Edinburgh School of Economics

Discussion Paper Series

Number 37

Why do lions get the lion's share? A Hobbesian theory of agreements

Joan Esteban (CSIC)
Jozsef Sakovics (University of Edinburgh and CSIC)

Date
November 1999

Published by
School of Economics
University of Edinburgh
30 -31 Buccleuch Place
Edinburgh EH8 9JT
+44 (0)131 650 8361
http://www.ed.ac.uk/schools-departments/economics

THE UNIVERSITY *of* EDINBURGH

Why do lions get the lion's share? A Hobbesian theory of agreements*

Joan Esteban§

Institut d'Anàlisi Econòmica, CSIC

József Sákovics

University of Edinburgh and Institut d'Anàlisi Econòmica, CSIC

November 25, 1999

Abstract

We propose a novel approach for N-person bargaining, based on the idea – borrowed from Hobbes– that the agreement reached in a negotiation should be determined by how the direct conflict resulting from disagreement would be resolved. The explicit modeling of the *conflict game* directly leads to the observation that the outcome of conflict is a function of the stakes. Thus, our basic building block is the *disagreement function*, which maps each set of feasible agreements into a disagreement point. Using this function and a weakening(!) of the Independence of Irrelevant Alternatives axiom, based on individual rationality, we reach a unique solution. The main feature of the solution is that it is reached via a sequence of partial agreements. We also give three alternative characterizations; two based on multi-stage, strategic bargaining games and one on the possibility of renegotiation.

Keywords: Bargaining, conflict, disagreement, Hobbes, social contract.

JEL Numbers: C78, D74.

* We are thankful to Salvador Barberà, Jordi Brandts, Yeon-Koo Che, Joe Harrington, Carmen Herrero, Marco Mariotti, Rich McLean, Clara Ponsatí, Debraj Ray and especially to Andreu Mas-Colell, as well as to seminar participants at Alicante, Barcelona Jocs, CORE, the Kenilworth ESRC Game Theory Meeting, NYU, Rutgers and St. Andrews for most helpful discussions.

§ Corresponding author: IAE (CSIC), Campus de la UAB, 08193 Bellaterra, Barcelona, SPAIN. E-mail: esteban@cc.uab.es The first draft was written while J. Esteban visited CREI at Universitat Pompeu Fabra. He also gratefully acknowledges financial support from Fundación Pedro Barrié de la Maza and research grant DGICYT PB96-0897.

«The rich get the law passed by means of force and arms or get it accepted by fear to their might, aren't things this way? » Plato, *Republic*.

«What I am arguing here is that in order to explain the substantive content of social institutions and, therefore, completely explain institutional development and change, our theories must focus primarily on the strategic conflict itself and on the mechanisms by which this conflict is resolved», Knight (1992), p. 123.

1. Introduction

In this paper, we develop a novel approach to the theory of negotiation, inspired by Hobbes' theory of the social contract. We view bargaining as a *process*,[1] driven by the fear of the conflictual resolution that would result in case of disagreement. To capture this phenomenon, we complement the standard definition of a bargaining problem with the description of a *conflict game*, the one that would ensue upon disagreement. It is through this auxiliary game that the distribution of power of the participants manifests itself throughout the entire negotiation process. Indeed, our Hobbesian agreement provides a rationale as to why lions get the "lion's share" in the distribution of a pie/carcass.

Our Hobbesian agreement comes about as a sequence of partial agreements, where these are reached by granting each player her current outside/conflict payoff. The driving force behind this progression is based on two factors. First, the players always prefer to honor their partial agreements along the sequence and, thereby, to limit conflict. Second, we assume that conflict does not exhaust the entire surplus and, therefore, the payoffs to conflict always dominate the current *status quo*. This makes the menace of conflict always credible and pushes the players towards further concessions, until they finally reach an efficient allocation. Observe, that the efficiency of the agreement is derived, not assumed.

To fix ideas, consider the simple case of splitting an inheritance, say, ten dollars, between two siblings (who do not fancy each other). The siblings can either agree on a particular split at no cost, or disagree and engage in a costly dispute over the money. Suppose that, if players engaged in conflict, in equilibrium seven dollars would be wasted (on, say,

[1] This process may be an actual one or just a thought process, which directly leads the players to agreement. See also Subsections 4.1 and 4.2.

lawyers' fees), while of the remaining three dollars one player would expect to obtain two and the other one. This allocation may reflect the fact that, for instance, one's lawyer is "twice" as influential as the other's. As a result of the expected outcome of conflict, any agreement must give to the siblings at least two and one dollars, respectively. Recognizing this, they are willing to get to a partial agreement, which guarantees them these outside payoffs. Consequently, the effective area of dissent shrinks to the remaining seven dollars, which are precisely the benefits from cooperation. On the division of these seven dollars the siblings may again either agree or disagree and engage in a dispute. In the dispute, say, four dollars would be wasted and the strong sibling would obtain two and the weak one. Notice that even if they disagree, both siblings are better off by respecting their partial agreement and restricting the dispute to the distribution of the seven-dollar surplus. It thus follows that any agreement must give to the siblings at least four and two dollars, respectively. This observation generates a new partial agreement. Applying the argument repeatedly, we reach the final agreement, where the ten dollars are distributed accordingly with the power of the parties in the conflict game:[2] 2/3 and 1/3.

The idea that the opposing parties willingly choose to limit the extent of conflict in case of disagreement is crucial to our theory. The best known historical example for this is described in the Old Testament. When the Philistines and the Jews gather to fight on the battlefield, instead of a full-scale battle they "agree" to reduce conflict into a fight between one representative agent from each side (Goliath and David). Perhaps, another, hypothetical example can be more illuminating: Consider two countries of comparable military strength disputing a territory that has two large and one small oil fields. Assume that their cost of fighting in an all-out war is about the value of one large oil field each. Then they should –and would– agree on keeping one large oil field each in order to avoid conflict. However, they would not be able to agree on the ownership of the small well (assumed indivisible). *Should* they go to full-fledged conflict over it? Obviously, not. It is Pareto improving if they limit themselves to some minor border dispute. What keeps the conflict from escalation is the separation between the agreement and conflict *games:* not respecting a (partial) agreement is not a unilateral deviation *in* the conflict game; instead it is a unilateral deviation provoking a

[2] In contrast, both the Nash (1950) and the Kalai-Smorodinsky (1975) solutions would predict that the seven-dollar surplus over and above the (total) disagreement point would be brotherly shared by the two players. They would obtain 5.5 and 4.5 dollars in total, respectively.

transition *to* the conflict game. This way such a deviation is observable: the countries foresee each other's reaction to a unilateral deviation.[3]

We start our formal analysis by positing the existence of a *disagreement function*, which maps the set of available agreements into a disagreement outcome. We wish to emphasize that the disagreement function is not freely chosen by the modeler. Rather, it forms part of the description of the bargaining situation being modeled. As we mentioned above, this function derives from the equilibrium outcome of a non-cooperative model of the conflictual resolution of opposing interests. In general, we expect this equilibrium to vary with the stakes of the negotiation, hence the introduction of a function, as opposed to a point. Note that the conflictual resolution may take various forms: going to court, to strike, to call an arbitrator, to lobby, to cut prices, to waste time, to fight etc.

Equipped with the disagreement function, we turn to the axiomatic characterization of the Hobbes solution and show that there exists a unique agreement satisfying a weakening of Nash's axiom of Independence of Irrelevant Alternatives: Independence of Individually Irrational Alternatives. This new axiom simply states that the agreement should not depend on the availability of alternatives which give to at least one player strictly less than what she would get in disagreement. The key observation driving our result is that once we eliminate the individually non-rational agreements, the bargaining problem becomes a different one. This is so, not only because the set of available agreements has changed, but also because (consequently) via the disagreement function the threat point has changed too. Since our axiom applies to all bargaining problems it also applies to this new one, and further reduces the set of feasible agreements. What we show is that the repeated application of the axiom to the resulting sequence of bargaining games converges to a situation where the disagreement outcome is efficient, thus pinpointing a unique solution.

In the spirit of the "Nash program",[4] we also display justifications of our agreement concept in a purely strategic context. First, we present two types of multi-stage, non-cooperative bargaining games that, under mild behavioral assumptions, yield the same

[3] For example, according to our solution, in a complete information Cournot model two identical firms would each agree to produce half the monopoly quantity, which is indeed the optimal colluding outcome (for them). The Nash equilibrium would correspond to unrestricted conflict (that is, competition) in this case. See Subsection 4.2 for a more detailed discussion of the Cournot example.

[4] In his own words (Nash, 1953, p. 129): "The two approaches to the problem, via the negotiation model or via the axioms, are complementary; each helps to justify and clarify the other."

agreement point as our axiomatic solution, as the unique allocation implementable by a subgame-perfect Nash equilibrium. The main difference between the two games is whether or not it is possible to retract past offers in case of a deviation. Without the possibility of retracting, the Hobbesian agreement cannot be implemented in a single stage. Rather, the players need to go through the sequence of partial agreements that lead us to the axiomatic solution. On the other hand, if the players can choose not to ratify their latest proposals, we show that the Hobbesian agreement can be reached immediately.

In addition, we also present a characterization exclusively based on the possibility of renegotiation of any disagreement outcome. This approach assumes that the players are collectively rational, in the sense that they would not implement an inefficient outcome. Thus, upon conflict they would renegotiate the disagreement outcome using the disagreement function. We show that the only renegotiation-proof solution is the Hobbesian one.

The paper is structured as follows. In Section 2 we put our approach in perspective, by discussing its relation to Political Philosophy, Political Economy, and the standard bargaining approach. Section 3 provides the axiomatic characterization of the Hobbesian agreement. In Section 4 we present the three strategic characterizations of the Hobbes solution just mentioned above. In Section 5 we elucidate our theory by contrasting it to the related bargaining literature. Finally, Section 6 provides some concluding remarks.

2. Agreeing in the shadow of conflict

The view we develop here is to a good extent inspired in Hobbes' theory of social agreements. Well before Economics developed the theory of bargaining, Political Philosophy had addressed the question of social agreements in its inquiry on the foundations of the state. Thomas Hobbes (1588-1679) was possibly the first modern political philosopher who formulated an articulated theory of the *social contract*. In his analysis of the foundations of the social contract,[5] he viewed the possibility of a collective agreement as a case of "conditional cooperation" (in Taylor's, 1987, words), constrained by what individuals can obtain in the *state of nature*. The state of nature is the outcome that would ensue from a non-cooperative, rule-free interaction among utility maximizing, selfish individuals (Hobbes' *first axiom*). The outcome of this interaction is resource consuming and is governed by the

[5] See Taylor (1987) and Gauthier (1990).

differences in endowments across individuals. His *second axiom* asserts that there exist agreements that Pareto dominate the allocation achieved under the state of nature. Finally, according to his *third axiom*, agreements should be conditioned by the allocation resulting in the state of nature. Therefore, social agreements, being certainly preferable to anarchy, were not the outcome of an idealistic introspection on how things ought to be, but rather the viable outcome of a process conditioned by the might of the parties.

This view was largely shared by Jean-Jacques Rousseau (1782), for whom the formation of the political society and the establishment of laws «gave new constraints to the weak and new forces to the rich, irreversibly destroyed natural freedom, established forever property law and inequality» (p.170-1). Adam Smith (1776) also conceived the state as the creature of the mighty, specifically designed to give stability to the unequal distribution of wealth. In his own words: «The rich, in particular, are necessarily interested to support that order of things, which can alone secure them in the possession of their own advantages. (…) Civil government, so far as it is instituted for the security of property, is, in reality, instituted for the defence of the rich against the poor, or those who have some property against those who have none at all». (Book V, Chap. 1, Part II)

Despite the dramatically different normative positions as to what a "social contract" ought to be, all of them coincide in the positive analysis: in actual social agreements the mighty obtain a preferential treatment.[6] That actual social agreements will, at least partly, reflect the initial distribution of power is to be expected as long as a social contract is to be found acceptable by all parties. Therefore, and this is one of Hobbes' characteristic themes, we cannot develop a theory of social agreements without reference to the power of the parties in the non-cooperative scenario. The state of nature not only determines the size of the potential surplus to be shared, but also the shares themselves.

The logic of modern bargaining theory is rather different from the Hobbesian view. Positive and normative arguments are deeply interwoven. Bargaining theory, as we know it today, is entirely influenced –and, in our opinion, even constrained– by the path-breaking papers of John Nash (1950, 1953). In the first of these, he defined what a bargaining problem was and since then his set-up has not been questioned. Thus, a bargaining problem is a pair,

[6] One of the lines along which the position held by Rousseau (and by Smith) departs from Hobbes' views is on whether the inequality that forces a biased social contract is innate to humans or it is acquired.

(S,d), where *S* denotes the utility vectors attainable by agreement and *d* is the utility allocation in case no agreement is reached. Equipped with this stylized information, bargaining theory tries to identify an element of *S* as the solution to the bargaining problem. There are two complementary approaches to the characterization of a solution to the bargaining problem. On the one hand, the axiomatic, normative theory imposes certain desiderata on the solution and analyzes when are these compatible and when do they yield a unique agreement.[7] On the other hand, the strategic approach focuses on the positive analysis of which agreement(s) can be reached as a result of the bargainers' strategic interaction in some well-defined non-cooperative game.[8] Since Nash, the goal of bargaining theory has been to reconcile the positive and normative approaches and to prove that the solutions obtained from the two angles coincide. Thus, accordingly with standard bargaining theory, "respectable" non-cooperative solutions yield *a fortiori* fair agreements.

In contrast, we develop a positive theory of agreements, adopting Hobbes' position that takes the initial conditions as given and focuses on *reachable* social agreements, quite independently of the moral judgment they might deserve. We reserve normative considerations for the "state of nature," the initial conditions under which a particular agreement has been reached.[9] This view is consistent with Roemer's (1996) reservations about the moral content of a bargaining agreement obtained without a prior redistribution of the initial endowments.

It is our opinion that standard bargaining theory has been driven to the use of normative axioms because the description of the bargaining problem was so stylized that there were no bases left for a positive derivation of the corresponding agreement. Note that, implicit in Nash´s formulation there is the assumption that all the information concerning the background non-cooperative game, which is not embodied in *S* and *d*, is irrelevant to the characterization of an agreement. Under this assumption, any pair of bargaining problems with the same *S* and the same *d* are identical and, therefore, have to yield the same agreement,

---

[7] See Thomson (1994) for an overview of this literature.

[8] See Osborne and Rubinstein (1990) for a survey.

[9] Consider the parallel case of assigning the gains from exchange. Economics takes a *positive* stand and investigates the terms of trade that will actually take place, resulting from different market structures and characteristics of the traders. It does not inquire about which would have been the "fair" terms of trade. The normative valuations are reserved for the comparison of the distribution of the characteristics that condition the trade (distribution of endowments, for instance).

irrespective of the particular characteristics of the background situation.[10] For example, the disagreement outcome, d, is exogenously given and, therefore, it does not depend on what is at stake in the negotiation. More specifically, d is assumed independent of the possible partial agreements that can be reached on the way to a complete agreement. In the world created by Nash, the players are either in agreement or in disagreement, no middle ground is considered. From this assumption it logically follows that the implicit underlying process which determines d is supposed to have no bearing on the agreement the players will reach over the remaining surplus. The fact that one player might be a lion and the other a hyena is considered irrelevant when establishing the lion's share of the surplus (in excess of d), irrespective of whether the approach is normative or positive.

We explore whether a solution can be characterized saving on axioms and making a more intensive use of the information contained in the description of the background game (and not merely S and d). This approach is in line with the growing literature on the explicit modeling of the conflictual resolution of opposing interests. The works by Becker (1983) on pressure groups and Tullock (1980) on rent-seeking, are the predecessors of the more recent papers by Grossman (1991, 1994), Hirshleifer (1991, 1995), Horowitz (1993), and Skaperdas (1992) among many others.[11] The common feature of all these models is that the opposition of interests is resolved via conflict. Players expend resources into trying to make their preferred option prevail. The equilibrium outcome entails waste of resources and the particular allocation reached critically depends on what is at stake as well as on the relative power, among other relevant characteristics, of the players.

In view of this literature, it seems natural to inquire why there is conflict to start with, could there be a plausible conflict-avoiding agreement in this scenario? An agreement would save resources and, therefore, the crucial issue is how to share this surplus. However, potential agreements are not a central issue for most of these papers. On the other hand, the few who deal with it, obtain agreements that are influenced by the power of the parties. This is the case, for instance, of the papers by Grossman (1994) and Horowitz (1993) on land reform. In Grossman (1994), landowners voluntarily give away land in order to decrease the probability

[10] Roemer (1988) and, more recently, Chen and Maskin (1999) have also expressed their reservations about the standard description of a bargaining problem, pointing out that Nash's abstraction might be dispensing with essential information.

of an expropriatory revolution and to save on protective expenditures. The size of the redistribution depends on the effectiveness of each party in rebelling or preventing it, as well as on the initial degree of inequality. Horowitz's (1993) approach is different. Landlords and peasants start from a status quo distribution and reach a sequence of interim agreements. At each stage, if they fail to reach a new interim agreement, either party can expropriate the other with some given probability (reflecting their relative power) or the status quo stays (again with some exogenously given probability). The economy follows a sequence of interim agreements converging to a steady state distribution that exactly reflects the power of the parties. Our theory of agreements is in accordance with the behavior predicated in this class of conflict models.

3. A disagreement theory of bargaining

In this section, we present our formal analysis. We start by defining our generalized version of the bargaining problem, incorporating into it –via the disagreement function– a reduced form of the conflict game. Having done that, we proceed to the characterization of the Hobbes solution.

3.1. Bargaining in the shadow of conflict

Suppose that there are *N* players, who wish to reach an agreement in $S^0 \in \mathbf{S}$, where $\mathbf{S}$ is the set of compact subsets of the utility[12] space, $\Re_+^N$. Assume further the existence of a *disagreement function*, *D(.)*, which assigns a disagreement point, *d*, to every compact subset of S^0. That is, if the set of alternatives considered were *S*, the outcome of disagreement would be $d = D(S)$. This mapping is to be interpreted as shorthand for the solution[13] to an underlying *conflict game*. We would like to stress that *D(S)* may depend on additional parameters, especially those related to the players' "strength", which form part of the

[11] Models of the conflictual resolution of opposing interests have been developed in areas such us growth, international trade, industrial organization, organizational design, patent races, or economics of litigation, to mention just a few.

[12] Actually, for our analysis it is not necessary that preferences satisfy the von Neumann-Morgenstern axioms. We could directly phrase our model in terms of money, prestige or the like.

[13] This solution maybe a unique Nash (subgame-perfect?) equilibrium, but uniqueness of equilibrium is not necessary. In case of multiplicity, the "disagreement outcome" can be defined as the meet of the utilities gained at the different equilibria.

description of this conflict game.[14] A bargaining problem in the shadow of conflict (BPSC) is then completely described by the pair $(S^0, D(.))$. Let B denote the set of all BPSCs. A *bargaining solution* for BPSCs is then a mapping, $f: B \to S$, satisfying $f(S^0,D(.)) \in S^0$. That is, the solution selects a subset of the alternatives as acceptable.

Note that, in principle, we need not impose any structure on *D(.)*, since it is meant to be a positive description of some real underlying conflict situation and therefore it cannot be freely chosen by the modeler. Nevertheless, to make the negotiation meaningful, we assume that disagreement can never be strictly Pareto optimal: for all $S \in S$, $\nexists s \in S$, such that $s \geq d = D(S)$.

3.2. The Hobbes solution

We require the Hobbes solution to satisfy a single axiom, based on the fundamental concept of individual rationality. In the context of a bargaining game that requires consensus to reach agreement, individual rationality implies that any solution should weakly Pareto dominate the disagreement outcome, since otherwise at least one player would prefer to provoke disagreement. The complement of the set of individually rational alternatives is then known not to be "eligible" for an agreement, so it is natural to expect that the shape/extension of this set should not affect the solution. Indeed, this is the unique assumption we make.

Let $S_x = \{s \in S \mid s \geq x\}$. That is, S_x is the subset of S which weakly Pareto dominates x. We impose the following axiom:

Independence of Individually Irrational Alternatives (IIIA):

$$f(S,D(.)) = f(S_{D(S)},D(.)) \text{ for all } (S,D(.)) \in B.$$

That is, the axiom requires that eliminating the feasible agreements which do not (weakly) Pareto dominate the disagreement point should not change the solution. Conceptually, IIIA is much weaker than Nash's Independence of Irrelevant Alternatives (IIA) axiom, since it only eliminates a subset of his "irrelevant alternatives". Thus, in a standard bargaining problem (SBP), IIIA simply eliminates the alternatives that are not individually rational with respect to

[14] An example of such conflict games is the family of games based on Tullock's (1980) *rent-seeking* model. In those games, the resolution of conflict is probabilistic, and the players can influence the probabilities via costly "effort." The players' strength is captured by their associated cost function.

the disagreement point.[15] However, in a BPSC the set of individually rational alternatives (coupled with the disagreement function) results, in general, in a different BPSC, to which the axiom also applies (note that, if $(S,D(.)) \in B$ then $(S_{D(S)},D(.)) \in B$ as well). Thus, as long as $D(.)$ is not constant (as in a SBP), the application of IIIA generates new BPSCs which, in turn, also have to satisfy the axiom. In view of all this, should we still find IIIA a plausible axiom? We certainly think so. The point of all "irrelevant alternatives" type axioms is to provide some consistency between solutions of the same underlying bargaining situation but with a different set of available agreements. In our view, the appropriate description of the bargaining situation should not be confined to a fixed disagreement point, since the outcome of disagreement is likely to depend on the alternatives available. Therefore, what should be kept fixed when carrying out the "consistency check" is the disagreement *function*, just as it is done in IIIA. That is, our assumption compares bargaining situations where the same set of players are bargaining in the shadow of the same conflict game but with different sets of feasible agreements.

We are now ready to define our solution concept.

Definition 1 *A bargaining solution to BSCPs is Hobbesian if and only if it satisfies IIIA.*

Our first result is that the requirements imposed on the solution are not too stringent – that is, there exist bargaining solutions that always select a non-empty set of agreements, satisfying them.

Proposition 1 *There exists a Hobbesian bargaining solution. Moreover, the set of Hobbesian agreements is always non-empty.*

Proof: Let $f_H(.,.)$ be a bargaining solution, $S^0 \in S$ an arbitrary bargaining set and $D(.)$ a disagreement function. Axiom IIIA implies that $f_H(S^0,D) = f_H(S^0_{d^0},D)$, where $d^0 = D(S^0)$. The disagreement point corresponding to the set $S^0_{d^0}$, however, is not d^0 but it is given by $d^1 = D(S^0_{d^0})$. Thus, the application of IIIA results in a new set, S^1. In general, repeatedly using the axiom, for the t-th iteration we will have

$$S^t = \{u \in S^{t-1} \mid u \geq d^{t-1}\}.$$

[15] Similarly, generically, no solution satisfies IIA in the context of a BPSC.

IIIA requires exactly that for all the elements of this sequence of sets, when coupled with *D(.),* the solution be the same. In other words, a bargaining solution satisfies IIIA if and only if $f_H(S^0,D) \subseteq S^* = lim_{T\to\infty} \bigcap_{t=0}^{T} S^t$. Note that, given the assumption that there always exist non-negative gains from agreement, the sets S^t are compact and nested. Therefore their intersection is uniquely defined and, by Tychonov's theorem, it is non-empty as well. Q.E.D.

Proposition 1 shows that, independently of the exact form it takes, just the conceptual increase in the informational content of the description of the bargaining problem is sufficient to provide us with a set of "acceptable" agreements. In general, these agreements are not unique. However, on the one hand, this indeterminacy of the solution may be perfectly acceptable in some situations. If, on the other hand, further refinement is required, there are two possible ways to proceed. Either we incorporate more information about the social situation into the description of the bargaining problem or we restrict the domain of BPSCs. In this paper, we take the latter route. We show that, for a broad class of disagreement functions, the above result can be strengthened: *the* unique Hobbes solution singles out a unique, Pareto efficient agreement. The assumptions we need to make are the following:

Assumption 1 *D is continuous in the Hausdorff topology: if a sequence of elements of* $\mathcal{S}$ *converges to S in the Hausdorff topology, then the corresponding sequence of disagreement points converges to D(S).*

Assumption 2 *Unless S is singleton, the disagreement outcome is strictly preferred to her worst agreement in S by at least one player: for all* $S \in \mathcal{S}^0$, such that $S \in \mathcal{S}$, *there exists* $z \in S$ *such that* $z_i < D_i(S)$ *for some* $i \in \{1,2,...,N\}$.

Assumption 1 is straightforward: it posits that small changes in the set of feasible utility allocations should not provoke major changes in the outcome of disagreement. Assumption 2 imposes that there exists some agreement to which at least one player strictly prefers the conflict outcome. That is, it requires that disagreement/conflict do not destroy all what is at stake, but it leave some positive part of the surplus for the players. This is a likely outcome if, for example, the surplus already exists without the cooperation of the players (as opposed to

gains from trade, for example). We will return to the relevance and meaning of this assumption after the proof of Proposition 2.[16]

Note that, for every $S^0 \in S$, the set of $D(.)$ satisfying Assumptions 1 and 2 is non-empty.

Proposition 2 *When Assumptions 1 and 2 hold, the Hobbesian bargaining solution is unique and it selects a unique and efficient agreement.*

Proof: To see that S^* has a unique element, note that, by the continuity of $D(.)$, $\lim_{t\to\infty} D(S^t) = D(S^*)$, and thus $S^* = S^*_{D(S^*)}$. Suppose that S^* is not a singleton. Then, by Assumption 2, $D(S^*)$ does dominate some points in S^*. Contradiction.

By construction, each set S^t contains the points of the weak Pareto frontier of S^0 that dominate $D(S^{t-1})$. Therefore, the point S^* is on the frontier of S^0. This proves the efficiency of the solution. Q.E.D.

In view of their critical role, let us discuss our assumptions on the disagreement function in more detail. Note first that without continuity, even in the presence of Assumption 2, the Hobbes solution could be set valued, since the sequence of disagreement points starting from d^0, might converge to an interior point of S. On the other hand, if we imposed that IIIA had to apply to the limit set, S^*, as well, we could drop the continuity assumption. However, as a principle, we prefer to put more structure on the (empirically testable) disagreement game rather than to increase our normative requirements (no matter how reasonable) on the solution.

Assumption 2 is a more delicate issue. It is immediate that, even if the disagreement function is continuous, without Assumption 2 the sequence of disagreement points mentioned in the proof may not converge to the Pareto frontier, and in fact, it may not even move away from d^0. In such a case, the Hobbesian set of agreements would still be well defined. Since the corresponding disagreement function yields less than the worst possible agreement as disagreement, the solution set would be the entire subset of S that dominates the players' outside options. This indeterminacy seems perfectly acceptable to us. It simply means that

[16] Esteban and Ray (1999) show that for a generalized version of the rent-seeking model (c.f. footnote 14), there always exists a *unique* Nash equilibrium and at this equilibrium each contending party expends strictly positive amounts of resources. It is straightforward to show that the disagreement point generated by the Nash equilibrium satisfies our Assumptions 1 and 2.

without further information, (like a given bargaining procedure) there is no basis to select any specific agreement. Nash's original demand game, for example, clearly supports this view. Even more importantly, however, we believe that Assumption 2 *is* satisfied in most settings. Examples abound: in pretrial bargaining the lawyer's fees are often set as a percentage of the amount under dispute; in collusive agreements in a market setting, even if there is cut-throat Bertrand competition, unless the firms are identical, there is always positive profits for the more efficient firm; in conflict models with endogenous choice of effort there is usually a unique interior Nash equilibrium, etc.

Nevertheless, there is an important class of situations, bargaining over the price to be paid for an object to be traded, which, in principle, violates Assumption 2. Under this scenario, if agreement is not reached, it is usually assumed that, since trade has not occurred, all the gains from trade vanish. In this specific setup, however, we need not consider the disagreement function as the outcome of a disagreement game. Instead, it should have the interpretation that it describes a kind of social norm, which imposes on the bargainers some minimal amounts of mutual concessions in order to show good faith.[17] Then, just as with the previous interpretation, unless (partial) agreements are superior to the "disagreement provoking point" they cannot be feasible since the negotiation would break up. We will return to this line of argument in Subsection 4.1.

4. Alternative characterizations of the Hobbes solution

In the previous section, we provided an axiomatic treatment of Bargaining Problems in the Shadow of Conflict. We have shown that the compounded effect of two innocent looking assumptions –namely, the existence of a disagreement function and the IIIA axiom– is sufficient to isolate a unique bargaining solution. Now, we follow the "Nash program" and complement our arguments with the exhibition of some plausible strategic bargaining games, for which it can be demonstrated that the Hobbes solution is their unique non-cooperative solution.

Corresponding to the fact that our analysis is now fully non-cooperative, we need to impose an additional assumption on the disagreement function. Namely we require $D(.)$ to satisfy the following monotonicity property:

Assumption 3 *Let* $\mathbf{x}$ *and* $\mathbf{y}$ *be elements of* S^0.

i) If $x_i \leq y_i$ *and* $\mathbf{x}_{-i} = \mathbf{y}_{-i}$ *then* $D_i(S_x) \leq D_i(S_y)$ *and* $D_{-i}(S_x) \geq D_{-i}(S_y)$, *while*

ii) if $x_i = y_i$ *and* $\mathbf{x}_{-i} \leq \mathbf{y}_{-i}$ *then* $D_i(S_x) \geq D_i(S_y)$ *and* $D_{-i}(S_x) \leq D_{-i}(S_y)$.

The first condition is quite natural, and one would expect it to be fulfilled in most applications. It simply states that, if we give part of the pie to a player before conflict over the remaining pie ensues, she must be no worse off and her adversaries should be no better off, than if conflict started without this transfer. For more than two players – note that when there are only two players, the two conditions coincide–, the second condition does entail a real loss of generality though. It states that, if we distribute part of the pie among a subset of the players before conflict, than all of the receptors are weakly better off, while the outsider is weakly worse off, than if the transfer has not been made. Note that, in general, since the relative magnitudes of the individual transfers are unconstrained, even a receptor could be hurt by the vector of transfers. This possibility is ruled out by the second condition of Assumption 3, everybody who receives a transfer should be better off in the conflict game. A relevant family of disagreement functions that satisfy both conditions is the one of homogeneous[18] *D(.)*.

Let $x \in S^0$ and $y = \{x_1, x_2, \ldots, x_i+\varepsilon, \ldots, x_N\} \in S^0$, with $\varepsilon \geq 0$. Define two sequences in S^0, a^t and b^t, as $a^1 = D(S_x), b^1 = D(S_y), a^t = D(S_{a^{t-1}}), b^t = D(S_{b^{t-1}}), t = 2,3,\ldots$. The following result will be used repeatedly in this section.

Lemma 1 *For all t,* $a_i^t \geq b_i^t$.

Proof: By Assumption 3 i), $a_i^1 \geq b_i^1$ and $a_{-i}^1 \leq b_{-i}^1$. We now proceed by induction. Assume that $a_i^t \geq b_i^t$ and $a_{-i}^t \leq b_{-i}^t$. Define $z = \{b_1^t, b_2^t, \ldots, a_i^t, \ldots, b_N^t\}$. Note that, $D_i(S_{b^t}) \leq D_i(S_z) \leq D_i(S_{a^t})$, where the first and the second inequality follows from Assumption 3 i) and ii), respectively. Similarly, we have that $D_{-i}(S_{b^t}) \geq D_{-i}(S_z) \geq D_{-i}(S_{a^t})$. Consequently, we have shown that $a_i^{t+1} \geq b_i^{t+1}$ and $a_{-i}^{t+1} \leq b_{-i}^{t+1}$, and the proof is complete. Q.E.D.

---

[17] Imagine, for example, the reaction of a prospective buyer of, say, a car when the vendor offers a discount of 1 dollar!

[18] Let $a > 0$ and $b \in \Re_+^N$. *D(.)* is homogeneous if and only if $D(aS+b) = aD(S)+b$, for all $S \in \mathbf{S}$.

4.1. A simple concession game with endogenous outside options

Consider the following multi-period game without discounting. In each period, the players –in some given order (possibly simultaneously)– state what is the lowest utility they are willing to accept at that time. These offers are binding. That is, once a player has made a claim, she cannot increase it later. Consequently, a vector of claims can be interpreted as a partial agreement.[19] In the same vein, offers can be (re)interpreted as concessions: if a player asks for x, then she is willing to accept any division which gives her at least x; she is willing to *concede* the rest to the other players.[20] Once every player has made a concession, each decides sequentially whether to continue negotiating or to provoke conflict over the remaining surplus. In case of conflict, they earn their partial agreement payoffs, plus their conflict payoff as determined by the disagreement function. The game ends[21] when either the offers are compatible (jointly feasible) or conflict is provoked.

Proposition 3 *When Assumptions 1-3 are satisfied, the Hobbesian agreement can be supported by a subgame-perfect equilibrium in the concession game with endogenous outside options.*

Proof: Take the following strategy profile, which does yield (in the limit) the Hobbesian agreement: each player concedes the sum of the corresponding disagreement payoffs of the rest of the players in each period and provokes conflict whenever her opponents concede her less than her current disagreement payoff. To see that this profile constitutes a subgame-perfect equilibrium note that, by conceding less in any round, conflict would ensue, which can give no better payoff. If a player conceded more then, by Lemma 1, she would not be able to improve her payoff either. Q.E.D.

Note that Proposition 3 does not guarantee that the Hobbesian agreement is the only subgame-perfect outcome. To ensure this, we need to make further assumptions. First, we strengthen Assumption 2, by imposing that the disagreement outcome should dominate the *status quo* for all players, not just one of them:[22]

---

[19] Note that, given a vector of offers, the point of partial agreement is always well defined, since it is the solution of N (linear) equations with N unknowns. Here is an example: Assume there are two players dividing a surplus of 1. If they propose, .7 and .8 then they concede .3 and .2, respectively. Thus the partial agreement is at (.2, .3).

[20] Note that this interpretation of incompatible claims is not new. It already appears in the Talmud (c.f. the "Contested Garment Principle" in Aumann and Maschler, 1985).

[21] The game may not end in finite time. This is not a problem, however, since we do not assume that time is valuable (no discounting).

[22] Beware that for some degenerate bargaining sets there may not exist $D(.)$ satisfying Assumption 2'.

Assumption 2' *Unless S is singleton, the disagreement outcome is strictly preferred to her worst agreement in S by all the players: for all* $S \in S^0$, such that $S \in S$, *there exists* $z \in S$ *such that z << D(S).*

Second, we impose a plausible behavioral assumption:

Assumption 4 *If in any period a player is offered less than her corresponding conflict payoff, she triggers conflict.*

A justification for this assumption can be a social norm which interprets it unacceptable to offer to the others –even as a partial deal– less than their outside payoffs.[23] We then have the following result:

Proposition 4 *When Assumptions 1, 2', 3 and 4 are satisfied, the concession game with endogenous outside options has a unique subgame-perfect equilibrium. In it, the Hobbesian agreement is reached via a sequence of partial agreements at the disagreement outcome sequence of the proof of Proposition 1,* d^t, $t=0,1,2,\ldots$.

Proof: First, note that the profile exhibited in the proof of Proposition 3 continues to be a subgame-perfect equilibrium (SPE) under Assumption 4. Therefore, we only need to show that there exists no other SPE outcome. Next, observe that disagreement cannot be supported by any SPE. To see this, note that provoking conflict –either by making too small a concession or by breaking up negotiations– can never be optimal,[24] since, by Assumption 2', continuing (and therefore making the minimal concession) is always strictly better.

Now, assume that there exists an agreement, different from the Hobbesian one, which can be supported by a SPE. Note that, by using the above strategy, any player can secure her Hobbesian payoff against the purportedly equilibrium strategies of the rest of the players. The key observation is that, by the previous argument, the other players will not provoke conflict. Therefore, little by little, they must yield the deviant player at least[25] her Hobbesian payoff.

---

[23] There is ample experimental evidence that people are willing to forgo economic gains in order to "punish" others –this is the so-called "spiteful behavior" (see Rabin (1993) and Levine (1998) for the theoretical background). Perhaps the study that best supports our assumption is Forsythe et al. (1994). In that paper it is shown that while in a standard "Ultimatum game" –one where a player can make a take-it-or-leave-it offer to the other about the division of a fixed surplus– the proposers give away a significant portion of the pie, if the game is made into a dictator game –where the receiver cannot veto the proposal– they tend to claim most of the surplus. Consequently, in the original game their generosity is not due to altruism or fairness, rather to fear of rejection.

[24] Note that, by the sequential nature of the opting-out stage, each player is pivotal.

[25] Note that, by Lemma 1, conceding more than necessary is always counterproductive. Thus, if the rest of the players concede more than the minimum in any period, it can only benefit the deviator.

Since the Hobbesian agreement is Pareto efficient, there always is a player who would prefer to deviate from the "equilibrium" supporting the alternative agreement.

The slowest rate at which the Hobbesian agreement can be reached is the one tracked by the sequence, d^t, since otherwise conflict would be triggered. In order to increase this rate, some player(s) have to concede more in at least one period. By the homogeneity of the disagreement function, no strict subset of the players would want to do so unilaterally, since at least one of them would end up with a lower payoff. Assume, therefore, that there is a coordinated increase in concessions. Since offers are binding, the last player to make a proposal would prefer to make the minimal concession instead, presenting her adversaries with a *fait accompli,* and improving her payoff. Q.E.D.

A salient example of a concession game with endogenous outside options, is a variant of final-offer arbitration, a widely used practice in Labor negotiations. In this game, the players make concessions repeatedly, with the option to call an arbitrator at any time. For simplicity, let us assume that the arbitrator keeps half of the remaining surplus and distributes the other half equally among the players. Recall that in line with the assumptions of this section, the arbitrator is always called when the concessions are not sufficient to give every player her expected share through arbitration. It is immediate that the Hobbesian agreement of this specific "bargaining game in the shadow of arbitration" is to divide the entire surplus equally. This, however, cannot be done in a single round. To see this, assume that there are two players and one of them concedes half the surplus in the first round. Following suit, the opponent would earn one half. Instead, she could deviate and concede one quarter. Note that this deviation cannot provoke the other player into calling the arbitrator, since by doing that he would earn one quarter, while ratifying the concessions, his continuation payoff would be at least 1/4 + 1/16 = 5/16, since he could always call the arbitrator in the following period (without making any further concessions). But then, the deviation is strictly profitable, since the deviator's continuation payoff is at least 1/2 + 1/16 = 9/16, which again can be assured by calling the arbitrator in the next period.

4.2. Negotiations with periodic ratification

Consider the following class of bargaining procedures. The players engage in some arbitrary form of exchange of proposals (utility demands) with the only requirement that each of them

should have a well-defined proposal at every point in time. At some (possibly random, but certainly finite) point, this exchange is interrupted and the players are asked (in sequence) whether they ratify their current proposals. If at least one player refuses to ratify, all the proposals made since the last ratification are null, and the players sequentially decide whether to provoke conflict. If they all decide to continue, they return to negotiation. On the other hand, if every player ratifies, a partial agreement is reached at the concessions embodied in the ratified proposals. Following a partial agreement, the players return to negotiate the division of the remaining surplus. This is again interrupted for ratification, and so on. We assume that, in case conflict arises when a partial agreement is already in place, that agreement is respected and the conflict is restricted to the distribution of the remaining surplus.[26] The game ends if either the players go to conflict or they get to a full agreement (that is, their current proposals are jointly feasible) or there is a perpetual lack of ratification following some partial agreement. As a typical element of this class, consider the Israeli-Palestinian negotiations.

Proposition 5 *When Assumptions 1-3 are satisfied, any member of the above mentioned family of bargaining games has the Hobbesian agreement as a subgame-perfect outcome.*

Proof: Take the following strategy: Player i never provokes conflict and offers the sum of the corresponding disagreement payoffs of the rest of the players in each period and refuses to ratify whenever her opponents offer her less than her current disagreement payoff. Note that if each player adheres to this strategy, they reach the Hobbesian agreement *and* the profile constitutes a SPE. To see the latter, note that, conceding less in any round would not change the payoffs, since the proposals would not be ratified. If a player conceded more or ratified an unfavorable partial agreement then, by Lemma 1, she could not improve her payoff. Finally, provoking conflict is also dominated by the equilibrium continuation. Q.E.D.

Note that Proposition 5 holds under Assumption 2. In order to obtain uniqueness, we need again Assumption 2' but we can do with a weaker version of the behavioral assumption made in the concession game above. Namely,

[26] Note, however, that even if the players were allowed to provoke total conflict at any stage of the game, they would not use this privilege.

Assumption 4' *If in any period a player is offered less than her corresponding conflict payoff, she refuses to ratify any concession (yet unratified) that she might have made.*

That is, given an unsatisfactory offer, conflict is not triggered, simply no concession whatsoever is made in return. We then have the following:

Proposition 6 *When Assumptions 1, 2', 3 and 4' are satisfied, the Hobbesian agreement is the unique outcome that can be supported by a subgame-perfect equilibrium in the periodic ratification game.*

Proof: First, note that the profile exhibited in the proof of Proposition 5 continues to be a SPE under Assumption 4'. Therefore, we only need to show that there exists no other SPE outcome. Next, observe that disagreement cannot be supported by any SPE. To see this, first note that provoking (perpetual) non-ratification, by making too small a concession, can never be part of an equilibrium, since, conditional on the others making the minimal concession[27] it is a unique best response to make it too. Given ratification, however, conflict is dominated.

Now, assume that there exists an agreement, different from the Hobbesian one, which can be supported by a SPE. Note that, by using the above strategy, any player can secure her Hobbesian payoff against the purportedly equilibrium strategies of the rest of the players. The key observation is that, by the previous argument, the other players will not provoke permanent non-ratification. Therefore, little by little (even if there are some periods of voluntary non-ratification), they must yield the deviant player at least her Hobbesian payoff. Since the Hobbesian agreement is Pareto efficient, there always is a player who would prefer to deviate. Q.E.D.

While Proposition 6 shows that there is a unique subgame-perfect agreement, the equilibrium path need not be unique. When there is a ratification stage, the moral hazard problem of the previous subsection does not arise, since upon a deviation the rest of the players can costlessly retract their concessions. As a result, coordination becomes feasible:

Corollary 1 *In the periodic ratification game, any subsequence of d^t, t=0,1,2,…, can serve as the series of partial agreements, including the full Hobbesian agreement in the first round.*

[27] Note that, by the sequential nature of the ratification stage, each player is pivotal.

For an example that shows the robustness of the above result to minor changes in the procedure, consider quantity-setting, non-differentiated oligopolists who are trying to collude in a market. Here, offers are self-imposed quantity caps, while disagreement is Cournot competition. In this setup a producer always has time to react to any increase in production of his competitors, before the market closes, exogenously ratifying the quantities. Consequently, by dividing up (equally) the monopoly quantity, the Hobbesian agreement is directly implementable, since all the producers know that by unilaterally increasing their production they would trigger a response by their competitors, making the deviation unprofitable.

4.3. Renegotiation-proof bargaining

In this subsection, we provide an independent justification of the Hobbes solution, based exclusively on the possibility to renegotiate the disagreement outcome. The concept of renegotiation has proved to be a powerful tool in implementation theory, greatly reducing the number of implementable outcomes. It is basically a concept of collective sequential rationality: the players as a group are assumed never to implement a Pareto inefficient outcome. This assumption reduces the severity of the feasible punishment strategies for deviant behavior, thereby destroying a number of potential equilibria. When the players cannot commit not to renegotiate a contract, this necessarily becomes incomplete. We take this observation to the extreme and actually assume that there is no contract signed. While it has been shown that such a "null contract" can actually be optimal from a mechanism design perspective (see Hart and Moore, 1999), our motivation is quite different, as we explain it below.

To date, renegotiation has always built upon bargaining theory, since it was considered (even in its etymology) as something that is posterior, more evolved than that. In this subsection, we invert this order of hierarchy. We derive a theory of bargaining from the mere possibility of renegotiation. At first blush, this may sound a bit circular: how can we have a theory of *re*-negotiation, before we have one of negotiation? Alternatively, in other words: how can we renegotiate a "null contract"? Nevertheless, this is where the Hobbesian approach proposed before comes to our rescue. For, even in the absence of a theory of agreement, we do have one of *dis*-agreement, embodied in the disagreement function. Therefore, we can meaningfully talk about our players renegotiating about how to *disagree* even if no contract is in place. That is, in case of disagreement, instead of implementing the corresponding disagreement outcome (which is presumably Pareto inefficient), they can agree

on a new disagreement outcome *and* return to the negotiation, since the one that got stalled was a different one. Now, what should be the new threat-point agreed upon? Well, our function *D(.)* readily provides us with a disagreement point. Of course, this new disagreement point is still inefficient, so further renegotiations are inevitable.

To shortcut such a sequence of renegotiations, Maskin and Moore (1987/1999) have introduced the concept of a *renegotiation function*, which is assumed to yield an efficient (and individually rational) outcome at every end-node of a game form. Then this function is applied to every outcome of the game-form, yielding a mechanism that is immune to renegotiation. Note that the use of the renegotiation function is more than just hiding a (possibly infinite) game in a black box. It incorporates into the game description that the players can individually compute where would the infinite game take them ("renegotiation is predictable") and can directly agree on the limit point.

Given how well the Hobbesian approach fits the implicit process of renegotiation, our obvious choice is to use the Hobbes solution as the renegotiation function for any BPSC.

Let a bargaining theory, $\boldsymbol{t}$, be a mapping from the set of BPSC to points in S^0. Call $\boldsymbol{t}$ renegotiation-proof if for every BPSC all players (weakly) prefer to implement the agreement prescribed by $\boldsymbol{t}$ to triggering a Hobbesian renegotiation. Since the Hobbes renegotiation function (as any other) yields a point on the Pareto frontier, the following proposition is straightforward.

Proposition 7 *In any extensive form game of negotiation where any player can unilaterally provoke total disagreement, the only possible (Hobbesian) renegotiation-proof bargaining theory is the one which prescribes the Hobbes agreement.*

Proposition 7 is not simple tautology. It serves the important purpose of translating some known mechanics of disagreement into a theory of agreement. That is, having identified a disagreement function, the mere fact that we assume that players can renegotiate any "outcome" of some non-cooperative game of negotiation identifies a unique candidate for an agreement, which will not be renegotiated.

There is an important remark in order. Note that, while it is true that varying the renegotiation function we can support different bargaining theories, this does not imply that our

solution is indeterminate. As we have emphasized in the previous section, the disagreement function is not a choice of the modeler, rather it is part of the description of the problem.

5. A comparative analysis

In this section, we clarify our theory by contrasting it to the papers and ideas, which are closest to it. While our approach is quite novel, there are a number of ideas in the literature to which it is related.

i) Hobbes and the theory of bargaining.

Binmore (1994) has also related Hobbes with bargaining theory. However, he identified Hobbes' state of nature with the *status quo* point, not with the disagreement point (as we do). The question here is not who is right and who is wrong. Simply, the different interpretations correspond to different social situations. Binmore has in mind a bargaining problem which is about possible improvements over an already existing contract. In that case, if the agents do not reach agreement, they continue respecting the old contract. We, on the other hand, are thinking of an incomplete contract scenario, where there is no fall back option and thus the conflict of interests must be resolved: either by consensus or by conflict.

ii) Endogenous determination of the disagreement point.

In his 1953 paper, Nash proposed a generalization of his original model of 1950. In this game, known as the "variable threat" model of bargaining, the players choose threats before the actual bargaining phase, of which they serve as the disagreement point. At first blush, our model may seem just like Nash's one, with a specific, well motivated threat game. Actually, however, our contribution goes well beyond that. There are two important differences between the models that we would like to underline:

a) Nash needs to employ an "umpire" to oblige the players to carry out their threats (in case of disagreement). We do without a $n+1^{st}$ party. The underlying reason for this is quite relevant. Nash thinks of the threat phase as one preceding the Nash bargaining game. Therefore, this phase has no interpretation on its own; it is simply a –perhaps realistic– way to make the bargaining game more detailed. In contrast, we think of our conflict subgame as one posterior to bargaining. By invoking sequential rationality, we can then analyze the players' optimal behavior in that subgame without any additional commitment device. Apart

from the obvious difference in philosophy, the technical difference is also apparent, since in Nash's game by a well-chosen threat (which she would prefer not to carry out) a player can improve on his share, without his bluff ever being called. Thus, even if we used our conflict game as the threat game, the equilibria would differ, since the players, in general, would not use a threat that forms part of an equilibrium of the conflict game.

b) When Nash's players generate a disagreement point, he considers the bargaining problem properly defined and proceeds to its solution (according to his 1950 paper). In contrast, we argue that they have simply arrived at a new bargaining situation, where they might wish to employ different threats than before. To put it another way: while in the Nash model the demand phase depends on the outcome of the threat phase, in our model the conflict game is supposed to depend on the demands made (when they are not compatible).

iii) Step-by-step resolution.

Kalai (1977) introduced the axiom of decomposability. This assumption requires that if we break up the set of available agreements, *S*, into two subsets, *X* and *Y*, then using the solution of (either) one of these as a partial agreement to subsequently bargain over the rest, $(S\text{-}f(X, d)) \bigcap \Re_+^N$, should give the same result as applying the solution directly. Note that Kalai's model agrees with ours in the idea that partial agreements are only renegotiated if this yields a Pareto improvement. On the other hand, Kalai does not propose a well-defined solution: he only establishes that the solution should be "proportional," without identifying what should these proportions be. In addition, Kalai's model has two caveats, first pointed out by Ponsati and Watson (1997). The first of these is that when agreeing on the first sub-problem, the bargainers of Kalai are not supposed to take into account the effect of today's agreement on tomorrow's one. This is not true in our model. Second, there seems to be an inconsistency between the assumption that the agreement on the first subproblem is binding, but at the same time can be renegotiated –since the second sub-problem is not $S \backslash X = Y$ but $(S\text{-}f(X, d)) \bigcap \Re_+^N$. In our model, however, these two sets coincide so we avoid any confusion.

iv) Additional parameters, representing the players' bargaining power.

Our explicit modeling of the conflict game allows us to introduce information about the players' "strength" to the bargaining problem. The closest reference here is the asymmetric Nash

solution (see Harsányi and Selten, 1972). This concept incorporates an additional exogenous vector of parameters to the bargaining problem, meant to represent the players' relative ability to bias the agreement in their favor. We have two important comments to make:

a) To our knowledge, it has never been argued that the disagreement point and the bargaining weights should be jointly determined.[28]

b) Even if this restriction were imposed on the asymmetric Nash solution, the Hobbes solution would not be a special case. That is, while for a given S^0 and $D(.)$, –with $d = D(S^0)$ – there always exist bargaining weights which make the Nash and the Hobbes solution coincide, if we vary S^0 the bargaining weights leading to equivalence also change.

v) An auxiliary function, which maps each bargaining problem into a point in the utility space.

Thomson (1981) introduced the concept of a *reference function*, the purpose of which is to summarize the relevant features of a bargaining problem. This function maps every bargaining problem into a reference point, which is then used to calibrate the bargaining power of the players. While, at first blush, our disagreement function may sound just like a special case, actually the two approaches are diametrically opposed. The role of a reference function is to summarize information that is already present in the bargaining problem. In contrast, our approach complements the originally available information with the outcome of conflict, which possibly depends on additional factors.

vi) Bargaining under the threat of some outside enforcement mechanism.

This topic has been extensively dealt with in the applied literature (pretrial negotiations, strikes, arbitration etc.). Perhaps, the piece closest to our approach is Powell (1996). Powell sets up a non-cooperative bargaining game where the players can choose to force a (probabilistic) settlement at some cost. The important difference with respect to our approach is that, in his model, forcing the settlement is equivalent to taking an *outside option*. However, outside options do not determine, in general, the outcome of a bargaining game. Therefore, Powell needs to rely on the solution to the bargaining game, which would come about in the absence

[28] The bargaining weights are often interpreted as parameters representing negotiating "skill." We do not find such an assumption justified in the framework of a normative theory based on fully rational players playing a game of complete information.

of outside options. In our case, in contrast, the solution of the game cannot be dissociated from the underlying conflict situation.

vii) Recursive solutions.

We are not the first ones to use a recursive application of some rule in bargaining theory. Let us mention just a couple. Raiffa (1953) proposes a method where the players first pocket half of their most preferred allocation, then half of their most preferred allocation in the remainder... etc. While in (its recursive) structure his procedure is very much like ours, the important difference is that he has no justification other than some vague consideration of "fairness" for the fifty percent rule. van Damme (1986) considers a recursivity axiom which imposes that if the players are making demands according to some individual theories, then in every step of the iteration, as a function of these demands some subset of S is to be discarded, and the negotiation resumed. Technically, the IIIA assumption is very similar, with the important difference that we only invoke individual rationality for discarding "irrelevant alternatives."

viii) The local shape of the Pareto frontier matters.

The Hobbes solution relaxes Nash's Independence of Irrelevant Alternatives (IIA) axiom to a large extent, since an infinite number of (endogenously determined) points of the Pareto frontier affect it. While most non-Nash bargaining solutions also relax IIA, the one that comes closest to ours in this respect is the Perles and Maschler (1981) solution. According to this concept, the players start at their most preferred outcome and trace the Pareto frontier by simultaneously lowering their demands at the speed that corresponds to the slope of the Pareto frontier at their current proposal. Our main criticism of the Perles-Maschler solution is that, while the rate of concession equaling the "marginal rate of substitution" is an appealing idea, it continues to be arbitrary.

ix) Disagreement modeled as a non-cooperative game.

Busch and Wen (1995) introduce a normal form game to determine the period payoffs in case of a *temporary* disagreement in an alternating offers bargaining game with flow payoffs. The incorporation of a disagreement game in their context has markedly different consequences than in our model. Namely, the issue in their case is that given the repeated nature of the strategic interaction, they obtain multiple equilibria in many cases. These equilibria are supported by a history dependent variation of the play in the disagreement game. However,

when the payoffs of the disagreement game are such that no multiplicity is generated, the outcome is qualitatively similar to Rubinstein's (1982).

6. Concluding remarks

In this paper we have presented a new approach to the theory of negotiation and have introduced the corresponding agreement concept. The two cornerstones of our theory are on the one hand, that we consider the disagreement points of the bargaining problem endogenous; and on the other hand, that we conceive of bargaining as a sequence of partial agreements. We have shown that coupling these ideas with a mild generalization of individual rationality one can obtain a natural agreement solution. Our results prove the power of focusing on the "state of nature" in order to understand social agreements, as proposed by Hobbes.

The question whether the sequence leading to agreement is interpreted as a succession of partial agreements or as one of partial disagreements is not just semantic. With our approach, we were able to derive a bargaining solution based on the knowledge of what a once for all disagreement would bring about. The *dual* approach, however, would not be able to derive a solution exclusively based on the outcome of partial disagreements, it would also need as input the solution of the problem for at least some subset of the bargaining set. (c.f. point *vi)* in the previous section.) From the non-cooperative view, this caveat of the dual approach can be overcome by considering a specific extensive form game: see Rubinstein's (1982) seminal solution based on time preference. There, if players do not agree, disagreement is only partial: some utility is destroyed (by delaying the agreement) and the players return to negotiation. While time preferences are useful in determining the losses due to partial disagreement, it is less than obvious that, in practice, they really are the driving force to agreements. We consider it an achievement that our model needs not employ such a construction.

Grossly oversimplifying our approach, one could try to argue that all we are doing is to substitute the description of a specific bargaining procedure with the description of a conflict game. So, what is the gain? We claim that there is an important difference, since the proper mechanics of negotiation have proven to be very elusive,[29] while modeling conflict turned out to be much more straightforward (c.f. Section 2). Moreover, our "roundabout" approach

[29] Largely due to the many behavioral issues that are difficult to consider in a mathematical model.

actually did prove useful in taking steps towards finding a qualitatively right strategic model of bargaining (c.f. Subsections 4.1 and 4.2).

In the same vein, while our main characterization is carried out in an axiomatic –and, therefore, intrinsically normative– framework, we consider our overall approach positive. We do not inquire about which "ought to be" the terms of an agreement. Instead, we ask the modeler for an enhanced description of the bargaining problem and derive our solution based on that, and a single axiom, which incorporates the notion that players are (individually) rational. That is, considering the trade-off between imposing axioms versus detailing the social situation, we decidedly favor the second option.

Finally, one could argue that the fact that our theory does not ever predict disagreement decreases its credibility; after all, we do observe wars, strikes and litigation in real life. Note however, that our model is one of complete information (unlike the real world), while disagreement should be expected to arise in situations of asymmetric information (see, for example, Crawford, 1982). Extending our approach to an incomplete information setting certainly seems to be an intriguing and worthwhile –as well as challenging– task.

References

Aumann, R. and M. Maschler (1985), "Game Theoretic Analysis of a Bankruptcy Problem *Journal of Economic Theory* 36, 195-213.

Becker, G. (1983), "A Theory of Competition among Pressure Groups for Political *Quarterly Journal of Economics* 98, 371-400.

Binmore, K. (1994), *Playing Fair: Game Theory and the Social Contract*, The MIT Press, Cambridge (Mass.).

Busch, L-A. and Q. Wen (1995), "Perfect Equilibria in a Negotiation Model," *Econometrica* 63(3), 545-565.

Chen, M. and E. Maskin (1999), "Bargaining, Production, and Monotonicity in Economic Environments", *Journal of Economic Theory* 89, 140-147.

Crawford, V. (1982), "A Theory of Disagreement in Bargaining," *Econometrica* 50(3), 607-637.

van Damme, E. (1986), "The Nash Bargaining Solution is Optimal," *Journal of Economic Theory* 38, 78-100.

Esteban, J. and D. Ray (1999), "Conflict and Distribution," *Journal of Economic Theory* 87, 379-415.

Forsythe, R., Horowitz, J., Savin, N. and M. Sefton (1994), "Fairness in Simple Bargaining *Games and Economic Behavior* 6, 347-369.

Gauthier, D. (1990), *Moral Dealing. Contracts, Ethics and Reason*, Cornell University Press, Ithaca N.Y..

Grossman, H.I. (1991), "A General Equilibrium Model of Insurrections," *American Economic Review* 81, 912-921.

Grossman, H.I. (1994), "Production, Appropriation and Land Reform," *American Economic Review* 84, 705-712.

Harsányi, J. and R. Selten (1972), "A Generalized Nash Solution for Two-Person Bargaining Games with Incomplete Information," *Management Science* 18, 80-106.

Hart, O. and J. Moore (1999), "Foundations of Incomplete Contracts," *Review of Economic Studies* 66(1), 115-138.

Hirshleifer, J. (1991), "The Paradox of Power," *Economics and Politics* 3, 177-200.

Hirshleifer, J. (1995), "Anarchy and its Breakdown," *Journal of Political Economy* 103, 26-52.

Horowitz, A. (1993), "Time Paths of Land Reform: A Theoretical Model of Reform Dynamics," *American Economic Review* 83(4), 1003-1010.

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

Kalai, E. (1977), "Proportional Solutions to Bargaining Situations: Interpersonal Utility *Econometrica* 45(7), 1623-1630.

Knight, J. (1992), *Institutions and Social Conflict*, CUP Cambridge.

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

Maskin, E, and J. Moore (1999), "Implementation and Renegotiation," *Review of Economic Studies* 66(1), 39-56.

Nash, J. (1950), "The Bargaining Problem", *Econometrica* 18, 155-162.

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

Osborne, M.J. and A. Rubinstein (1990), *Bargaining and Markets*, Academic Press, San Diego.

Perles, M. and M. Maschler (1981), A Super-Additive Solution for the Nash Bargaining Game," *International Journal of Game Theory* 10, 163-193.

Ponsati, C. and J. Watson (1997), "Multiple-Issue Bargaining and Axiomatic Solutions," *International Journal of Game Theory* 26, 501-524.

Powell, R. (1996), "Bargaining in the Shadow of Power," *Games and Economic Behavior* 15, 255-289.

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

Raiffa, H. (1953), "Arbitration Schemes for Generalized Two-Person Games," in Kuhn and Tucker (eds.) *Contributions to the theory of games II*, Annals of Mathematics Studies #28. Princeton University Press.

Roemer, J. (1988), "Axiomatic Bargaining Theory on Economic Environments" *Journal of Economic Theory* 45, 1-31.

Roemer, J. (1996), *Theories of Distributional Justice*, Harvard University Press, Cambridge (Mass.).

Rousseau, J.J. (1782), *Discours sur l'Origine et les Fondements de l'Inégalité parmi les Hommes*, London.

Rubinstein, A. (1982), "Perfect Equilibrium in a Bargaining Model," *Econometrica* 50, 97-109.

Skaperdas, S. (1992), "Cooperation, Conflict, and Power in the Absence of Property *American Economic Review* 82, 720-739.

Smith, A. (1776), *The Wealth of Nations*.

Taylor, M. (1987), *The Possibility of Cooperation*, Cambridge University Press.

Thomson, W. (1981), "A Class of Solutions to Bargaining Problems," *Journal of Economic Theory* 25, 431-441.

Thomson, W. (1994), "Cooperative Models of Bargaining," Chapter 35 in Aumann and Hart (eds.) *Handbook of Game Theory*, Elsevier Science.

Tullock, G. (1980), "Efficient Rent Seeking," in J.M. Buchanan, R.D. Tollison and G. Tullock (eds.) *Toward a Theory of the Rent-Seeking Society*, College Station: Texas A&M University Press, 97-112.