{1}{2}, \frac{1}{2})$.

Remark 1. *Why should it be more costly to maintain a social system in which there are more profitable deviations but those are offset by more severe punishments for deviators?*

If m represents how deeply internalized a norm is (i.e. how much anxiety or guilt an individual feels when breaking the norm) it appears natural to assume that inducing higher m is more costly. Similarly, if m represents a cost that an individual needs to incur to avoid that a deviation is later detected and sanctioned, higher m will correspond to better monitoring and thus be more costly. What if m represents costs that a deviator incurs as a result of actual social sanctions, say if deviators are shunned or bullied by other members of the society and higher m correspond to more intense bullying or a longer time period in which the individual is shunned after a deviation? After all, if the norm is successfully sustained, no deviations will occur and, therefore,

nobody will need to be bullied or shunned! Note that, even in this case, higher m will typically require more monitoring as individuals doing the shunning or bullying need to be monitored. Indeed, if shunning a deviator results in opportunity costs for other members of society, those agents will need to monitored and incentivized to make sure that they indeed do shun a deviator and the longer the time period in which a deviator is excluded from interactions with others the more monitoring will be necessary. Similarly, if, for example, bullying an individual is pleasurable for other members of society[2]*, monitoring will be required to make sure that only deviators get bullied - if both deviators and non-deviators get bullied, bullying no longer would work as a sanction. Since higher levels of monitoring are required for higher m (and this monitoring needs to take place even if nobody finds it optimal to deviate), also here it appears natural that a social system with larger sanctions will require some additional costs.*

Note that the fact that the allocation $(\frac{1}{2}, \frac{1}{2})$ does not depend on the utility functions is not surprising - since a constant monetary cost does not involve any uncertainty, players attitude toward risk is irrelevant.

2.2. **Example: Norms Sustained through Threat of Permanent Disagreement.** In this subsection we will assume that the cost a player incurs if he wants to deviate from the prevailing standard is not monetary but rather is derived from the fact that with an exogenous fixed positive probability p the bargaining process will permanently terminate in disagreement. Again, p allows for a number of interpretations. For instance, it could that if one player tries to impose an allocation that deviates from the norm, especially if this involves an act which is seen as immoral - like breaking a previously given promise - with probability p the other player has internalized the norm so strongly that she will stop the cooperation even if it is costly for her to do so. Alternatively, again if imposing a different allocation involves an act like lying or committing a fraud, it could be that the other player will always respond by terminating the cooperation but such acts are only detected with probability p

[2]For instance, "bullying" could involve a transfer of wealth or services from the bullied person to the individual doing the bullying.

given the level of monitoring in a given society. Finally, it could be that if a player tries tries to impose an allocation that deviates from the norm, with probability p the other player will stop cooperation as making a deal in which he accepts less than the valid norm sometimes would make him loose face in front of others or expose himself or herself to social sanctions.

Formally, assume that whenever a player i tries to impose an allocation $x' \in \mathcal{X}$ that allocation will be implemented with probability $1-p$ and with probability p the outcome will be permanent disagreement giving i a payoff of $u_i(0) = 0$. Thus, if player i tries to impose an outcome x' his expected payoff will be equal to $(1-p) \cdot u_i(x'_i)$. This means that a player i who expects to receive x_i under some allocation x would have no incentive to try to impose an alternative allocation x' if and only if

$$u_i(x_i) \geq (1-p) \cdot u_i(x'_i). \tag{2}$$

Analogously as in the last subsection, we will say that an allocation $x \in \mathcal{X}$ *can be sustained as a norm for a given* p if and only if, for all players $i \in \{1,2\}$ and $x' \in \mathcal{X}$, inequality (2) holds. For any $x \in \mathcal{X}$, the set of all numbers $p \in [0,1]$ such that x can be sustained a norm will be denoted by $\mathcal{S}(x)$. Again, an allocation $x \in \mathcal{X}$ is said to be *easier to sustain* than an allocation $x' \in \mathcal{X}$ if and only if $\mathcal{S}(x') \subsetneq \mathcal{S}(x')$. Like in the last subsection, we will say that an allocation $x \in \mathcal{X}$ is said to be *easiest to sustain as a norm* if it is easier to sustain than any other allocation $x' \neq x$.

Note that, since p does not depend on x', whenever inequality (2) is not satisfied for some player i and $x' \in \mathcal{X}$, it will also not hold for that player i and a division x' in which i gets the entire dollar. Thus, $p \in \mathcal{S}(x)$ if and only if

$$u_i(x_i) \geq (1-p) \cdot u_i(1)$$

for $i \in \{1,2\}$. This means that

$$\mathcal{S}(x) = [1 - \min(\frac{u_1(x_1)}{u_1(1)}, \frac{u_2(x_2)}{u_2(1)}), 1].$$

In other words, the minimal value of p needed to sustain an allocation x is given by $1 - \min(\frac{u_1(x_1)}{u_1(1)}, \frac{u_2(x_2)}{u_2(1)})$.

It is straightforward to see that $\min(\frac{u_1(x_1)}{u_1(1)}, \frac{u_2(x_2)}{u_2(1)})$ achieves its maximum for the Kalai-Smorodinsky solution, i.e. the unique allocation $x^{K.S.}$ satisfying $\frac{u_1(x_1)}{u_1^{K.S.}(1)} = \frac{u_2(x_2^{K.S.})}{u_2(1)}$. We have shown the following result.

Proposition 2. *For the deviation costs considered in this subsection, there exists an allocation of surplus that is easier to sustain as a norm than any other allocation of surplus and that allocation is the Kalai-Smorodinsky solution, i.e. the unique allocation $x^{K.S.}$ such that $\frac{u_1(x_1^{K.S})}{u_1(1)} = \frac{u_2(x_2^{K.S.})}{u_2(1)}$.*

Again, if it is more costly for a society to implement higher p, the Kalai-Smorodinsky solution will be the unique allocation that is cheapest to sustain.

2.3. **Example: Nash Punishments.** In this subsection we will consider an example where the probability of permanent disagreement p considered in the last section does depend on how large the deviation from the norm is and more extreme deviations result in a higher chance that cooperation breaks down permanently. The variable probability p in this subsection can be interpreted similarly as the fixed probability p from Subsection 2.2, except that the probability with which interactions are terminated now depends on how large the deviation was. If, for instance, after a deviation from a norm by one player there is a chance that the other player has internalized the norm so strongly that he will stop all cooperation, this chance would now be higher the more extreme the deviation is.

More formally, assume that if the norm is x and player i tries to impose an alternative allocation x' with $x_i' > x_i$, permanent disagreement happens with probability $p(x_i' - x_i)$, where $p : [0,1] \to [0,1]$ is a differentiable function satisfying $p(0) = 0$, $p' > 0$, and $p'' \geq 0$. The conditions $p(0) = 0$, $p' > 0$, and $p'' \geq 0$ capture the idea that there is little cost of imposing an allocation "close to the norm x", costs increase the more excessive x' becomes, and "marginal costs are increasing". Let P be the set of functions $p : [0,1] \to [0,1]$ that satisfy the above three properties.

Analogously as in the last two subsections, we will say that an allocation $x \in \mathcal{X}$ *can be sustained as a norm for a given* $p \in P$ if and only if

$$u_i(x_i) \geq (1 - p(x'_i - x_i)) \cdot u_i(x'_i), \tag{3}$$

holds for all players $i \in \{1, 2\}$ and all allocations $x' \in \mathcal{X}$ such that $x'_i > x_i$.[3]

Again, for any $x \in \mathcal{X}$, the set of all $p \in P$ such that x can be sustained a norm will be denoted by $\mathcal{S}(x)$. Like in the last two subsections, an allocation $x \in \mathcal{X}$ is said to be *easier to sustain* than an allocation $x' \in \mathcal{X}$ if and only if $\mathcal{S}(x') \subsetneq \mathcal{S}(x')$. Finally, we will say that an allocation $x \in \mathcal{X}$ is said to be *easiest to sustain as a norm* if it is easier to sustain than any other allocation $x' \neq x$.

We claim that, for any allocation $x \in \mathcal{X}$, the set $\mathcal{S}(x)$ satisfies

$$\mathcal{S}(x) = \{p \in P : p'(0) \cdot u_1(x_1) \geq u'_1(x_1) \text{ and } p'(0) \cdot u_2(x_2) \geq u'_2(x_2)\}. \tag{4}$$

To see that (4) holds, consider any allocation $x \in \mathcal{X}$ and any function $p \in P$. For $i \in \{1, 2\}$, define $f_i : [x_i, 1] \to \mathbb{R}$ by $f_i(x'_i) = (1 - p(x'_i - x_i)) \cdot u_i(x'_i)$. Now inequality (3) can be rewritten as

$$u_i(x_i) \geq f_i(x'_i). \tag{5}$$

Note that, since $p(0) = 0$ for any $p \in P$, inequality (5) holds with equality if $x'_i = x_i$. Since the functions f_i are concave,[4] this implies that $f'(x_i) \leq 0$ is a sufficient and necessary condition for inequality (5) to hold for all $x'_i \in (x_i, 1]$, whenever $x_i < 1$. Since $f'(x_i) \leq 0$ is equivalent to $p'(0) \cdot u_i(x_i) \geq u'_i(x_i)$, this proves (4) for the case where $x_1 < 1$ and $x_2 < 1$. However, for the case where $x_1 = 1$ or $x_2 = 1$, equation (4) holds since $\mathcal{S}(x) = \emptyset$ and the right hand side of (4) is also equal to the empty set.[5]

[3]We only consider deviations to x' which are more favorable for player i in the sense that $x'_i > x_i$. If imposing an alternative allocation of surplus is related with additional costs, a player would never have an incentive to impose an allocation x' which gives him less than x.

[4]To see that f_i is concave note that $u_i \geq 0$, $u'_i > 0$, $u''_i \leq 0$, $p \leq 1$, $p' > 0$, $p'' \leq 0$ imply that $f''_i = -p'' \cdot u_i - p' \cdot u'_i + (1 - p) \cdot u''_i \leq 0$.

[5]Assume $x_i = 1$ for some $i \in \{1, 2\}$. $\mathcal{S}(x) = \emptyset$ must hold since for any $p \in P$, it will be the case that (3) does not hold for player j with $x_j = 0$ and positive x'_j that are sufficiently close to zero. To see that the right hand side of (4) is also equal to the empty set, note that, if $x_j = 0$, then, for any $p \in P$, $p'(0) \cdot u_j(x_j) = 0 < u'_j(x_j)$ as $u_j(0) = 0$ and $u'_j > 0$.

Equation (4) implies that an allocation $x \in \mathcal{X}$ is easier to sustain as a norm than an allocation $y \in \mathcal{X}$ if and only if

$$\max(\frac{u_1'(x_1)}{u_1(x_1)}, \frac{u_2'(x_2)}{u_2(x_2)}) < \max(\frac{u_1'(y_1)}{u_1(y_1)}, \frac{u_2'(y_2)}{u_2(y_2)})$$

or, equivalently,

$$\max(\frac{u_1'(x_1)}{u_1(x_1)}, \frac{u_2'(1-x_1)}{u_2(1-x_1)}) < \max(\frac{u_1'(y_1)}{u_1(y_1)}, \frac{u_2'(1-y_1)}{u_2(1-y_1)}).$$

Consider

$$\max(\frac{u_1'(x_1)}{u_1(x_1)}, \frac{u_2'(1-x_1)}{u_2(1-x_1)}) \tag{6}$$

as a function of x_1. Since the functions u_1 and u_2 are concave and increasing, $\frac{u_1'(x_1)}{u_1(x_1)}$ is decreasing in x_1 and $\frac{u_2'(1-x_1)}{u_2(1-x_1)}$ is increasing in x_1. Thus, there is a single allocation for which (6) is minimized and that allocation is the unique solution of the equation

$$\frac{u_1'(x_1)}{u_1(x_1)} = \frac{u_2'(1-x_1)}{u_2(1-x_1)}. \tag{7}$$

However, the unique allocation for which equation (7) holds is the symmetric Nash Bargaining Solution.[6] Therefore, we obtained the following result.

Proposition 3. *For the deviation costs considered in this subsection, there exists an allocation of surplus that is easier to sustain as a norm than any other allocation of surplus and that allocation is the symmetric Nash Bargaining solution, i.e. the unique solution to the problem* $\max_x u_1(x) \cdot u_2(x)$.

The reason for the name "Nash Punishments" in the title of this subsection is that the deviation costs considered in this example had a very particular property. The set $\mathcal{S}(x)$ depended only on the local properties of the utility functions u_1 and u_2 around x and the disagreement payoff, which is reminiscent of Nash's (1950) Independence of Irrelevant Alternatives Axiom.

[6]The symmetric Nash Bargaining Solution is the unique allocation solving $\max_x u_1(x_1) \cdot u_2(1-x_1)$. It is straightforward to see that this problem has an interior solution. However, the first order condition $u_1'(x_1) \cdot u_2(1-x_1) - u_1(x_1) \cdot u_2'(1-x_2) = 0$ is equivalent to equation (7).

2.4. **More General Framework.** In this section we will consider a more general framework in which norms can be sustained through a mixture of monetary sanctions after deviations and threats of permanent disagreement. We will still assume that players are symmetric in all aspects except their attitude towards risk.[7]

Let $\mathcal{C}$ be a set whose elements are pairs (p, m) where $p : [0,1] \to [0,1]$ and $m : [0,1] \to [0,\infty)$ are continuous functions. We call such a set $\mathcal{C}$ a *deviation cost set* and interpret it as set of possible ways in which a society can make it costly for agents to deviate from social norms. More precisely, a pairs $(p, m) \in C$ corresponds to a situation in which if, the social norm specifies $x \in X$ and a player i deviates from the norm attempting to impose an allocation in which he gets $x_i' \geq x_i$, player i's expected payoff will be given by

$$(1 - p(x_i' - x_i)) \cdot u_i(x_i' - m(x_i' - x_i)) + p(x_i' - x_i') \cdot u_i(0).$$

Here, $m_i(x_i' - x_i')$ represents a monetary cost the agent has to incur and $p(x_i' - x_i')$ the probability of permanent disagreement.[8]

As in subsections 2.1 - 2.3, we will say that an allocation can be sustained given a cost $c \in \mathcal{C}$ if no player has an incentive to impose an alternative allocation if the costs of doing so are given by c.

Definition 1. *An allocation $x \in \mathcal{X}$ is* can be sustained (as a norm) for given deviation costs $(p, m) \in \mathcal{C}$ *if and only if*

$$u_i(x_i) \geq (1 - p(x_i' - x_i)) \cdot u(x_i' - m(x_i' - x_i)) + p(x_i' - x_i) \cdot u_i(0) \tag{8}$$

[7]Considering the case where players are identical in all aspects except one seems like a natural starting point. In addition, it makes it easier to compare our results with classic symmetric bargaining concepts.

[8]Note that we assume that the monetary cost $m(x_i' - x_i)$ does not have to be paid if bargaining ends in permanent disagreement. This appears to be a natural assumption if only the final bargaining outcome is observable to other individuals as in this case other individuals will not know why no cooperation took place and who, if anybody, is to blame.

for all players $i \in \{1,2\}$ *and allocations* $x' \in \mathcal{X}$ *such that* $x'_i > x_i$.[9] *For any allocation* $x \in \mathcal{X}$*, denote the set of costs* $c \in \mathcal{C}$ *for which* x *can be sustained by* $\mathcal{S}_{\mathcal{C}}(x)$.

As in Subsections 2.1 - 2.3, we will compare allocations based on the set of $c \in \mathcal{C}$ for which they can be sustained as a norm.

Definition 2. *Fix a deviation cost set* $\mathcal{C}$*. An allocation* $x \in \mathcal{X}$*,* is easier to sustain (as a norm) than $y \in \mathcal{X}$ *(or* $x \succ_{\mathcal{C}} y$*) if and only if* $\mathcal{S}_{\mathcal{C}}(y) \subsetneq \mathcal{S}_{\mathcal{C}}(x)$.

The above definition immediately implies that the binary relation $\succ_{\mathcal{C}}$ on $\mathcal{X}$ is irreflexive (i.e. there is no x with $x \succ_{\mathcal{C}} x$) and transitive (i.e. for $x, y, z \in \mathcal{X}$, $y \succ_{\mathcal{C}} x$ and $z \succ_{\mathcal{C}} y$ implies $z \succ_{\mathcal{C}} x$). Thus, $\succ_{\mathcal{C}}$ is a strict partial order on $\mathcal{X}$.[10] As in sections 2.1 - 2.3 we will say that an allocations $x \in \mathcal{X}$ is easiest to sustain if it is largest with respect to that partial order.

It will be useful to introduce some more terminology to compare allocations.

Definition 3. *Fix a deviation cost set* $\mathcal{C}$*. An allocation* $x \in \mathcal{X}$ *is* equally easy to sustain (as a norm) as y *if* $\mathcal{S}_{\mathcal{C}}(x) = \mathcal{S}_{\mathcal{C}}(y)$.

Similarly as before, we will say that an allocation is easiest to sustain as a norm if and only if there is no allocation that is easier to sustain.

Definition 4. *Fix a deviation cost set* $\mathcal{C}$*. An allocation* $x \in \mathcal{X}$ *is* easiest to sustain (as a norm) *if and only* $\mathcal{S}_{\mathcal{C}}(x') \subsetneq \mathcal{S}_{\mathcal{C}}(x)$ *for any* $x' \in \mathcal{X}$ *such that* $x' \neq x$.

Note that the framework is rich enough to incorporate the settings of Subsections 2.1, 2.2, and 2.3. Of course, there are other natural generalizations of the three settings presented in Subsections 2.1, 2.2, and 2.3 that could be used. An earlier version of this paper derived very similar results as will be derived here for the case where the functions p and m were allowed to depend both on $x'_i - x_i$ and x_i and both functions were assumed to be monotonically non-decreasing in both $x'_i - x_i$ and x_i.

[9]We only consider deviations to x' which are more favorable for player i in the sense that $x'_i > x_i$. If imposing an alternative allocation of surplus is related with additional costs, a player would never have an incentive to impose an allocation x' which gives him less than x.

[10]A binary relation that is irreflexive and transitive is called a strict partial order.

A crucial property that both frameworks have is that a player's incentive to deviate from the norm will be lower if he or she receives a higher fraction of the surplus.

Since, for general deviation cost sets $\mathcal{C}$, it can be that there is no allocation that is easiest to sustain[11] we will often consider elements that are maximal rather than largest with respect to the partial order $\succ_{\mathcal{C}}$.

Definition 5. *An allocation $x \in \mathcal{X}$ is* undominated *if and only if there is no allocation $x' \in \mathcal{X}$ that is easier to sustain. For a given deviation cost set $\mathcal{C}$, the set of undominated allocations will be denoted by $\mathcal{U}_{\mathcal{C}}$.*

Note that, if for some deviation cost set $\mathcal{C}$, there is an allocation x^* that is easiest to sustain, the fact that it is easier to sustain than any other allocation y implies that, for that deviation cost set $\mathcal{C}$ it is also the unique allocation that is undominated. Thus, in Subsections 2.1, 2.2, and 2.3, the three propositions characterizing the allocations that are easiest to sustain would remain true if the words "there exists an allocation of surplus that is easier to sustain as a norm than any other allocation of surplus" would have been replaced by "there exists a unique allocation that is undominated".

Remark 2. *Note that the function $\mathcal{S}_{\mathcal{C}}$ defined in Definition 1 only depends the preferences of both players over monetary lotteries and, therefore, does not depend on which utility function is used to represent those preferences. In particular, $\mathcal{S}_{\mathcal{C}}$ will not be affected if positive affine transformations are applied to the utility functions u_1 and u_2. As a result, the same is true for the derived concepts in Definitions 2, 4 and 5. Since positive affine transformations of utility functions will not affect $\mathcal{U}_{\mathcal{C}}$, we can without loss of generality assume that $u_1(0) = u_2(0) = 0$ and $u_1(1) = u_2(1) = 1$ when proving statements about the set $\mathcal{U}_{\mathcal{C}}$.*

[11] Assume the utility functions u_1 and u_2 are such that the Nash bargaining solution does not coincide with the allocation which gives each player 50 cents. Now let $\mathcal{C}$ be the deviation cost set which includes both the deviation costs considered in Subsection 2.1 and the deviation costs considered in Subsection 2.2. The results from Subsections 2.1 and 2.2 imply that both the allocation giving each player 50 cents and the allocation corresponding to the Nash bargaining solution are maximal with respect to the partial order $\succ_{\mathcal{C}}$. In particular, there is no allocation of surplus that is easier to sustain than any other.

Remark 3. *We use the term "deviation cost set" for the set $\mathcal{C}$ and elements $c \in \mathcal{C}$ "deviation costs" rather than "sanction set" for the cost $\mathcal{C}$ and "sanctions" for its elements. A deviation from an existing norm can be costly because deviators face sanctions. However, there can be other reasons why deviations from a norm requires costs - in particular, it may be that c represents* costs that the individual has to incur to avoid sanctions.

Imagine, for example, that to gain a more beneficial allocation of surplus an action like theft or fraud is required that, if detected, will result in very severe legal or social consequences but will remain undetected (and unpunished) if the agent incurs a cost of m, where m depends on the level of social monitoring in that society. If the punishments after a detected case of theft or fraud are sufficiently severe, stealing a part of the surplus without incurring the cost m to keep the crime undetected will never be optimal. Thus, in such a case, the agent will deviate from the norm if and only if m, the cost he needs to incur to avoid sanctions is sufficiently low.

2.5. **Existence.** Our first theorem establishes the existence of undominated allocations.

Theorem 1. *For any deviation cost set $\mathcal{C}$, there exists an undominated allocation x^*, i.e. the set $\mathcal{U}_{\mathcal{C}}$ is non-empty.*

Proof. See Appendix. □

Despite the fact that the cost functions appearing in $\mathcal{C}$ are assumed to be continuous functions, the partial order $\succ_{\mathcal{C}}$ in general does not need to be a continuous binary relation and, in particular, the lower contour sets $L(y) = \{x \in \mathcal{X} : y \succ_{\mathcal{C}} x\}$ do not need to be open for all allocations y.[12] As a result, the existence of a maximal

[12]Let us sketch some intuition why the sets $L(y) = \{x \in \mathcal{X} : y \succ_{\mathcal{C}} x\}$ do not need to be open in general. In each of the examples in subsections 2.1 – 2.3, we had a set $\mathcal{C}$ and a unique element in $x_{\mathcal{C}} \in V^*_{\mathcal{C}}$, moreover, that $x_{\mathcal{C}}$ had the property that there was a cost $c \in \mathcal{C}$ such that $x_{\mathcal{C}}$ was the unique allocation that can be sustained given costs c. (For example, for the case of constant monetary costs, $(\frac{1}{2}, \frac{1}{2})$ was the unique allocation that can be sustained for fixed monetary costs of $\frac{1}{2}$.) Imagine that, for some utility functions u_1 and u_2, there is a non-empty open set $U \subsetneq \mathcal{X}$ such that for each $x \in U$ there is a cost $c_x = (p_x, m_x)$ such that x is the unique allocation that can be sustained under costs c_x. Let $y, z \in U$ be two different allocations. Consider $\mathcal{C} = \{c_x : x \in F - \{z\}\}$

element does not follow immediately from the compactness of $\mathcal{X}$ and the proof of Theorem 1 has to rely on a more subtle argument that uses the Kuratowski-Zorn lemma. The compactness of $\mathcal{X}$ and the fact that the functions appearing in $\mathcal{C}$ are continuous nevertheless are important in the proof.

2.6. **The sets $\mathcal{X}^*$ and $\mathcal{X}^{**}$.** All three examples considered in Subsections 2.1 – 2.3 had the property that there was a single allocation x that is easiest to sustain. One way to see "what is possible more generally" would be to ask what the set of allocations x is such that there is some deviation cost set $\mathcal{C}$ such that x is easiest to sustain if that deviation cost set $\mathcal{C}$ is considered.

Definition 6. *Let $\mathcal{X}^*$ be the set of all allocations $x^* \in \mathcal{X}$ such that there is a deviation cost set $\mathcal{C}$ with the property that x^* is easiest to sustain, i.e. such that $\mathcal{S}_{\mathcal{C}}(y) \subsetneq \mathcal{S}_{\mathcal{C}}(x^*)$ for any $y \in \mathcal{X}$ such that $y \neq x^*$.*

Note that $\mathcal{X}^*$ can be seen as a lower solution (see Myerson (1991), p. 107-108) in the sense that if $x^* \in \mathcal{X}^*$ then there will be environment where the allocation x^* is easiest (and cheapest in terms of social costs) to sustain. Of course, environments where there is an allocation that is easiest to sustain are rather specific. Can we derive an upper solution (again in the sense of Myerson (1991), p. 107-108) to say something about the possible allocations one should expect more generally?

The problem is that there are deviation cost sets $\mathcal{C}$ for which the set of undominated allocations $\mathcal{U}_{\mathcal{C}}$ is quite large. Consider the case where the deviation cost set is $\mathcal{C} = \{(p_0, m_1)\}$, with $p_0 \equiv 0$ and $m_1 \equiv 1$. In this case, clearly $\mathcal{S}_{\mathcal{C}}(x) = \mathcal{C}$ for any $x \in \mathcal{X}$, i.e. all allocations are equally easy to sustain. The problem here is that $\mathcal{C}$ is not rich enough to distinguish different allocations in $\mathcal{X}$.

Let us think again about the case where the deviation cost set is $\mathcal{C} = \{(p_0, m_1)\}$, with $p_0 \equiv 0$ and $m_1 \equiv 1$ a bit more. As was noted above, for this deviation cost set, $\mathcal{S}_{\mathcal{C}}(x) = \mathcal{C}$ for any $x \in \mathcal{X}$, i.e. all allocations are equally easy to sustain. If it would be guaranteed that the norm will only be used in situations where the cost

and note that, by construction, $L(y) = \{x \in \mathcal{X} : y \succ_{\mathcal{C}} x\}$ is equal to $\mathcal{X} - U \cup \{z\}$. Since U was open and $z \in U$, $L(y) = \mathcal{X} - U \cup \{z\}$ is not open.

set is $\mathcal{C} = \{(p_0, m_1)\}$, indeed, the allocation $(\frac{1}{2}, \frac{1}{2})$ would be as easy (and as costly) to sustain as the allocation $(1, 0)$ as the only way to sustain either uses (p_0, m_1). However, if the is any chance that the same norm would also be used in situations where the deviation cost set has the form considered in one of the examples from subsections 2.1 – 2.1, using $(\frac{1}{2}, \frac{1}{2})$ would be more efficient than using $(1, 0)$ in the sense that required sanctions/internalization would be smaller. Of course, the cost sets corresponding to the examples in sections 2.1 – 2.1 where just three possible cost sets. What if we found out, that also in any other cost set which is rich enough to that it can distinguish between $(1, 0)$ and $(\frac{1}{2}, \frac{1}{2})$ the allocation $(\frac{1}{2}, \frac{1}{2})$ is easier to sustain? This motivates the following definition.

Definition 7. *Let $\mathcal{X}^{**}$ be the set of all allocations $x \in \mathcal{X}$ such that there is no allocation $y \in \mathcal{X}$ with the property that, for any cost set $\mathcal{C}$, y is either easier or equally easy to sustain as x and, for some deviation cost set $\mathcal{C}$, y is easier to to sustain than x.*

2.7. **Characterization of $\mathcal{X}^*$ and $\mathcal{X}^{**}$.** A fundamental question that solution concepts like the Nash bargaining solution or the Kalai-Smorodinsky solution try to address is how different attitudes toward risk affect bargaining outcomes. A basic intuition is that, if one player is more risk averse than the other player, he or she will be more timid when making demands, allowing the player who is more aggressive in his demands to achieve a more favorable outcome.

Our approach will predict that exactly those allocations x are possible bargaining outcomes that are not unbalanced in the sense that, *given the allocation x*, one player will be strictly more risk averse than the other when making demands.

Definition 8. *Let $i \in \{1, 2\}$ be a player and j his opponent. Define $\mathcal{D}_i \subset \mathcal{X}$ to be the set of allocations x such that, for any $q \in (0, 1)$ and $\Delta \in (0, 1)$,*

$$x_i + \Delta \leq 1 \quad \textit{and} \quad u_i(x_i) < q \cdot u_i(x_i + \Delta) + (1 - q) \cdot u_i(0)$$

is implied by

$$x_j + \Delta \leq 1 \quad \textit{and} \quad u_j(x_j) \leq q \cdot u_j(x_j + \Delta) + (1 - q) \cdot u_j(0).$$

To understand Definition 8,[13] imagine that a player contemplates whether to accept the allocation x or appeal against x and demand some $x' \in \mathcal{X}$ which gives the player Δ more, a demand which will be accepted only with some probability q and result in disagreement with probability $1 - q$. The set $\mathcal{D}_i$ is the set of allocations that are unbalanced in the sense that player i would have higher incentives to make such demands than his opponent.

Note that, on an intuitive level, it seems natural that a player will be more willing to appeal an allocation x and demand some x' with $x'_i > x_i$ if x is an allocation that only gives player i a small share of the surplus. This suggests that if $x \in \mathcal{D}_i$, then for any allocation y with $y_i < x_i$ it must be that $y \in \mathcal{D}_i$. Lemma 3 in the appendix formally shows that, for each player $i \in \{1, 2\}$, there exists a number $\bar{x}_i = \sup_{x \in \mathcal{D}_i} x_i \in (0, \frac{1}{2}]$ such that the set $\mathcal{D}_i$ either satisfies $\mathcal{D}_i = \{x \in \mathcal{X} : x_i < \bar{x}_i\}$ or satisfies $\mathcal{D}_i = \{x \in \mathcal{X} : x_i \leq \bar{x}_i\}$.

Theorem 2. $\mathcal{X}^* = \mathcal{X}^{**} = \{x \in \mathcal{X} : x_1 \geq \bar{x}_1 \textit{ and } x_2 \geq \bar{x}_2\}$, *where* $\bar{x}_i = \sup_{x \in \mathcal{D}_i} x_i > 0$ *for* $i = 1, 2$.[14]

Proof. See Appendix. □

In particular, $\mathcal{X}^* = \mathcal{X}^{**}$ is a set valued solution concept that generalizes the equal division from Subsection 2.1, the Kalai-Smorodinsky solution from Subsection 2.2, the Nash Bargaining solution from Subsection 2.3, as well as any other single-valued bargaining solution which for the considered bargaining problem returns an allocation that is easiest to sustain for some deviation cost set $\mathcal{C}$.

The proof of Theorem 2 is in the appendix, here we just mention some basic ideas used in the proof. The definition of the sets $\mathcal{X}^*$ and $\mathcal{X}^{**}$ immediately implies that

[13] See Rubinstein, Safra, and Thomson (1992) for a characterization of the Nash Bargaining solution in similar terms.

[14] The sets $\mathcal{D}_i$ were defined in Definition 8.

$\mathcal{X}^* \subset \mathcal{X}^{**}$. Define $\bar{x}_1$ and $\bar{x}_1$ as in the theorem. Since $\mathcal{X}^* \subset \mathcal{X}^{**}$, to prove Theorem 2, it is enough to show that: (1) $\mathcal{X}^{**} \subset \{x \in \mathcal{X} : x_1 \geq \bar{x}_1 \text{ and } x_2 \geq \bar{x}_2\}$ and (2) $\{x \in \mathcal{X} : x_1 \geq \bar{x}_1 \text{ and } x_2 \geq \bar{x}_2\} \subset \mathcal{X}^*$.

To provide some intuition for the result, let us sketch why (1) holds. To prove (1) it is enough to show that for any allocation $x \in \mathcal{X}$ such that $x_i < \bar{x}_i$ for some player i, it is the case that $x \notin \mathcal{X}^*$. Now, note that if x is an allocation such that $x_i < \bar{x}_i$ then there is an allocation y such that $x_i < y_i < \bar{x}_i$. Since $x_i < \bar{x}_i$ and $y_i < \bar{x}_i$, both x and y lie in $\mathcal{D}_i$. But $\mathcal{D}_i$ was the set of allocations z, such that, given z, player i has strictly higher incentives to make demands than the other player j. This suggests that for the allocations x and y the sets $\mathcal{S}_{\mathcal{C}}(x)$ and $\mathcal{S}_{\mathcal{C}}(y)$ will be equal to the set of costs $c \in \mathcal{C}$ such that player i would not want to impose some alternative allocation – if i does not want to impose an alternative the same is true for the other player j as he has strictly weaker incentives to make demands.[15] But if $\mathcal{S}_{\mathcal{C}}(x)$ and $\mathcal{S}_{\mathcal{C}}(y)$ are both determined only by player i's incentives than we expect that since $x_i < y_i$, player i will more satisfied under y than under x and will have weaker incentives to make demands, and, therefore, $\mathcal{S}_{\mathcal{C}}(x) \subset \mathcal{S}_{\mathcal{C}}(y)$ for any $\mathcal{C}$ and $\mathcal{S}_{\mathcal{C}}(x) \subsetneq \mathcal{S}_{\mathcal{C}}(y)$ if $\mathcal{C}$ is sufficiently rich. This, however, implies $x \notin \mathcal{X}^{**}$. To prove statement (2) for each $x \in \{x \in \mathcal{X} : x_1 \geq \bar{x}_1 \text{ and } x_2 \geq \bar{x}_2\}$ a cost set $\mathcal{C}$ is constructed such that x is easiest to sustain for the cost set $\mathcal{C}$.

The set $\mathcal{X}^* = \mathcal{X}^{**}$ defined in Theorem 2 in general depends on the risk preferences of the two players. For instance, it is straightforward to show that, for the case where both players have the same preferences over lotteries, $\mathcal{X}^* = \mathcal{X}^{**} = \{(\frac{1}{2}, \frac{1}{2})\}$ holds. From Theorem 2 and the examples considered in Subsections 2.1 – 2.3 we know that, in general, the set $\mathcal{X}^* = \mathcal{X}^{**}$ is convex and contains the Nash bargaining solution, the Kalai-Smorodinsky solution, and the equal division $(\frac{1}{2}, \frac{1}{2})$. The reader might wonder how much larger $\mathcal{X}^* = \mathcal{X}^{**}$ is compared to the convex hull of those three allocations. The following proposition addresses this question.

[15] See Lemma 4 in the appendix for a formal statement.

Proposition 4. *For $i = 1, 2$, let $\bar{y}_i \in [0, 1]$ be the unique solution of*

$$\frac{\frac{u_i(2\bar{y}_i) - u_i(\bar{y}_i)}{\bar{y}_i}}{u_i(\bar{y}_i) - u_i(0)} = \frac{u_j'(1 - \bar{y}_i)}{u_j(1 - \bar{y}_i) - u_j(0)},$$

where j stands for the other player. Define $\mathcal{Y}^{**} \subset \mathcal{X}$ by

$$\mathcal{Y}^{**} = \{x \in \mathcal{X} : x_1 \geq \min(\bar{y}_1, \frac{1}{2}) \text{ and } x_2 \geq \min(\bar{y}_2, \frac{1}{2}).\}$$

Then, $\mathcal{X}^* = \mathcal{X}^{**} \subset \mathcal{Y}^{**}$.

Proof. See Appendix. □

Consider the equation defining $\bar{y}_i$. Note that if we replace $\frac{u_i(2\bar{y}_i) - u_i(\bar{y}_i)}{\bar{y}_i}$ with $u_i'(\bar{y}_i) \geq \frac{u_i(2\bar{y}_i) - u_i(\bar{y}_i)}{\bar{y}_i}$ we obtain an equation which characterizes the payoff of player i under the Nash bargaining solution.[16] Thus with $\mathcal{Y}^{**}$ we have an outer solution concept whose extreme points can be directly related to the Nash bargaining solution.

One can show that for the case where one of the players is risk neutral the set $\mathcal{Y}^{**}$ defined in Proposition 4 actually coincides with $\mathcal{X}^* = \mathcal{X}^{**}$. This implies that if player i is risk neutral then $\mathcal{X}^* = \mathcal{X}^{**} = \{x \in \mathcal{X} : x_j \in [\bar{y}_i, \frac{1}{2}]\}$ where $\bar{y}_i$ is defined as in Proposition 4.

2.8. **Comparative Statics.** By Theorem 2, $\mathcal{X}^* = \mathcal{X}^{**} = \{x \in \mathcal{X} : x_1 \geq \bar{x}_1 \text{ and } x_2 \geq \bar{x}_2\}$, where $\bar{x}_i = \sup_{x \in \mathcal{D}_i} x_i$ for $i = 1, 2$. The definition of the sets $\mathcal{D}_i$ implies some immediate comparative statics results.

Consider, for instance, how the set $\mathcal{X}^* = \mathcal{X}^{**}$ changes if player 1 would became more risk averse in the sense that his preference over risk is no longer given by u_1 but instead $\hat{u}_1 = v \circ u_1$, where v is an increasing, strictly concave function. The definition of the sets $\mathcal{D}_i$ together with Jensen's inequality immediately implies that $\bar{x}_1$ would weakly decrease and $\bar{x}_2$ would weakly increase. Thus, $\mathcal{X}^* = \mathcal{X}^{**}$ would "shift" in player 2's favor.

[16] The Nash bargaining is the allocation x that maximizes $(u_1(x_1) - u_1(0)) \cdot (u_2(x_2) - u_2(0)$. It therefore satisfies the first order condition $\frac{u_1'(x_1)}{u_1(x_1) - u_1(0)} = \frac{u_2'(x_2)}{u_2(x_2) - u_2(0))}$. Replacing x_j with $1 - x_i$ we obtain the equation.

3. Bargaining Between More than Two Individuals

3.1. **Coalitional Games.** When bargaining involves three or more individuals, a fundamentally new aspect may appear that is not present in bargaining between two individuals. It can be the case that there are some subgroups of players which can collaborate for mutual benefit among themselves, even if others choose not to participate. In this section, the focus will be on this aspect of multi-player bargaining.[17]

A *coalitional game with transferable payoffs* consists of a set of players $N = \{1, .., n\}$ and a function $v : 2^N \to [0, \infty)$ such that $v(\emptyset) = 0$.[18] The function v is called the *value function*. A non-empty subset of the set of players N will be called a *coalition*.

While coalitional games with transferable payoffs allow for a number of possible interpretations, we will for now narrowly interpret the value $v(S)$ for a coalition $S \subset N$ as the *monetary surplus* coalition S can achieve when acting on its own. Since the set N can be reconstructed from the value function v we will often identify a coalitional game with its value function.

We will restrict attention to games v such that, for any non-empty coalition $S \subset N$,

$$v(S) + v(N - S) \leq v(N). \tag{9}$$

This captures the idea that the grand coalition N can achieve at least the same monetary surplus as the sum of what S and $N - S$ could achieve on their own.

Note that, unlike in Section 2, we have not specified attitudes towards risk. In the following we will assume that players are risk neutral. This is so that we can focus our analysis on one single aspect (in this case coalitional bargaining) without introducing other factors. It will also make our results more easily comparable with other results obtained for coalitional games with transferable payoffs.

[17]It is possible to generalize the results from Section 2, to address the question how individual attitudes toward risk affect bargaining outcome if n players have to unanimously decide how to divide a dollar. However, that generalization does not involve anything conceptually new and the same methods can be applied that have been used to prove the results of Section 2.

[18]Here 2^N denotes the set of all subsets of N.

3.2. **Allocations.** Fix a game v. An *allocation* x is a vector in $[0,\infty)^N$ such that $\sum_{i\in N} x_i = v(N)$. An allocation describes a possible division of the surplus $v(N)$ among the players in the game v. The set of all allocations will be denoted $\mathcal{X}$. Whenever x is an allocation we will write x_S for $\sum_{i\in S} x_i$.

Let $\mathcal{C}$ be a deviation cost set as defined in Section 2. The next definition is a straightforward generalizations of Definition 1.

Definition 9. *An allocation $x \in \mathcal{X}$* can be sustained given costs $(p,m) \in C$ *if and only if*

$$x_S \geq (1-p(x'_S - x_S)) \cdot (x'_S - m(x'_S - x_S)) + p(x'_S - x_S) \cdot v(S) \tag{10}$$

holds for all non-empty $S \subset N$ and all allocations $x \in \mathcal{X}$ such that $x'_S \in (x_S, v(N) - v(N-S)]$. For any allocation $x \in \mathcal{X}$, denote the set of costs $c \in \mathcal{C}$ for which x can be sustained by $\mathcal{S}_{\mathcal{C}}(x)$.

Note that Definition 9 the cost of a deviation by a coalition does not depend on the size of the coalition. We think about this case as a very natural benchmark which, in particular, allows meaningful comparisons with solution concepts that implicitly assume that larger coalitions are not per se more or less effective in negotiations than smaller ones. At the same time, we would like to point out that, of course, the approach proposed in this paper can also be easily applied if the costs appearing in Definition 9 do, for instance, depend on the size of the coalition S.[19]

After having defined $\mathcal{S}_{\mathcal{C}}$ in Definition 9, we can apply Definitions 2 to 5 from Section 2 without any changes. In particular, we will say that an allocation $x \in \mathcal{X}$ *is easier to sustain than* an allocation $x' \in \mathcal{X}$ if and only if $\mathcal{S}_{\mathcal{C}}(x') \subsetneq \mathcal{S}_{\mathcal{C}}(x)$ and say that an allocation $x \in \mathcal{X}$ is *undominated* if and only if there is no allocation $x' \in \mathcal{X}$ that is easier to sustain. As in Section 2, the set of undominated allocations will be denoted by $\mathcal{U}_{\mathcal{C}}$.

[19]As an illustration, it is straightforward to generalize the arguments and results that follow to address the case where the chance of permanent disagreement does not depend on the size of the coalition S but the monetary costs are linear in the number of players that are in S.

3.3. **Nash Punishments.** Let $\mathcal{C}^{Nash}$ be the deviation cost set containing all pairs (p, m) such that $p : [0, 1] \to [0, 1]$ is differentiable with $p(0) = 0$, $p' > 0$, and $p'' > 0$ and $m : [0, 1] \to [0, \infty)$ satisfies $m \equiv 0$. This is the deviation cost set that was considered in Subsection 2.3 in which we obtained the symmetric Nash Bargaining solution as a unique prediction. The set $\mathcal{U}_{\mathcal{C}^{Nash}}$ can, therefore, be seen as the "analogue" of the Nash bargaining solution for coalitional games with transferable payoffs in the sense that the monitoring and sanctioning technology is the same as those that yielded the Nash bargaining solution in Subsection 2.3.

To characterize $\mathcal{U}_{\mathcal{C}^{Nash}}$, recall that *the core of a coalitional game* v is defined as

$$Core(v) = \{x \in \mathcal{X} : x_S \geq v(S) \text{ for all } S \subset N\}.$$

On an intuitive level, an allocation is in the core if and only if there is no coalition that could achieve a higher payoff on its own.

Proposition 5. *If $Core(v)$ is empty, then $\mathcal{U}_{\mathcal{C}^{Nash}} = \mathcal{X}$. If $Core(v)$ is non-empty, then the set $\mathcal{U}_{\mathcal{C}^{Nash}}$ consists of exactly those allocations $x \in Core(v)$ that solve the problem*

$$\max_{x \in Core(\mathcal{X})} \min_{S} x_S - v(S),$$

where the minimum is over all coalitions $S \subset N$ such that $v(S) + v(N - S) < v(N)$.

Proof. See Appendix. □

Remark 4. *The set $\mathcal{U}_{\mathcal{C}^{Nash}}$ always contains the nucleolus of the game v.*[20] *In particular, whenever $\mathcal{U}_{\mathcal{C}^{Nash}}$ is a singleton, the unique element of $\mathcal{U}_{\mathcal{C}^{Nash}}$ must be the nucleolus.*

The fact that $\mathcal{U}_{\mathcal{C}^{Nash}} = \mathcal{X}$ if the core of the game v is empty, is due to the fact that all the costs in $\mathcal{C}^{Nash}$ are non-monetary in the sense that for any $(p, m) \in \mathcal{C}^{Nash}$, it is the case that $m \equiv 0$. To see this more generally, let $\mathcal{C}$ be an arbitrary *non-monetary deviation cost set*, i.e. a deviation cost set such that, for every $(p, m) \in \mathcal{C}$, it is the case

[20]See Schmeidler (1969) for a definition of the nucleolus of a coalitional game with transferable payoffs.

that $m \equiv 0$. Now, think about a coalition S that, for a given $(p, m) \in \mathcal{C}$, contemplates whether it prefers to receive x_S with certainty or prefers to try to impose an allocation x' with $x'_S > x_S$. Since $m \equiv 0$, the only "downside" of trying to impose such an x' is that sometimes the coalition will receive $v(S)$ instead of x'_S. But this means that, if it happens to be that $x_S < v(S)$, coalition S w will always find it in its interest to try to impose such an x'.[21] Thus $\mathcal{S}_{\mathcal{C}}(x) = \emptyset$ whenever $x \notin Core(v)$. This immediately implies that $\mathcal{U}_{\mathcal{C}} = \mathcal{X}$ if the game v has an empty core.

To get a sense of the above result for the case where the core is non-empty, consider the following *factory game* as an example. A capitalist owning a factory and n workers are bargaining how to divide a surplus that they can jointly create. A coalition consisting of the capitalist and k workers, can on its own produce a surplus of $k \cdot \theta$, where $\theta > 0$ is the marginal product of an additional worker. A coalition that does not include the capitalist can not produce any surplus. To define a formal coalitional game, let $N^{fact} = \{1, 2, ..., n, n+1\}$ be the set of players, where we think of players $1, 2, ..., n$ as workers and player $n+1$ as the capitalist, and let $v^{fact} : 2^N \to [0, \infty)$ be given by $v^{fact}(S) = (|S| - 1) \cdot \theta$ if $n + 1 \in S$ and $v^{fact}(S) = 0$ if $n + 1 \notin S$. For this game, the core is given by

$$Core(v^{fact}) = \{x \in \mathcal{X} : x_i \leq \theta \text{ for } i \in \{1, 2, ..., n\}\}.$$

In contrast, $\mathcal{U}_{\mathcal{C}^{Nash}}$ yields a much sharper prediction as

$$\mathcal{U}_{\mathcal{C}^{Nash}}(v^{fact}) = \{x^*\} \tag{11}$$

where $x^* \in \mathcal{X}$ is the allocation in which each worker $i \in \{1, .., n\}$ receives a payoff of $w^* = \frac{\theta}{2}$ and the capitalist receives $f(n) - n \cdot w^*$. To prove (11) note first that

$$\min_{S: v(S) + v(N-S) < v(N)} x^*_S - v(S) = w^*.$$

[21] Note that, if $x(S) < v(S)$, then $x(S) < v(N) - v(N - S)$ by (9) and thus there will exist x'_S in the required interval $(x_S, v(N) - v(N - S)]$.

Therefore, (11) will follow from Proposition 5 if we show that for each $x \neq x^*$

$$\min_{S:v(S)+v(N-S)<v(N)} x_S - v(S) < w^*. \tag{12}$$

Consider an allocation $x \neq x^*$. Note that there must be a worker i such that $x_i \neq w^*$.[22] However, if there is a worker i who receives $x_i < w^*$ then for $S = \{i\}$ we have $x_S - v(S) = x_i < w^*$ which implies (12). If, on the other hand, there is a worker i who receives $x_i > w^*$, then for $S = N - \{i\}$ we have $x_S - v(S) = (n \cdot \theta - x_i) - (n-1) \cdot \theta < w^*$ which again implies (12). Since we have shown that (12) holds for any allocation $x \neq x^*$, (11) follows from Proposition 5.

In this example, our approach yielded much sharper predictions than the core concept, predicting that all workers will receive the same wage and pinning down the exact size of this wage. Of course, the reader may wonder to what extend that prediction depended on the nature of the assumed costs. The next two subsections address this question.

3.4. **The sets $\mathcal{X}^*(v)$ and $\mathcal{X}^{**}(v)$.** The definition of the sets $\mathcal{X}^*$ and $\mathcal{X}^{**}$ from Subsection 2.6 can be immediately extended to the current setting.

Definition 10. *Fix a coalitional game v. Let $\mathcal{X}^*(v)$ be the set of all allocations $x^* \in \mathcal{X}$ such that there is a deviation cost set $\mathcal{C}$ with the property that x^* is easiest to sustain, i.e. such that $\mathcal{S}_{\mathcal{C}}(y) \subsetneq \mathcal{S}_{\mathcal{C}}(x^*)$ for any $y \in \mathcal{X}$ such that $y \neq x^*$.*

Definition 11. *Fix a coalitional game v. Let $\mathcal{X}^{**}(v)$ be the set of all allocations $x \in \mathcal{X}$ such that there is no allocation $y \in \mathcal{X}$ with the property that, for any cost set $\mathcal{C}$, y is either easier or equally easy to sustain as x and, for some deviation cost set $\mathcal{C}$, y is easier to to sustain than x.*

In the following subsection we will characterize the set $\mathcal{X}^{**}(v)$. In the context of general coalitional games $\mathcal{X}^*(v)$ is a less interesting object than for the bargaining problem considered in Section 2 because for many natural deviation cost sets there

[22] If all workers would receive the same payoff as in x^* then also the capitalist would receive the same payoff as in x^* as payoffs must add up to $v(N) = f(n)$.

will be no unique allocation that is easiest to sustain. This is not only the case for the Nash Punishments considered in the last subsection. As a matter of fact, there are games v for which $\mathcal{X}^{**}(v)$ is empty, i.e. there is no deviation cost set $\mathcal{C}$ for which there is an allocation that is easiest to sustain.[23] For this reason, we will focus our attention on the set $\mathcal{X}^{**}(v)$.

3.5. **Characterization of $\mathcal{X}^{**}(v)$.** Before we can state Theorem 3 that provides a characterization of the set $\mathcal{X}^{**}(v)$ we need to introduce some notation.

Definition 12. *We will say that an allocation $x \in \mathcal{X}$ is* weakly less extreme than *an allocation $y \in \mathcal{X}$ if and only if for any coalition S such that $x_S - v(S) < x_{N-S} - v(N-S)$ there exists a coalition $S' \subset N$ such that*

$$y_{S'} - v(S') \leq x_S - v(S) \text{ and } x_{N-S} - v(N-S) \leq y_{N-S'} - v(N-S').$$

An allocation $x \in \mathcal{X}$ is equally extreme as *$y \in \mathcal{X}$ if and only if x is weakly less extreme than y and y is weakly less extreme than x. An allocation $x \in \mathcal{X}$ is* strictly less extreme than *$y \in \mathcal{X}$ if and only if x is weakly less extreme than y and not equally extreme as y.*

To understand Definition 12, it is worthwhile to think about a game v as representing a framework where any coalition S can bargain with coalition $N-S$ on how to divide the surplus $v(N) - v(N-S) - v(S)$ which remains if S receives $v(S)$ and $N-S$ receives $v(N-S)$. Then, $x_S - v(S) < x_{N-S} - v(N-S)$ means that, under allocation x, the surplus $x_S - v(S)$ received by coalition S is strictly smaller than the surplus $x_{N-S} - v(N-S)$ received by coalition $N-S$. An allocation x is weakly less extreme than an allocation y if, for any case where the division between S and $N-S$ under x is "unfair", i.e. $x_S - v(S) < x_{N-S} - v(N-S)$ we can find a coalition

[23] An example of such a game is the 4 player game with set of players $N = \{1,2,3,4\}$ and valuation function v given by:

$$v(S) = \begin{cases} 1 \text{ if } S = \{1\},\ S = \{2,3\},\ \text{or } S = \{1,2,3,4\}, \\ 0 \text{ if } S = \emptyset, S = \{1,4\},\ \text{or } S = \{2,3,4\} \\ \frac{1}{2} \text{ otherwise.} \end{cases}$$

The proof that $\mathcal{X}^{*}(v) = \emptyset$ is left as a fun exercise for the reader.

S' such that under y the division between S' and $N - S'$ is "equally or more unfair" in the sense that that coalition S' under y gets even less surplus than S under x (i.e. $y_{S'} - v(S') \leq x_S - v(S)$) and $N - S'$ gets even more surplus under y than $N - S$ under x (i.e. $x_{N-S} - v(N - S) \leq y_{N-S'} - v(N - S')$). An allocation x is strictly less extreme than y if and only if the above holds and in addition there are coalitions S' such that $y_{S'} - v(S') < y_{N-S'} - v(N - S')$ but there is no coalition S for which the division under x would be "equally or more unfair", i.e. for which $x_S - v(S) \leq y_{S'} - v(S')$ and $y_{N-S'} - v(N - S') \leq x_{N-S} - v(N - S)$ both hold.

Theorem 3. *The set $\mathcal{X}^{**}(v)$ is equal to set of allocations $x \in \mathcal{X}$ such that there exist no $y \in \mathcal{X}$ that is strictly less extreme than x.*

Proof. The proof is in the appendix. □

The proof of the theorem is in the appendix. Two key observations are Lemma 8 and Lemma 9. Together, those to lemmas imply that, for any allocations $x, y \in \mathcal{X}$, x is weakly less extreme than y if and only if $\mathcal{S}_{\mathcal{C}}(y) \subset \mathcal{S}_{\mathcal{C}}(x)$ holds for all deviation cost sets C.

It is interesting to consider the set $\mathcal{X}^{**}(v)$ for the case where the core of a game is empty. An allocation that is not in the core is usually regarded as "unstable".[24] The monetary components in the deviation cost sets allow us to quantify exactly how "unstable" allocations outside the core are and, in particular, Theorem 3 can yield tight predictions also for games that have an empty core.

As a trivial illustration consider a "majority game" between 5 political parties active in a parliament with 100 seats, where party 1 controls 40 seats in parliament and parties 2, 3, 4, and 5 control 15 seats each. Let $N = \{1, 2, 3, 4, 5\}$ and define the value function v^{maj} by $v^{maj}(S) = 1$ if the sum of the seats controlled by parties in S is larger than 50 and $v^{maj}(S) = 0$ otherwise. Since, for any coalition S the number of controlled seats is either larger or smaller than 50, for any coalition S we have

[24] See, for instance, Perry and Reny (1994).

$v^{maj}(S) + v^{maj}(N-S) = v^{maj}(N)$ and, therefore,

$$x_S - v^{maj}(S) = v^{maj}(N) - x_{N-S} - v^{maj}(S) = -(x_{N-S} - v^{maj}(N-S)).$$

This means that for the game v^{maj}, the set of allocations x such that there is no y that is strictly less extreme than x (which, by Theorem 3 is equal to $\mathcal{X}^{**}(v)$) is equal to the set of allocations $x \in \mathcal{X}$ that maximize

$$\min_S x_S - v^{maj}(S).$$

It is straightforward to check that, for the concrete numbers of seats given above, this problem has a unique solution in which party 1 receives $\frac{3}{7}$ and each of the smaller parties receives $\frac{1}{7}$.

3.6. **Purely Monetary Costs.** The reader may wonder what results can be obtained for the case where the deviation cost set $\mathcal{C}$ is *purely monetary* in the sense that any $(p, m) \in \mathcal{C}$ satisfies $p \equiv 0$.

For any coalitional game v, the *dual game* $v^* : 2^N \to [0, \infty)$ is defined by

$$v^*(S) = v(N) - v(N-S).$$

Note that the above definition together with (9) implies that, for any coalition S, $v^*(S) \geq v(S)$. In particular, $Core(v^*) \subset Core(v)$ must hold.

Given that we already derived a number of similar results, it is not hard to verify the following.

Proposition 6. *Let $\mathcal{C}^{mon}$ be the set of pairs (p, m) where $p : [0,1] \to [0,1]$ satisfies $p \equiv 0$ and $m : [0,1] \to [0, \infty)$ is a continuous function.*

If $Core(v^*) \neq \emptyset$, then $\mathcal{U}_{\mathcal{C}^{mon}} = Core(v^*) \subset Core(v)$. If $Core(v^*) = \emptyset$ then $\mathcal{U}_{\mathcal{C}^{mon}}$ is equal to the set of allocations maximizing $\min_{S \subset N} x_S - v^*(S)$.

Proof. Straightforward. □

As an example, consider again the "majority game" from the last subsection. Note that, for that game we have $(v^{maj})^*(S) = v^{maj}(S)$, and thus Proposition 6 implies that

$\mathcal{U}_{\mathcal{C}^{mon}}$ is equal to the set of allocations maximizing $\min_{S \subset N} x_S - v^{maj}(S)$. In particular, for the concrete numbers considered in the last subsection, $\mathcal{U}_{\mathcal{C}^{mon}}$ is a singleton and consists of the allocation in which party 1 receives $\frac{3}{7}$ and all other parties receive $\frac{1}{7}$.

4. Conclusion

This paper studied bargaining outcomes in societies where cooperation is governed by social norms. It proposed a new approach in which bargaining outcomes were analyzed based on how difficult they are to sustain as part of a social norm. It then used two classes of bargaining problems – bargaining between two players which differed in their attitude towards risk in Section 2 and the coalitional bargaining problems between many risk neutral players in Section 3 – to explore the link between the way social norms are sustained and the type of bargaining outcomes that can be expected.

Understanding the relationship between bargaining outcomes when cooperation is governed by social norms and the way in which social norms are sustained is not only interesting when one knows something about how social norms are enforced and wants to make predictions about bargaining outcomes. By their very nature, mechanisms sustaining a social norm may be very hard to observe. This is clear if norms are internalized. However, also if norms are sustained through sanctions, it typically will be difficult to observe those sanctions directly, given that nobody will be actually sanctioned if everybody adheres to the norm. Thus, understanding the relationship between bargaining outcomes and the way social norms are enforced may also be helpful if one can observe bargaining outcomes that are part of a social norm and wishes to understand better how the underlying social norm is sustained through sanctions. Finally, note that understanding this relationship may allow us to link behavior in bargaining situations that a priori appear very different like the two-player bargaining problems considered in Section 2 and the coalitional bargaining problems considered in Section 3 if the norms governing those situations are sustained in a similar way. As a matter of fact, we saw this when we derived "the analogues" of

the Nash Bargaining solution, the Kalai-Smorodinsky solution, or the equal monetary split for coalitional games with transferable payoffs.

It has been observed by a number of authors that social norms play a less important role in the economic literature than in some of the other social sciences. For example, Elster (1989) writes:

> One of the most persistent cleavages in the social sciences is the opposition between two lines of thought conveniently associated with Adam Smith and Emilie Durkheim, between *homo economicus* and *homo sociologicus*. Of these, the former is supposed to be guided by instrumental rationality, while the behavior of the latter is dictated by social norms. The former is "pulled" by the prospect of future rewards, whereas the latter is "pushed" from behind by quasi-inertial forces (Gambetta, 1987). [...] The former is easily caricatured as a self-contained, asocial atom, and the latter as the mindless plaything of social forces.

This paper is motivated by the idea, eloquently expressed by Arrow (1971) in the quote given in in the introduction, that many norms may be a "reaction of society to compensate for market failure", i.e. that the function of social norms is to overcome inefficiencies that would occur given that agents are individually rational.[25] Thus, in the language of Elster, here norms – whether internalized or enforced through outside sanctions – are used to turn the ingenious *homo economicus* who at every turn is cleverly looking which action may be most advantageous into a *homo sociologicus* who just mindlessly follows the social norm and has given up any hope that doing something unorthodox may be to his or her advantage. Nevertheless, the approach proposed in this paper shows how the norms that govern the *homo sociologicus* may be related to the individual incentive problems as we ask for which allocations it is

[25]This is not a function that is typically assigned to social norms in the modern economic literature where social norms - if studied at all - are often analyzed as equilibrium selection problems of non-cooperative games. For instance, in the first sentence of the entry on social norms in the New Palgrave Dictionary of Economics, Young (2008) says "The function of a social norm is to coordinate people's expectations in interactions that possess multiple equilibria".

easiest to turn that *homo economicus* into a *homo sociologicus*. Thus it may allow us economists to say something about the norms the *homo sociologicus* follows despite the fact that he or she may very much look like a "mindless plaything of social forces".

Mathematical Appendix

4.1. **Proof of Theorem 1.** By Remark 2 it is enough to prove the theorem for the case where $u_i(0) = 0$ for $i \in \{1, 2\}$.

For any player $i \in \{1, 2\}$ and allocation $x \in \mathcal{X}$, let $\mathcal{S}^i_{\mathcal{C}}(x)$ be the set of cost parameters $(p, m) \in \mathcal{C}$ such that

$$u_i(x_i) \geq (1 - p(x'_i - x_i)) \cdot u_i(x'_i - m(x'_i - x_i)) \tag{13}$$

holds for all allocations $x' \in \mathcal{X}$ with $x'_i > x_i$. On an intuitive level, $\mathcal{S}^i_{\mathcal{C}}(x)$ is exactly equal to the set of cost parameters for which player i would not want to impose some alternative outcome $x' \in \mathcal{X}$. The definition of $\mathcal{S}_{\mathcal{C}}(x)$ immediately implies that $\mathcal{S}_{\mathcal{C}}(x) = \mathcal{S}^1_{\mathcal{C}}(x) \cap \mathcal{S}^2_{\mathcal{C}}(x)$.

Lemma 1. *For any $c \in \mathcal{C}$, the set $\{x \in \mathcal{X} : c \notin \mathcal{S}^i_{\mathcal{C}}(x)\}$ is an open subset of $\mathcal{X}$.*

Proof of Lemma 1. Let $c = (p, m) \in \mathcal{C}$ and $x \in \mathcal{X}$. $c \notin \mathcal{S}^i_{\mathcal{C}}(x)$ means that there is an $x' \in \mathcal{X}$ with $x'_i > x_i$ such that

$$u_i(x_i) < (1 - p(x'_i - x_i)) \cdot u_i(x'_i - m(x'_i - x_i)).$$

Consider the above expression if we replace x_i with some other value $\hat{x}_i$ that is close enough to x_i so that $x'_i > \hat{x}_i$. Note that since the functions u_i, p, and m are all continuous, if the above inequality does hold for some x it will also hold if we replace x with any allocation $\hat{x}_i$ that is sufficiently close to x. □

The next lemma formalizes the intuition that a player who receives more will have smaller incentives to impose an alternative allocation.

Lemma 2. *Let $i \in \{1, 2\}$. If $x, y \in \mathcal{X}$ satisfy $x_i < y_i$ then $\mathcal{S}^i_{\mathcal{C}}(x) \subset \mathcal{S}^i_{\mathcal{C}}(y)$.*

Proof of Lemma 2. Let $x, y \in \mathcal{X}$ such that $x_i < y_i$. We need to show that $\mathcal{S}^i_{\mathcal{C}}(x) \subset \mathcal{S}^i_{\mathcal{C}}(y)$. Assume that this is not the case, i.e. there exists a $c = (m, p) \in \mathcal{C}$ such that $c \in \mathcal{S}^i_{\mathcal{C}}(x)$ and $c \notin \mathcal{S}^i_{\mathcal{C}}(y)$. Since $c \notin R_i(y)$, there exists an $y' \in \mathcal{X}$ with $y'_i > y_i$ such that

$$u_i(y_i) < (1 - p(y'_i - y_i)) \cdot u_i(y'_i - m(y'_i - y_i))$$

or, equivalently,[26]

$$p(y'_i - y_i) < \frac{u_i(y'_i - m(y'_i - y_i)) - u_i(y_i)}{u_i(y'_i - m(y'_i - y_i))}. \tag{14}$$

Let $x' \in \mathcal{X}$ be given by $x'_i - x_i = y'_i - y_i$.[27] Note that, the fact that u_i is increasing implies that $u_i(x' - m(x'_i - x_i)) < u_i(y' - m(y'_i - y_i))$ and the fact that u_i is concave implies $u_i(y'_i - m(y'_i - y_i)) - u_i(y_i) < u_i(x'_i - m(x'_i - x_i)) - u_i(x_i)$. Thus, inequality (14) implies

$$p(x'_i - x_i) < \frac{u_i(x'_i - m(x'_i - x_i)) - u_i(x_i)}{u_i(x'_i - m(x'_i - x_i))}.$$

or, equivalently,

$$u_i(x_i) < (1 - p(x'_i - x_i)) \cdot u_i(x'_i - m(x'_i - x_i))$$

which contradicts $c \in R_i(x)$. □

Define the binary relation $\succeq_{\mathcal{C}}$ on $\mathcal{X}$ by the requirement that, for any $x, y \in \mathcal{X}$, $x \succeq_{\mathcal{C}} y$ if and only if either $x \succ_{\mathcal{C}} y$ or $x = y$. The fact that $\succ_{\mathcal{C}}$ is a strict partial order immediately implies that $\succeq_{\mathcal{C}}$ is a partial order.

To prove the theorem it is enough to show that $\succeq_{\mathcal{C}}$ has a maximal element. By the Kuratowski-Zorn lemma, this will be the case if every subset $A \subset \mathcal{X}$ that is totally ordered with respect to $\succeq_{\mathcal{C}}$ has an upper bound.

Let $A \subset \mathcal{X}$ be a totally ordered subset of $\mathcal{X}$, i.e. a set such that for any $x, y \in A$ either $x \succeq_{\mathcal{C}} y$ or $y \succeq_{\mathcal{C}} x$. We will show that the set A has an upper bound, i.e. there exists an allocation $z \in \mathcal{X}$ such that $z \succeq_{\mathcal{C}} x$ for all $x \in A$. Since the case where A is empty is trivial, consider the case where A is non-empty.

[26] Note that the last inequality implies $u_i(y'_i - m(y'_i - y_i)) > 0$.
[27] Such x' exists since we $x_i < y_i$.

Define $R = \cup_{x \in A} \mathcal{S}(x)$. If A contains an element z with $\mathcal{S}_{\mathcal{C}}(z) = R$ this element z is an upper bound for A.[28] Assume, therefore, that this is not the case, i.e.

$$R(x) \subsetneq R \text{ for all } x \in A. \tag{15}$$

For $i = 1, 2$ define[29]

$$\bar{x}_i = \inf\{x_i : x \in \mathcal{X} \text{ and } R \subset \mathcal{S}^i_{\mathcal{C}}(x)\}.$$

Note that by Lemma 1, for $i \in \{1, 2\}$ and $x \in \mathcal{X}$, $x_i = \bar{x}_i$ implies $R \subset \mathcal{S}^i_{\mathcal{C}}(x_i)$.[30] Lemma 2 (together with the definition of $\bar{x}_i$) now implies that for $i \in \{1, 2\}$ and $x \in \mathcal{X}$, $R \subset \mathcal{S}^i_{\mathcal{C}}(x)$ if and only if $x_i \geq \bar{x}_i$.

Note next that

$$\bar{x}_1 + \bar{x}_2 \leq 1.$$

To see that this is indeed the case, assume $\bar{x}_1 + \bar{x}_2 > 1$ and let $\varepsilon = \frac{\bar{x}_1 + \bar{x}_2 - 1}{2}$. Then, for any $x \in \mathcal{X}$ it is either the case that $x_1 \leq \bar{x}_1 - \varepsilon$ or $x_2 \leq \bar{x}_2 - \varepsilon$. The definition of $\bar{x}_1$ implies that there exists a $c_1 \in R$ such that $c_1 \notin \mathcal{S}^1_{\mathcal{C}}((\bar{x}_1 - \varepsilon, 1 - (\bar{x}_1 - \varepsilon)))$. Similarly, definition of $\bar{x}_2$ implies that there exists a $c_2 \in R$ such that $c_2 \notin \mathcal{S}^2_{\mathcal{C}}((1 - (\bar{x}_2 - \varepsilon), \bar{x}_2 - \varepsilon)))$. Now, since $c_1, c_2 \in R$ and R was defined as $R = \cup_{x \in A} \mathcal{S}(x)$, there must exist an $x^1 \in A$ with $c_1 \in \mathcal{S}_{\mathcal{C}}(x^1)$ and there must exist an $x^2 \in A$ with $c_2 \in \mathcal{S}_{\mathcal{C}}(x^2)$. Since A is a totally ordered subset of $\mathcal{X}$, it must be the case that either $\mathcal{S}_{\mathcal{C}}(x^1) \subset \mathcal{S}_{\mathcal{C}}(x^2)$ or $\mathcal{S}_{\mathcal{C}}(x^2) \subset \mathcal{S}_{\mathcal{C}}(x^2)$. Assume $\mathcal{S}_{\mathcal{C}}(x^1) \subset \mathcal{S}_{\mathcal{C}}(x^2)$ holds. (The argument if $\mathcal{S}_{\mathcal{C}}(x^2) \subset \mathcal{S}_{\mathcal{C}}(x^2)$ holds is analogous.) $\mathcal{S}_{\mathcal{C}}(x^1) \subset \mathcal{S}_{\mathcal{C}}(x^2)$ together with $c_1 \in \mathcal{S}_{\mathcal{C}}(x^1)$ and $c_2 \in \mathcal{S}_{\mathcal{C}}(x^2)$ implies that $\{c_1, c_2\} \subset \mathcal{S}_{\mathcal{C}}(x^2)$. This, however, is not possible as, for any $x \in \mathcal{X}$, it is either the case that $x_1 \leq \bar{x}_1 - \varepsilon$ or $x_2 \leq \bar{x}_2 - \varepsilon$ and, therefore (by Lemma 2) for any

[28]Since A is totally ordered and $z \in A$, for any $x \in A$ it has to be the case that either $x \succeq_{\mathcal{C}} z$ or $z \succeq_{\mathcal{C}} x$. Since $R = \cup_{x \in A} \mathcal{S}(x)$ and $\mathcal{S}_{\mathcal{C}}(z) = R$, $x \succeq_{\mathcal{C}} z$ can never be true for $x \in A - \{z\}$. Thus, it must be that $z \succeq_{\mathcal{C}} x$ for all $x \in A$.

[29]To see that the set $\{x_i : x \in \mathcal{X} \text{ and } R \subset \mathcal{S}^i_{\mathcal{C}}(x)\}$ is nonempty and, therefore, the infimum is well defined note that $\mathcal{S}_1((1, 0)) = \mathcal{S}_2((0, 1)) = \mathcal{X}$ which means that, for any $R \subset \mathcal{X}$, $1 \in \{x_i : x \in \mathcal{X} \text{ and } R \subset \mathcal{S}^i_{\mathcal{C}}(x)\}$.

[30]Indeed, if there was a $c \in R$ such that $c \notin \mathcal{S}^i_{\mathcal{C}}(x_i)$, then by Lemma 1 there is a neighbourhood V of x such that, for $y \in V$, $c \notin \mathcal{S}^i_C(y)$ and, therefore, $R \not\subset \mathcal{S}^i_C(y)$. This would contradict the definition of $\bar{x}_i$ as $\inf\{x_i : x \in \mathcal{X} \text{ and } R \subset \mathcal{S}^i_{\mathcal{C}}(x)\}$.

$x \in \mathcal{X}$ either $c_1 \notin \mathcal{S}^1_{\mathcal{C}}(x)$ (and thus $c_1 \notin \mathcal{S}_{\mathcal{C}}(x) = \mathcal{S}^1_{\mathcal{C}}(x) \cap \mathcal{S}^2_{\mathcal{C}}(x)$) or $c_2 \notin \mathcal{S}^2_{\mathcal{C}}(x)$ (and thus $c_2 \notin \mathcal{S}_{\mathcal{C}}(x) = \mathcal{S}^1_{\mathcal{C}}(x) \cap \mathcal{S}^2_{\mathcal{C}}(x)$).

Let $z \in \mathcal{X}$ be any allocation such that $z_1 \geq \bar{x}_1$ and $z_2 \geq \bar{x}_2$. Since, $R \subset \mathcal{S}^i_{\mathcal{C}}(x_i)$ if and only if $x_i \geq \bar{x}_i$, we can conclude that $R \subset \mathcal{S}(z)$. Then (15) yields that $z \succ_{\mathcal{C}} x$ for all $x \in A$.

Since we have shown that every chain has an upper bound, the Kuratowski-Zorn Lemma implies that $\succeq_{\mathcal{C}}$ (and therefore also $\succ_{\mathcal{C}}$) has a maximal element.

4.2. **Proof of Theorem 2.** As noted in the main part of the paper, the definition of the sets $\mathcal{X}^*$ and $\mathcal{X}^{**}$ immediately implies that $\mathcal{X}^* \subset \mathcal{X}^{**}$. Define $\bar{x}_1$ and $\bar{x}_1$ as in the statement of the theorem. Since $\mathcal{X}^* \subset \mathcal{X}^{**}$, to prove Theorem 2, it is enough to show that:

(1) $\mathcal{X}^{**} \subset \{x \in \mathcal{X} : x_1 \geq \bar{x}_1 \text{ and } x_2 \geq \bar{x}_2\}$ and
(2) $\{x \in \mathcal{X} : x_1 \geq \bar{x}_1 \text{ and } x_2 \geq \bar{x}_2\} \subset \mathcal{X}^*$.

By Remark 2 it is enough to prove (1) and (2) for the case where $u_i(0) = 0$ for $i \in \{1, 2\}$. We start with a lemma that characterizes the sets $\mathcal{D}_i$ for $i = 1, 2$.

Lemma 3. *Let $i \in \{1, 2\}$. There exists a number $\bar{x}_i \in (0, \frac{1}{2}]$ such that either*

$$\mathcal{D}_i = \{x \in \mathcal{X} : x_i < \bar{x}_i\}$$

or

$$\mathcal{D}_i = \{x \in \mathcal{X} : x_i \leq \bar{x}_i\}.$$

In particular, the set $\mathcal{D}_i$ is non-empty.

Proof of Lemma 3. Let $i \in \{1, 2\}$. We will organize the argument in several steps.

Step 1: Note that the definition of the set $\mathcal{D}_i$ immediately implies that $\mathcal{D}_i$ contains the allocation x with $x_i = 0$. Since $\mathcal{D}_i \subset \mathcal{X}$ is non-empty, we can define $\bar{x}_i$ by $\bar{x}_i = \sup_{x \in \mathcal{D}_i} x_i$.

Step 2: Note that if $x \in \mathcal{D}_i$, then also $x' \in \mathcal{D}_i$ for any $x' \in \mathcal{X}$ with $x'_i < x_i$. To see that this is the case, consider the definition of the set $\mathcal{D}_i$ and note that, for $q \in (0, 1)$

and $\Delta \in [0,1]$,

$$u_i(x_i) < q \cdot u_i(x_i + \Delta) + (1-q) \cdot u_i(0)$$

is equivalent to

$$\frac{u_i(x_i) - u_i(0)}{u_i(x_i + \Delta) - u_i(x_i)} < q$$

and

$$u_j(x_j) \leq q \cdot u_j(x_j + \Delta) + (1-q) \cdot u_j(0)$$

is equivalent to

$$\frac{u_j(x_j) - u_j(0)}{u_j(x_j + \Delta) - u_j(x_j)} \leq q$$

Our claim now follows immediately from the observation that, since the utility functions of both players are increasing and concave, for $k \in \{1,2\}$, $u_k(x_k + \Delta) - u_k(x_k)$ is non-increasing in x_k and $u_k(x_k) - u_k(0)$ is increasing in x_k.

Steps 1 and 2 together imply that for $\bar{x}_i = \sup_{x \in \mathcal{D}_i} x_i$ either $\mathcal{D}_i = \{x \in \mathcal{X} : x_i < \bar{x}_i\}$ or $\mathcal{D}_i = \{x \in \mathcal{X} : x_i \leq \bar{x}_i\}$ and $\bar{x}_i \geq 0$. All that remains to be shown is that $0 < \bar{x}_i \leq \frac{1}{2}$.

Step 3: To show that $\bar{x}_i \leq \frac{1}{2}$, assume that is not the case and let $x \in \mathcal{D}_i$ be an allocation with $x_i > \frac{1}{2}$. Let $j \in \{1,2\}$ with $j \neq i$. Note that for $\Delta = x_i$ and $q = u_j(x_j)/u_j(x_j + \Delta)$ we have $x_j + \Delta = 1 \leq 1$ and $u_j(x_j) = q \cdot u_j(x_j + \Delta)$. Since $x \in \mathcal{D}_i$, this implies that $x_i + \Delta = 2 \cdot x_i \leq 1$. This, however, contradicts $x_i > \frac{1}{2}$.

Step 4: To show that $\bar{x}_i > 0$, note that for any allocation x such that $x_i < \frac{1}{2}$ and $u_i(x_i) - u_i(0) < \frac{u_i'(1)}{u_j'(0)} \cdot (u_j(1) - u_j(0))$ it will be the case that $x \in \mathcal{D}_i$.

□

For any deviation cost set $\mathcal{C}$, define $\mathcal{S}^1_{\mathcal{C}}$ and $\mathcal{S}^2_{\mathcal{C}}$ as in the proof of Theorem 1. The next lemma relates the sets $\mathcal{D}_i$ to $\mathcal{S}^1_{\mathcal{C}}$ and $\mathcal{S}^2_{\mathcal{C}}$. This lemma is the key observation in the proof of statement (1).

Lemma 4. *Fix a deviation cost set $\mathcal{C}$. Let $i, j \in \{1,2\}$ with $i \neq j$. For any $x \in \mathcal{D}_i$ it is the case that*

$$\mathcal{S}^i_{\mathcal{C}}(x) \subset \mathcal{S}^j_{\mathcal{C}}(x).$$

Proof of Lemma 4. We will prove the statement in the lemma for the case where $i = 1$ and $j = 2$. The argument for the case where $j = 1$ and $i = 2$ is analogous. To show that for any $x \in \mathcal{D}_1$ it is the case that

$$\mathcal{S}^1_{\mathcal{C}}(x) \subset \mathcal{S}^2_{\mathcal{C}}(x)$$

it is enough to show that $(p, m) \notin \mathcal{S}^2(y)$ implies $(p, m) \notin \mathcal{S}^1(y)$. Assume, therefore, $(p, m) \notin \mathcal{S}^2(x)$.

Since, $(p, m) \notin \mathcal{S}^2(x)$ there must exist a $x_2' \in (x_2, 1]$ such that

$$u_2(x_2) < (1 - p(x_2' - x_2)) \cdot u_2(x_2' - m(x_2' - x_2)). \tag{16}$$

Note that (16) implies that $x_2' - m(x_2' - x_2) > x_2$. Set $\Delta = x_2' - m(x_2' - x_2) - x_2$ and $q = 1 - p(x_2' - x_2)$. Note that $x_2' - m(x_2' - x_2) \leq 1$ implies $x_2 + \Delta \leq 1$. We can now rewrite (16) as

$$u_2(x_2) < q \cdot u_2(x_2 + \Delta).$$

Since $x \in \mathcal{D}_1$, the last inequality together with $x_2 + \Delta \leq 1$ implies that $x_1 + \Delta \leq 1$ and

$$u_1(x_1) < q \cdot u_1(x_1 + \Delta). \tag{17}$$

Let x'' be the allocation characterized by $x_1'' - x_1 = x_2' - x_2$. Note that such an allocation does indeed exist as $x_1 \leq x_2$ follows from Lemma 3 given that $x \in \mathcal{D}_i$. Note that $x_1'' - x_1 = x_2' - x_2$ implies that $\Delta = x_2' - m(x_2' - x_2) - x_2 = x_1' - m(x_1'' - x_1) - x_1$ and $q = 1 - p(x_2' - x_2) = 1 - p(x_1'' - x_1)$. Thus inequality (17) can be rewritten as

$$u_1(x_1) < (1 - p(x_1'' - x_1)) \cdot u_1(x_1'' - m(x_1' - x_1)).$$

Since $x_1'' - x_1 = x_2' - x_2 > 0$ and $x_1'' = x_1 + x_2' - x_2 \leq x_2' \leq 1$, this proves that $(p, m) \notin \mathcal{S}^1(y)$, which is what we wanted to show.

□

Proof of Statement (1). To prove that (1) holds, assume (1) is not true, i.e. assume there exists an allocation $x \in \mathcal{X}^{**}$ such that such that either $x_1 < \bar{x}_1$ or $x_2 < \bar{x}_2$.

We will obtain a contradiction for the case where $x_1 < \bar{x}_1$. The argument for the case where $x_2 < \bar{x}_2$ is analogous. Assume, therefore, $x_1 < \bar{x}_1$ and let $y \in \mathcal{X}$ be an allocation such that $x_1 < y_1 < \bar{x}_1$.

Let $\mathcal{C}$ be the cost set consisting of all pairs (p, m) such that $p : [0,1] \to [0,1]$ and $m : [0,1] \to [0,\infty)$ are continuous functions. Note that $x \in \mathcal{X}^{**}$ implies that it cannot be that $\mathcal{S}_{\mathcal{C}}(x) \subsetneq \mathcal{S}_{\mathcal{C}}(y)$.[31]

By Lemma 3, $x, y \in \mathcal{D}_1$. Therefore , by Lemma 4,

$$\mathcal{S}^1_{\mathcal{C}}(y) \subset \mathcal{S}^2_{\mathcal{C}}(y)$$

and

$$\mathcal{S}^1_{\mathcal{C}}(x) \subset \mathcal{S}^2_{\mathcal{C}}(x).$$

Note that by Lemma 2, $\mathcal{S}^1_{\mathcal{C}}(x) \subset \mathcal{S}^1_{\mathcal{C}}(y)$ and $\mathcal{S}^2_{\mathcal{C}}(y) \subset \mathcal{S}^2_{\mathcal{C}}(x)$. Thus,

$$\mathcal{S}_{\mathcal{C}}(x) = \mathcal{S}^1_{\mathcal{C}}(x) \cap \mathcal{S}^2_{\mathcal{C}}(x) = \mathcal{S}^1(x) \subset \mathcal{S}^1(y) = \mathcal{S}^1_{\mathcal{C}}(y) \cap \mathcal{S}^2_{\mathcal{C}}(y) = \mathcal{S}_{\mathcal{C}}(y).$$

We have shown that $\mathcal{S}_{\mathcal{C}}(x) \subset \mathcal{S}_{\mathcal{C}}(y)$.

Let $(p, m) \in \mathcal{C}$ be the pair given by $p : [0,1] \to [0,1]$ satisfying $p \equiv 0$ and $m : [0,1] \to [0,\infty)$ satisfying $m \equiv y_2$. Since $\bar{x}_1 \leq \frac{1}{2}$ by Lemma 3, $x_1 < y_1 < \bar{x}_1$ implies that $\max(x_1, 1 - x_1) > \max(y_1, 1 - y_1) = y_2$ and thus $(p, m) \notin \mathcal{S}_{\mathcal{C}}(x)$ and $(p, m) \in \mathcal{S}_{\mathcal{C}}(y)$. Given that we have already shown that $\mathcal{S}_{\mathcal{C}}(x) \subset \mathcal{S}_{\mathcal{C}}(y)$, this means that $\mathcal{S}_{\mathcal{C}}(x) \subsetneq \mathcal{S}_{\mathcal{C}}(y)$. However, we have already noted that this cannot be if $x \in \mathcal{X}^{**}$. The contradiction proves that statement (1) holds. □

The next lemma characterizes the elements of the set $\{x \in \mathcal{X} : x_1 \geq \bar{x}_1 \text{ and } x_2 \geq \bar{x}_2\}$. It will be used to prove the that statement (2) holds.

[31] To prove that $\mathcal{S}_{\mathcal{C}}(x) \subsetneq \mathcal{S}_{\mathcal{C}}(y)$ cannot hold for $x \in \mathcal{X}^{**}$, it is enough to show that $\mathcal{S}_{\mathcal{C}}(x) \subsetneq \mathcal{S}_{\mathcal{C}}(y)$ implies $x \notin \mathcal{X}^{**}$.
Assume $\mathcal{S}_{\mathcal{C}}(x) \subsetneq \mathcal{S}_{\mathcal{C}}(y)$. Since any deviation cost set $\mathcal{C}'$ satisfies $\mathcal{C}' \subset \mathcal{C}$, it is the case that $\mathcal{S}_{\mathcal{C}'}(x) = \mathcal{S}_{C}(x) \cap \mathcal{C}'$ and $\mathcal{S}_{\mathcal{C}'}(y) = \mathcal{S}_{C}(y) \cap \mathcal{C}'$. Thus $\mathcal{S}_{\mathcal{C}}(x) \subsetneq \mathcal{S}_{\mathcal{C}}(y)$ implies $\mathcal{S}_{\mathcal{C}'}(x) \subset \mathcal{S}_{\mathcal{C}'}(y)$ for all deviation cost sets $\mathcal{C}'$. Thus, y is easier or equally easy to sustain as x for any cost set $\mathcal{C}'$ and there is a cost set (namely $\mathcal{C}$) for which y is easier to sustain. Thus, $x \notin X^{**}$.

Lemma 5. *Assume the utility functions are normalized such that* $u_1(0) = u_2(0) = 0$.[32] *For any* $x^* \in \{x \in \mathcal{X} : x_1 \geq \bar{x}_1 \text{ and } x_2 \geq \bar{x}_2\}$ *at least one of the following statements is true:*

(a) x^* *satisfies* $x_1^* = x_2^* = \frac{1}{2}$.

(b) There exist $i, j \in \{1, 2\}$ *such that* $x_i^* < x_j^*$ *and*

$$\frac{u_i(x_i^*)}{u_i(x_i^* + \Delta)} \geq \frac{u_j(x_j^*)}{u_j(x_j^* + \Delta)} \tag{18}$$

for some $\Delta \in (0, 1 - x_j^*]$.

(c) x^* *is the symmetric Nash bargaining solution.*

Proof of Lemma 5. Let $x^* \in \{x \in \mathcal{X} : x_1 \geq \bar{x}_1 \text{ and } x_2 \geq \bar{x}_2\}$. Clearly, the statement of the lemma is true if $x^* = (\frac{1}{2}, \frac{1}{2})$. We will prove the lemma for the case where $x_1^* < x_2^*$. The argument for the case where $x_2^* < x_1^*$ is analogous.

Since $x^* \in \{x \in \mathcal{X} : x_1 \geq \bar{x}_1 \text{ and } x_2 \geq \bar{x}_2\}$ implies that $x_1^* \geq \bar{x}_1$, it must be that either $x_1^* > \bar{x}_1$ or $x_1^* = \bar{x}_1$.

Consider first the case where $x_1^* > \bar{x}_1$. Note that in this case, by Lemma 3, $x^* \notin \mathcal{D}_1$. However, $x^* \notin \mathcal{D}_1$ implies that there must exist $q \in (0, 1)$ and $\Delta \in (0, 1)$ such that the statement

$$x_2^* + \Delta \leq 1 \text{ and } u_2(x_2^*) \leq q \cdot u_2(x_2^* + \Delta) \tag{19}$$

holds but the statement

$$x_1^* + \Delta \leq 1 \text{ and } u_1(x_1^*) < q \cdot u_1(x_1^* + \Delta) \tag{20}$$

does not hold. Note now that, since we assumed $x_1^* < x_2^*$, the inequality $x_2^* + \Delta \leq 1$ from (19) implies $x_1^* + \Delta \leq 1$ from (20). Thus, if (20) does not hold it must be that

$$u_1(x_1^*) \geq q \cdot u_1(x_1^* + \Delta)$$

[32] If this was not the case, the formula in statement (b) would need to be adjusted.

Combining the last inequality with the second inequality from (19) we obtain that

$$\frac{u_1(x_1^*)}{u_1(x_1^*+\Delta)} \geq \frac{u_2(x_2^*)}{u_2(x_2^*+\Delta)}.$$

Thus, in this case, x^* satisfies condition (b) in the statement of the lemma.

All that remains is to prove the lemma for the case where $x_1^* = \bar{x}_1$. To this end, let x^n be a sequence of allocations such that $\frac{1}{2} > x_1^n > \bar{x}_1$ and $x^n \to \bar{x}_1$. Note that since $x_1^n > \bar{x}_1$, the same reasoning that yielded (18) for x^* with $\frac{1}{2} > x_1^* > \bar{x}_1$ will yield that for each n there exists $\Delta^n \in (0, x_1^n]$ such that

$$\frac{u_1(x_1^n)}{u_1(x_1^n+\Delta^n)} \geq \frac{u_2(x_2^n)}{u_2(x_2^n+\Delta^n)} \tag{21}$$

Since $\Delta^n \in [0,1]$ for all n and $[0,1]$ is compact there exists a convergent sub-sequence Δ^{n_k}. Let $\Delta = \lim_{k\to\infty} \Delta^{n_k} \in [0, \bar{x}_1]$.

If $\Delta > 0$, inequalities (21) together with the fact that u_1 and u_2 are continuous implies that, in the limit,

$$\frac{u_1(x_1^*)}{u_1(x_1^*+\Delta)} \geq \frac{u_2(x_2^*)}{u_2(x_2^*+\Delta)}$$

Thus, in this case, x^* satisfies condition (b) in the statement of the lemma.

If $\Delta = 0$, then (21) implies that

$$\frac{u_1(x_1^*)}{u_1'(x_1^*)} \geq \frac{u_2(x_2^*)}{u_2'(x_2^*)}.$$

Note that if the last inequality is binding, then x^* satisfies condition (c) in the statement of the lemma.[33] On the other hand if the last inequality is strict then for all sufficiently small positive h, it will be the case that

$$\frac{u_1(x_1^*)}{u_1(x_1^*+h)} \geq \frac{u_2(x_2^*)}{u_2(x_2^*+h)}.$$

Thus, in this case x^* satisfies condition (b) in the statement of the lemma.

[33]The unique allocation satisfying

$$\frac{u_1(x_1^*)}{u_1'(x_1^*)} = \frac{u_2(1-x_1^*)}{u_2'(1-x_1^*)}.$$

is the Nash bargaining solution as the above equation is the first order condition for the problem $\max_{x_1\in[0,1]} u_1(x_1)\cdot u_2(1-x_1)$.

□

Proof of Statement (2). We will now show that for any $x^* \in \{x \in \mathcal{X} : x_1 \geq \bar{x}_1 \text{ and } x_2 \geq \bar{x}_2\}$, there exists a non-monetary deviation cost set $\mathcal{C}$ such that x^* is easiest to sustain as a norm. Recall that, by Remark 2, we can without loss of generality restrict attention to the case where $u_i(0) = 0$ for $i \in \{1, 2\}$.

Let $x^* \in \{x \in \mathcal{X} : x_1 \geq \bar{x}_1 \text{ and } x_2 \geq \bar{x}_2\}$. Note that x^* must then satisfy (a), (b), or (c) in Lemma 5. We will consider the three cases separately.

Step 1: Consider first the case where $x^* = (\frac{1}{2}, \frac{1}{2})$, i.e. condition (a) in Lemma 5 is satisfied. For $\varepsilon \in (0, \frac{1}{2})$, define $p^\varepsilon : [0, 1] \to [0, 1]$ by

$$p^\varepsilon(h) = \begin{cases} 1 & \text{for } h \in [0, \frac{1}{2}] \\ 1 - \frac{h - \frac{1}{2}}{\varepsilon} & \text{for } h \in [\frac{1}{2}, \frac{1}{2} + \varepsilon] \\ 0 & \text{for } h \in [\frac{1}{2} + \varepsilon, \infty). \end{cases}$$

Furthermore, let $m_o : [0, 1] \to (0, \infty)$ be defined by

$$m_o(h) = 0$$

for all $h \in [0, \infty)$ and let $\mathcal{C}$ be the deviation cost set given by

$$\mathcal{C} = \{(p^\varepsilon, m_o) : \varepsilon \in (0, \frac{1}{2})\}.$$

The fact that, for all $\varepsilon \in (0, \frac{1}{2})$, $p(h) = 1$ for $h \in [0, \frac{1}{2}]$, implies that $\mathcal{S}_\mathcal{C}((\frac{1}{2}, \frac{1}{2})) = \mathcal{C}$. On the other hand, for any, $x' \in \mathcal{X} - (\frac{1}{2}, \frac{1}{2})$, it will be the case that $(p^\varepsilon, m_o) \notin \mathcal{S}_\mathcal{C}(x')$ for $\varepsilon < \frac{|x'_1 - x'_2|}{2}$. Thus, for any such x', $\mathcal{S}_\mathcal{C}(x') \subsetneq \mathcal{S}(x)$. We have shown that, for the deviation cost set $\mathcal{C}$, $x^* = (\frac{1}{2}, \frac{1}{2})$ is easiest to sustain.

Step 2: Next, consider the case where condition (b) in Lemma 5 is satisfied for $i = 1$ and $j = 2$. The argument in the case where condition (b) in Lemma 5 is satisfied for $i = 2$ and $j = 1$ is analogous.

Let $m_o : [0, 1] \to (0, \infty)$ again be defined by

$$m_o(h) = 0$$

for all $h \in [0, \infty)$, define $p_A : [0,1] \to [0,1]$ by

$$p_A(h) = 1 - \min(\frac{u_1(x_1^*)}{u_1(x_1^* + h)}, \frac{u_2(x_2^*)}{u_2(x_2^* + h)})$$

and define $p_B^\varepsilon : [0,1] \to [0,1]$ for $\varepsilon \in (0, \frac{1}{2})$ by

$$p_B^\varepsilon(h) = \begin{cases} 1 & \text{for } h \in [0, 1 - x_1^*] \\ 1 - \frac{h-1-x_1^*}{\varepsilon} & \text{for } h \in [1 - x_1^*, 1 - x_1^* + \varepsilon] \\ 0 & \text{for } h \in [1 - x_1^* + \varepsilon, \infty). \end{cases}$$

Now, consider the deviation cost set $\mathcal{C}$ defined by

$$\mathcal{C} = \{(p_A, m_o)\} \cup \{(p_B^\varepsilon, m_o) : \varepsilon \in (0, \frac{1}{2})\}.$$

Note that $\mathcal{S}_\mathcal{C}(x^*) = \mathcal{C}$ as

$$(1 - (p_A(x_i' - x_i^*)) \cdot u_i(x_i') =$$

$$= \min(\frac{u_1(x_1^*)}{u_1(x_1^* + x_i' - x_i^*)}, \frac{u_2(x_2^*)}{u_2(x_2^* + x_i' - x_i^*)}) \cdot u_i(x_i') \leq$$

$$\leq \frac{u_i(x_i^*)}{u_i(x_i')} \cdot u_i(x_i') = u_i(x_i^*),$$

implies that $(p_A, m) \in \mathcal{S}_\mathcal{C}(x^*)$ and $p_B^\varepsilon(h) = 1$ for $h \in [0, 1 - x_1^*]$ and $\varepsilon \in (0, \frac{1}{2})$ implies that $(p_B^\varepsilon, m) \in \mathcal{S}_\mathcal{C}(x^*)$.

We will now show that $\mathcal{S}_\mathcal{C}(x^*) \subsetneq \mathcal{C}$ for $x \neq x^*$. For any x with $x_1 < x_1^*$ this is the case as $(p_B^\varepsilon, m) \notin \mathcal{S}_\mathcal{C}(x)$ for $\varepsilon < x_1^* - x_1$. We will show that, on the other hand, for any x with $x_1 > x_1^*$ (and therefore $x_2 < x_2^*$), it is the case that $(p_A, m) \notin \mathcal{S}_\mathcal{C}(x)$.

To see that this is the case, recall that in this step we have assumed that x^* satisfies (b) in Lemma 5. Let $\Delta \in (0, 1 - x_2^*]$ be such that equation (18) from Statement (b) in Lemma 5 is satisfied. Note that for $x' = (x_1 - \Delta, x_2 + \Delta)$ we have

$$(1 - p_A(x_2' - x_2)) \cdot u_2(x_2') =$$

$$= \min(\frac{u_1(x_1^*)}{u_1(x_1^* + \Delta)}, \frac{u_2(x_2^*)}{u_2(x_2^* + \Delta)}) \cdot u_2(x_2') =$$

$$\frac{u_2(x_2^*)}{u_2(x_2^*+\Delta)} \cdot u_2(x_2') = \frac{u_2(x_2^*)}{u_2(x_2^*+\Delta)} \cdot u_2(x_2+\Delta) > u_2(x_2).$$

where the second equality follows from inequality (18) and the last inequality holds as $\frac{u_2(z_2)}{u_2(z_2+\Delta)}$ is increasing in z_2.[34] Since $\mathcal{S}_{\mathcal{C}}(x^*) = \mathcal{C}$ and $\mathcal{S}_{\mathcal{C}}(x) \subsetneq \mathcal{C}$ for $x \neq x^*$, we have shown that x^* is easiest to sustain for the cost set $\mathcal{C}$.

Step 3: Next, consider the case where x^* satisfies condition (c) in Lemma 5, i.e. x^* is the Nash bargaining solution. For this case, we have already seen in Subsection 2.3 that there exists a non-monetary deviation cost set $\mathcal{C}$ such that x^* is easiest to sustain.

4.3. **Proof of Proposition 4.** Let us start by noticing that the values $\bar{y}_1$ and $\bar{y}_2$ are indeed well defined.

Lemma 6. *Let $i, j \in \{1,2\}$ such that $i \neq j$. Then the equation*

$$\frac{\frac{u_i(2\bar{y}_i)-u_i(\bar{y}_i)}{\bar{y}_i}}{u_i(\bar{y}_i)-u_i(0)} = \frac{u_j'(1-\bar{y}_i)}{u_j(1-\bar{y}_i)-u_j(0)},$$

has a unique solution $\bar{y}_i$. Moreover, if x^{NBS} is the Nash bargaining solution, $\bar{y}_i \leq x_i^{NBS}$ holds.

Proof of Lemma 6. Consider the equation in the statement of the lemma. Note that, for $\bar{y}_i \to 0$, the left hand side goes to infinity[35] and the left hand converges to $u_j'(1)/(u_j(1)-u_j(0))$. Similarly, for $\bar{y}_i \to 1$, the right hand side converges to infinity and the left hand side converges to some real number. Since the left hand side and the right hand side are continuous in $\bar{y}_i$ for $\bar{y}_i \in (0,1)$, this implies that the equation in the lemma has a solution.

The uniqueness of the solution follows from the fact that the right hand side of the equation in the lemma is increasing in $\bar{y}_i$ and the left hand side is decreasing in $\bar{y}_i$. Indeed, the fact that the right hand side is increasing follows immediately from the

[34] To prove that $\frac{u_2(z_2)}{u_2(z_2+\Delta)}$ is increasing in z_2 it is enough to show that $\frac{u_2(z_2+\Delta)}{u_2(z_2)}$ is decreasing in z_2 but $\frac{u_2(z_2+\Delta)}{u_2(z_2)} = 1 + \frac{u_2(z_2+\Delta)-u_2(z_2)}{u_2(z_2)}$, $u_2(z_2+\Delta)-u_2(z_2)$ is decreasing in z_2 (as u_2 is a concave function) and $u_2(z_2)$ is increasing in z_2 (as u_2 is an increasing function).

[35] This follows from the observation that $\frac{u_i(2\bar{y}_i)-u_i(\bar{y}_i)}{\bar{y}_i}$ converges to $u_i'(0)$ as $\bar{y}_i \to 0$.

fact that u_j is increasing and concave. The fact that the left hand side is decreasing in $\bar{y}_i$ follows from the observation that $\frac{u_i(2\bar{y}_i)-u_i(\bar{y}_i)}{u_i(\bar{y}_i)-u_i(0)}$ is decreasing, which is the case as

$$\frac{d}{dx_2}\frac{u_i(2\bar{y}_i)-u_i(\bar{y}_i)}{u_i(\bar{y}_i)-u_i(0)} = \frac{d}{dx_2}\frac{u_i(2\bar{y}_i)-u_i(0)}{u_i(\bar{y}_i)-u_i(0)} =$$

$$= \frac{u_i'(2\cdot\bar{y}_i)\cdot(u_i(\bar{y}_i)-u_i(0)) - (u_i(2\cdot\bar{y}_i)-u_i(0))\cdot u_i'(\bar{y}_i))}{(u_i(\bar{y}_i)-u_i(0))^2} < 0$$

where the last inequality follows from $u_i(\bar{y}_i)-u_i(0) < u_i(2\cdot\bar{y}_i)-u_i(0)$ and $u_i'(2\cdot\bar{y}_i) \leq u_i'(\bar{y}_i)$.

The fact that $\bar{y}_i \leq x_i^{NBS}$ follows from the above monotonicity results together with the observation that $\frac{u_i(2\bar{y}_i)-u_i(\bar{y}_i)}{\bar{y}_i} < u_i'(\bar{y}_i)$ and the fact that x_i^{NBS} solves

$$\frac{u_i'(x_i^{NBS})}{u_i(x_i^{NBS})-u_i(0)} = \frac{u_j'(1-x_i^{NBS})}{u_j(1-x_i^{NBS})-u_j(0)}.$$

□

Proof of Proposition 4. The fact that $\mathcal{X}^* = \mathcal{X}^{**} = \{x \in \mathcal{X} : x_1 \geq \bar{x}_1 \text{ and } x_2 \geq \bar{x}_2\}$ is a subset of Y^{**} now follows from Lemma 5 and Lemma 6. To see that this is indeed the case, assume without loss of generality that u_1 and u_2 have been normalized so that $u_1(0) = u_2(0) = 0$ and $u_1(1) = u_2 = 1$.[36] Let $x^* \in \mathcal{X}^{**}$. Since $x^* \in \mathcal{X}^{**}$, Lemma 5 tells us that x^* must satisfy at least one of the conditions (a), (b), or (c) in Lemma 5. Note that if x^* satisfies conditions (a) or (c), Lemma 6 immediately implies $x^* \in \mathcal{Y}^{**}$. Consider therefore the case where x^* satisfies condition (b), i.e. there exists $i, j \in \{1,2\}$ such that $x_i^* < x_j^*$ and (18) holds for some $\Delta \in (0, 1-x_j^*] = (0, x_i^*]$. Rearranging terms and subtracting 1 from both sides of the inequality note that (18) is equivalent to

$$\frac{u_i(x_i^*+\Delta)-u_i(x_i^*)}{u_i(x_i^*)} \leq \frac{u_j(x_j^*+\Delta)-u_j(x_j^*)}{u_j(x_j^*)}.$$

The fact that u_i is concave and $\Delta \leq x_i^*$ implies that

$$\frac{u_i(2\cdot x_i^*)-u_i(x_i^*)}{x_i^*} \leq \frac{u_i(x_i^*+\Delta)-u_i(x_i^*)}{\Delta}.$$

[36]This is without loss of generality as neither the set $\mathcal{X}^{**}$ nor the definition of $\bar{y}_1$ and $\bar{y}_2$ are affected by positive affine transformations of the utility functions.

The fact that u_j is concave implies that

$$\frac{u_j(x_j^* + \Delta) - u_j(x_j^*)}{\Delta} \leq u_j'(x_j).$$

Combining the last three inequalities and using $x_j^* = 1 - x_i^*$ yields

$$\frac{\frac{u_i(2 \cdot x_i^*) - u_i(x_i^*)}{x_i^*}}{u_i(x_i^*)} \leq \frac{u_j'(1 - x_i^*)}{u_j(1 - x_i^*)}.$$

In the proof of Lemma 6 we already argued that the left hand side of the above inequality is decreasing in x_i^* and the right hand side is increasing in x_i^*. This implies $x_i^* \geq \bar{y}_i$. But $x_i^* < x_j^*$ implies $x_j^* \geq \frac{1}{2}$. Thus, $x_i^* \geq \min(\frac{1}{2}, \bar{y}_i)$ and $x_j^* \geq \min(\frac{1}{2}, \bar{y}_j)$ which means that $x^* \in \mathcal{Y}^{**}$.

□

4.4. **Proof of Proposition 5.** An analogous reasoning as in Subsection 2.3 yields that

$$\mathcal{S}(x) = \{p \in P : p'(0) \cdot (x_S - v(S)) \geq 1 \text{ for all nonempty}$$

$$S \subset N \text{ such that } x_S < v(N) - v(N - S)\}.$$

Since $x_S < v(N) - v(N - S)$ is equivalent to $x_S - v(S) < v(N) - v(N - S) - v(S)$ we obtain

$$\mathcal{S}(x) = \{p \in P : \forall_{S \subset N, S \neq \emptyset} \ (x_S - v(S)) \geq \min(\frac{1}{p'(0)}, v(N) - v(N - S) - v(S))\}.$$

Note that if x is an allocation that does not lie in the core of the game v there exists a coalition S such that $x_S - v(S) < 0$. Since $p'(0) > 0$ for all $(p, m) \in \mathcal{C}^{Nash}$ and $v(N) - v(N - S) - v(S) \geq 0$ for any $S \subset N$ by (9), this means that, for $x \notin Core(v)$ it is the case that $\mathcal{S}_{\mathcal{C}^{Nash}}(x) = \emptyset$. This implies that if $Core(v)$ is empty, $\mathcal{U}_{\mathcal{C}^{Nash}} = \mathcal{X}$. We have shown the statement of the proposition for the case where the core is empty.

Now, consider the case where the core of the game v is non-empty and consider an $x \in Core(v)$, i.e. an x such that $x_S - v(S) \geq 0$ for all coalitions S. Note that, for

such an x,

$$(x_S - v(S)) \geq \min(\frac{1}{p'(0)}, v(N) - v(N-S) - v(S))$$

is always satisfied if $v(N) - v(N-S) - v(S) = 0$. On the other hand, if $v(N) - v(N-S) - v(S) > 0$ then either $x_S - v(S) < v(N) - v(N-S) - v(S)$ or $x_{N-S} - v(N-S) < v(N) - v(N-S) - v(S)$.[37] This means that, if $v(N) - v(N-S) - v(S) > 0$, then $\min(x_S - v(S), x_{N-S} - v(N-S)) < v(N) - v(N-S) - v(S)$. But this implies that, for $x \in Core(v)$,

$$\mathcal{S}(x) = \{p \in P : \min_{S:S\subset N, v(N)-v(N-S)-v(S)>0} x_S - v(S) \geq \frac{1}{p'(0)}\}.$$

Thus, for the case where the core is non-empty $\mathcal{U}_{\mathcal{C}^{Nash}}$ is equal to exactly those allocations for which

$$\min_{S:S\subset N, v(N)-v(N-S)-v(S)>0} x_S - v_S$$

is largest.

4.5. **Proof of Theorem 3.** We start with a lemma that gives a more convenient characterization of the sets $\mathcal{S}_{\mathcal{C}}(x)$.

Lemma 7. *Let $\mathcal{C}$ be a deviation cost set and x an allocation in $\mathcal{X}$. The set $\mathcal{S}_{\mathcal{C}}(x)$ is exactly equal to the set of $(p, m) \in \mathcal{C}$ such that, for each non-empty $S \subset N$,*

$$p(\Delta) \cdot (x_S - v(S)) \geq (1 - p(\Delta)) \cdot (\Delta - m(\Delta)) \tag{22}$$

for all $\Delta \in (0, x_{N-S} - V(N-S)]$.

Proof of Lemma 7. This follows immediately from Definition 9. Indeed, we defined $\mathcal{S}_{\mathcal{C}}(x)$ to be the set of $(p, m) \in \mathcal{C}$ such that inequality (10) holds for all coalitions S and $x'_S \in [x_S, v(N) - v(N-S)]$. Subtracting $v(S)$ from both sides of inequality (10) and substituting x'_S with $x_S + \Delta$, we see that inequality (10) holds for all coalitions

[37] As $x_S + x_{N-S} = v(N)$, $x_S - v(S) \geq v(N) - v(N-S) - v(S)$ and $x_{N-S} - v(N-S) \geq v(N) - v(N-S) - v(S))$ together would imply $v(N) - v(N-S) - v(S) \geq 2 \cdot (v(N) - v(N-S) - v(S))$ which cannot hold if $v(N) - v(N-S) - v(S) > 0$.

S and $x'_S \in [x_S, v(N) - v(N-S)]$ if and only if

$$(x_S - v(S)) \geq (1 - p(\Delta)) \cdot (x_S + \Delta - v(S) - m(\Delta))$$

holds for all coalitions S and $\Delta \in [0, V(N) - V(N-S) - x_S] = [0, x_{N-S} - V(N-S)]$. Rearranging the terms in the last inequality, yields the inequality in the lemma. □

The next two lemmas concern the relationship between $\mathcal{S}_{\mathcal{C}}(x)$ and $\mathcal{S}_{\mathcal{C}}(x)$ if x is weakly less extreme than y.

Lemma 8. *If an allocation $x \in \mathcal{X}$ is weakly less extreme than an allocation $y \in \mathcal{X}$, then, for any deviation cost set $\mathcal{C}$, $\mathcal{S}_{\mathcal{C}}(y) \subset \mathcal{S}_{\mathcal{C}}(x)$.*

Proof of Lemma 8. This follows immediately from Lemma 7 as the conditions in Lemma 7 depend only on $x_S - v(S)$ and $x_{N-S} - v(N-S)$ and are easier to satisfy if the value $x_{N-S} - v(N-S)$ is smaller or the value $x_S - v(S)$ is larger. □

Lemma 9. *Let $x, y \in \mathcal{X}$. If x not weakly less extreme than y, then there exists a deviation cost set $\mathcal{C}$ such that $\mathcal{S}_{\mathcal{C}}(y) \not\subset \mathcal{S}_{\mathcal{C}}(x)$.*

Proof of Lemma 9. Let $x, y \in \mathcal{X}$ be such that x is not weakly less extreme than y. The fact that x is not weakly less extreme than y implies that there exists a coalition S with

$$x_S - v(S) < x_{N-S} - v(N-S) \tag{23}$$

such that for all coalitions $S' \subset N$

$$y_{S'} - v(S') > x_S - v(S) \text{ or } y_{N-S'} - v(N-S') < x_{N-S} - v(N-S).$$

Since the set of all coalitions $S' \subset N$ is finite, there exists an $\varepsilon > 0$ such that

$$y_{S'} - v(S') > x_S - v(S) + \varepsilon \text{ or } y_{N-S'} - v(N-S') < x_{N-S} - v(N-S) - \varepsilon \tag{24}$$

holds for all coalitions $S' \subset N$.

Note that (9) together with $x_S + x_{N-S} = v(N)$ and (23) implies[38]

$$x_{N-S} - v(N-S) > 0.$$

Let $\mathcal{C} = \{(p, m)\}$, where $p : [0,1] \to [0,1]$ is given by $p(h) = \frac{1}{2}$ and $m : [0,1] \to [0, \infty)$ is given by

$$m(h) = \begin{cases} v(N) + h & \text{if } h \leq x_{N-S} - v(N-S) - \varepsilon \\ -x_S + v(S) - \varepsilon + h & \text{if } h \geq x_{N-S} - v(N-S) \\ v(N) + h + \frac{(h - (x_{N-S} - v(N-S) - \varepsilon)) \cdot (-x_S + v(S) - \varepsilon - v(N))}{\varepsilon} & \text{otherwise} \end{cases}$$

Note that, by Lemma 7, $(p, m) \notin \mathcal{S}(x)$. Indeed, for coalition S and $\Delta = x_{N-S} - v(N-S)$ inequality (22) does not hold as

$$\frac{1}{2} \cdot (x_S - v(S)) < \frac{1}{2} \cdot (x_S - v(S) + \varepsilon).$$

On the other hand, by Lemma 7, $(p, m) \in \mathcal{S}(y)$. To see that this is so, consider again inequality (22). By (24), for any coalition S' it is the case that either $y_{S'} - v(S') > x_S - v(S) + \varepsilon$ or $y_{N-S'} - v(N-S') < x_{N-S} - v(N-S) - \varepsilon$. If $y_{N-S'} - v(N-S') < x_{N-S} - v(N-S) - \varepsilon$, then

$$p(\Delta) \cdot (y_{S'} - v(S')) \geq (1 - p(\Delta)) \cdot (\Delta - m(\Delta)) \tag{25}$$

holds for any $\Delta \in (0, y_{N-S} - v(N-S)]$ as, for those Δ, $p(\Delta) = \frac{1}{2}$ and $m(\Delta) = v(N) + \Delta$. If, $y_{S'} - v(S') > x_S - v(S) + \varepsilon$ then (25) follows from the fact that $p(\Delta) = \frac{1}{2}$ and $m(\Delta) \geq -x_S + v(S) - \varepsilon + \Delta$. We have constructed a deviation cost set $\mathcal{C}$ such that $\mathcal{S}_{\mathcal{C}}(y) \not\subset \mathcal{S}_{\mathcal{C}}(x)$.

□

[38]If $x_{N-S} - v(N-S) \leq 0$ then $x_S - v(N-S) < 0$ by (23). But this would mean $x_{N-S} - v(N-S) + x_S - v(N-S) < 0$ or, given $x_S + x_{N-S} = v(N)$, $v(N) - v(N-S) - v(N-S) < 0$ which would contradict (9).

We are now ready to prove Theorem 3. Let Z be the set of allocations x such that there exists no allocation y that is strictly less extreme. To prove Theorem 3 it is enough to show that

(1) $\mathcal{X}^{**}(v) \subset Z$ and
(2) $Z \subset \mathcal{X}^{**}(v)$.

To prove (1) we will assume $x \notin Z$ and show that $x \notin \mathcal{X}^{**}$. Assume, therefore, $x \notin Z$. This means there is an allocation y that is strictly less extreme than x, i.e. an allocation y such that y is weakly less extreme than x and x is not weakly less extreme than x. Since y is weakly less extreme than x then, by Lemma 8, for any cost set $\mathcal{C}$ it is the case that $\mathcal{S}_{\mathcal{C}}(x) \subset \mathcal{S}_{\mathcal{C}}(y)$. Since x is not equally extreme as y, by Lemma 9, it is not the case that $\mathcal{S}_{\mathcal{C}}(x) = \mathcal{S}_{\mathcal{C}}(y)$ for all cost sets $\mathcal{C}$. Thus, we can conclude that, for all cost sets $\mathcal{C}$ it is the case that $\mathcal{S}_{\mathcal{C}}(x) \subset \mathcal{S}_{\mathcal{C}}(y)$ and there must be at least one cost set $\mathcal{C}$ such that $\mathcal{S}_{\mathcal{C}}(x) \subsetneq \mathcal{S}_{\mathcal{C}}(y)$. We conclude $x \notin \mathcal{X}^{**}$ holds which proves (1).

To prove (2) we will assume $x \notin X^{**}$ and show that $x \notin Z$. Assume, therefore, $x \notin X^{**}$. This means there exists an allocation y such that, for all cost sets $\mathcal{C}$, $\mathcal{S}_{\mathcal{C}}(x) \subset \mathcal{S}_{\mathcal{C}}(y)$ and at least for one cost set $\mathcal{C}$ it is the case that $\mathcal{S}_{\mathcal{C}}(x) \subsetneq \mathcal{S}_{\mathcal{C}}(y)$. Lemma 8 together with the fact that $\mathcal{S}_{\mathcal{C}}(x) \subset \mathcal{S}_{\mathcal{C}}(y)$ holds for all deviation cost sets $\mathcal{C}$ implies that y is weakly less extreme than x. Lemma 9 together with the fact that at least for one cost set $\mathcal{C}$ it is the case that $\mathcal{S}_{\mathcal{C}}(x) \subsetneq \mathcal{S}_{\mathcal{C}}(y)$ implies that it cannot be that x is weakly less extreme than x. Thus, y is strictly less extreme than x. This implies $x \notin Z$ holds which proves (2).

References

[1] Abreu, D. and F. Gul (2000) "Bargaining and Reputation," *Econometrica*, 68, 85 – 117

[2] Arrow, K. (1971) "Political and Economic Evaluation of Social Effects and Externalities," *Frontiers of Quantitative Economics.*, Amsterdam: North-Holland, p. 3-25

[3] Crawford, V (1982) "A Theory of Disagreement in Bargaining" *Econometrica*, 50:3, 607 – 637

[4] Compte, O. and P. Jehiel (2010) "The coalitional Nash Bargaining Solution," *Econometrica*, 78, 1593 – 1623

[5] Crawford, O. and P. Jehiel (2010) "The coalitional Nash Bargaining Solution," *Econometrica*, 78, 1593 – 1623

[6] Elster, J. (1989) "Social Norms and Economic Theory," *Journal of Economic Perspectives* 3:4, 99-117

[7] Kalai, E. and M. Smorodinsky (1975) "Other Solutions to Nash's Bargaining Problem," *Econometrica*, 43, 513 – 518

[8] Myerson, R. (1991) "Game Theory: Analysis of Conflict," *Harvard University Press*

[9] Perry, M. and P. Reny (1994) "A Nooncooperative View of Coalition Formation and the Core," *Econometrica*, 62, 795 – 817

[10] Nash, J. (1950) "The Bargaining Problem," *Econometrica*, 18, 155 – 162

[11] Nash, J. (1953) "Two-Person Cooperative Games," *Econometrica*, 21, 128 – 140

[12] Rubinstein, A. (1982) "Perfect Equilibrium in a Bargaining Model," *Econometrica*, 50, 97 – 109

[13] Rubinstein, A., Z. Safra, and W. Thomson (1992) "On the Interpretation of the Nash Bargaining Solution and Its Extensions to Non-Expected Utility Preferences," *Econometrica*, 60, 1171 – 1186

[14] Samuelson, L. (2004): "Information-Based Relative Consumption Effects," *Econometrica* 72:1, 93-118.

[15] Samuelson, L. and Swinkels, J. (2006): "Information, evolution, and utility," *Theoretical Economics* 1, 119-142.

[16] Schmeidler, D. (1969), "The nucleolus of a characteristic function game", *SIAM Journal on Applied Mathematics*, 17, 1163 – 1170

[17] Young, P. (1993): "An Evolutionary Model of Bargaining," *Journal of Economic Theory* 59:1, 145-168.

[18] Young, P. (2008): "Social Norms," in *The New Palgrave Dictionary of Economics*, Second Edition, Steven N. Durlauf and Lawrence E. Blume, eds., London, Macmillan