Synthese (2020) 197:407–425
https://doi.org/10.1007/s11229-018-1729-4

Bargaining and the dynamics of divisional norms

Justin P. Bruner[1]

Received: 1 September 2017 / Accepted: 12 February 2018 / Published online: 19 February 2018

Abstract Recently, philosophers have investigated the emergence and evolution of the social contract. Yet extant work is limited as it focuses on the use of simple behavioral norms in rather rigid strategic settings. Drawing on axiomatic bargaining theory, we explore the dynamics of more sophisticated norms capable of guiding behavior in a wide range of scenarios. Overall, our investigation suggests the utilitarian bargaining solution has a privileged status as it has certain stability properties other social arrangements lack.

Keywords Social contract theory · Game theory · Bargaining theory · Axiomatic bargaining · Social norms · Evolutionary game theory · David Gauthier · Utilitarianism · John Nash

1 Introduction

Broadly speaking, there are two influential approaches to the social contract. The more widely known of these considers the social arrangements rational and self-interested agents would agree to. This is the tradition of Hobbes, continued to this day by David Gauthier and James Buchanan, among others, and is typically discussed in terms of the theory of rational choice. An alternative approach, which harkens back to David Hume, conceives of the social contract as continually evolving. Work in this latter tradition focuses primarily on better understanding social contract formation and how beneficial social arrangements are maintained and modified over time. Contemporary

✉ Justin P. Bruner
justin.bruner@anu.edu.au

[1] School of Politics and International Relations, The Australian National University, Canberra, Australia

contributors such as Brian Skyrms, Robert Sugden and Kenneth Binmore continue Hume's proto-evolutionary exploration of the social contract with a suite of modern tools from the social and evolutionary sciences. This paper contributes to this latter 'dynamic' tradition. It will do so, however, by drawing on insights from the 'static' approach of Hobbes and Gauthier. Within philosophy, the dynamic and static approaches to the social contract have developed in relative isolation of each other.[1] I believe there are unappreciated benefits to integration.

We restrict our attention to one of the canonical problems of the social contract, that of division.[2] The problem of division pertains to how the benefits and burdens of social cooperation are to be distributed. Recently, tools from the social sciences and game theory have been brought to bear on this topic. Braithwaite (1955) was the first to note that the theory of games could be usefully employed by moral and political philosophers interested in social contract theory and distributive justice and drew on the then newly developed axiomatic theory of bargaining developed by John F. Nash. Nash's approach to the problem of division has been more recently utilized by political philosophers such as David Gauthier to frame the debate as to how rational individuals will divide the fruits of social cooperation.

As mentioned, a second strand in the literature tackles the social contract from a different perspective by drawing on models of biological and cultural evolution to investigate potential solutions to the problem of division. In particular, this work has focused on better understanding the emergence of divisional norms in simple bargaining scenarios. Skyrms (1996), for instance, has demonstrated that the even split is the likely outcome of bargaining under canonical models of cultural and biological evolution. Alexander (2007) has explored the surprising effect network structure has on the emergence of fair-dealing.[3]

Yet these accounts, while illuminating, deal with extraordinarily simple bargaining scenarios (or bargaining *games*) and even simpler bargaining solutions (or 'divisional norms').[4] For instance, agents considered by Skyrms are confronted by a scenario (the so-called 'mini-Nash demand game') in which they and their counterpart simultaneously demand either 4, 5 or 6 units of a resources from a total of 10 available units. Others have explored similarly simple strategic scenarios.[5] Missing from this picture

[1] This is in part due to the fact that the dynamic and static approach have different targets. Authors in the static approach to the social contract tend to have justification in mind. This is not always the case for those writing in the dynamic tradition, although Kenneth Binmore and Gerald Gaus may be two notable exceptions.

[2] This is of course just one of a slew of issues relating to the social contract. For instance, Skyrms (2013) conceives of the social contract as involving three stages. The first pertains to determining who to interact with (i.e., the problem of partner choice), the second pertains to whether the agents involved successfully cooperate (the problem of collective action or the 'stag hunt'), and the third addresses how to divide the fruits of cooperation (the problem of division). For simplicity, we focus only on this last problem. For more on these latter two steps, see Wagner (2012) and Bruner and O'Connor (2017).

[3] For more on the evolutionary approach to the social contract, see Bruner (2015), Zollman (2009), Sugden (1986) and Vanderschraaf (2016, forthcoming).

[4] We use the term 'norm' here rather loosely and do not presuppose a particular account of norms (although see Bicchieri 2006 for a rich theory of norms and conventions).

[5] See Bruner (2017), Huttegger and Smead (2011) and Wagner (2012).

are more sophisticated, all-purpose norms that can be flexibly applied to a variety of problems. As a result, extant work on the social contract has been limited to exploring the emergence of divisional norms in *very specific strategic contexts*, without an eye toward better understanding how such norms can be appropriated and employed in alternate scenarios.[6]

To address this oversight in the literature, we look to the 'static' tradition which has, over a number of decades, developed an elaborate and axiomatic theory of bargaining. In particular, axiomatic bargaining theory is helpful for it provides us with a suite of divisional norms untethered to a particular strategic scenario and can thus be fruitfully employed across a wide range of settings. We explore this further in the next section and introduce a few of the canonical bargaining solutions from philosophy and economics. We then consider the dynamics of these more sophisticated divisional norms.[7]

We find traditional stability concepts are somewhat limited and unable to provide a detailed picture of the dynamics of these more sophisticated divisional norms. After establishing these limitations in Sect. 3, we then consider two possible ways of proceeding. First, we draw on recent work in bargaining theory which provides an account of how play proceeds in the face of disagreement. In particular, individuals engage in so-called metabargaining, which, under certain conditions, results in agreement between initially conflicting bargaining recommendations. Our analysis indicates that when metabargaining is allowed, the utilitarian bargaining solution has a somewhat privileged status in the sense that it has certain stability properties alternative divisional norms lack.

In Sect. 5, we consider an alternative approach to the evolution of divisional norms inspired by Gauthier's 'conditionally just' agent from *Morals by Agreement*. Gauthier considers an individual who conditions her behavior on that of her counterpart, advancing Gauthier's favored bargaining solution (the minimax relative concession solution) when dealing with similarly disposed individuals, and acting so as to maximize expected utility when interacting with all others. We demonstrate that under certain conditions, Gauthier's community of just agents can be destabilized by those relying on alternative conditional bargaining solutions. This suggests Gauthier's choice to focus solely on the egoist when assessing the stability of the social contract is misguided. Diversity, in the form of alternate divisional norms, is also capable of destabilizing a community populated by Gauthier's conditionally just agents. Finally, we show that (conditional) utilitarians once again have certain stability properties competing divisional norms lack.

Taken together, then, this paper outlines two plausible avenues social contract theorists may venture down when considering the evolution of divisional norms. While

[6] Barrett (2014) has conducted work exploring how different signaling conventions can be appropriated to function in unfamiliar environments. I do not know of any formal work in political philosophy or ethics taking this approach, although Bednar and Page (2007) and Zollman (2009) are in the vicinity.

[7] Although we do not address this in much detail, it is also possible to see the present paper as a contribution to the 'static' approach to the social contract. In particular, we are providing an account of how one of the many rationalizable social contracts will come to fruition via cultural-evolutionary dynamics. This in some ways is similar to recent work on the social contract by Gaus (2010), who argues there are a variety of publicly justifiable social contracts and cultural evolution plays a crucial role in selecting a uniquely justified set of social-moral rules to govern the behavior of a community. See also Moehler (2018).

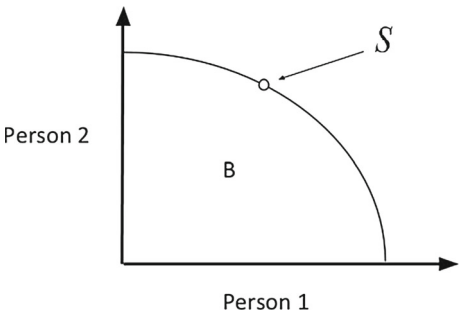

Fig. 1 Simple two-person bargaining problem. Example bargaining solution *S*

there are significant differences, both lead to the same tentative conclusion regarding the centrality of the utilitarian bargaining solution. That said, we take our analysis to be best conceived of as a proof of possibility, establishing a few means by which the dynamic and static traditions in social contract theory can be integrated so as to provide a richer account of the dynamics of divisional norms.

2 Bargaining games and bargaining solutions

Modern bargaining theory goes back to John Nash's famous work on bargaining and is primarily concerned with how rational agents divide goods amongst themselves (Nash 1950). Nash's axiomatic approach stipulates that such divisions should have certain characteristics. In this section, we flesh out this approach and discuss a few popular bargaining solutions which have been floated by Nash and others.

A *bargaining problem* consists of two elements: a disagreement point and a set of feasible alternatives. The disagreement point specifies how both individuals fare if bargaining breaks down and no agreement is made.[8] The set of feasible alternatives refers to all possible outcomes the agents could jointly bring about. A *bargaining solution* is very simply a function which takes as input a bargaining problem (a disagreement point and a set of feasible alternatives) and returns an element of the feasible set. In other words, using only the disagreement point and the set of possible arrangements, a bargaining solution specifies which, of the available alternatives, rational agents will settle on (see Fig. 1 for an illustration).

We first consider the so-called Nash bargaining solution advocated by Nash in his influential 1950 paper (Nash 1950). Simply put, the Nash bargaining solution selects the element of the feasible set which maximizes the product of the bargainers' payoffs (sans the disagreement point). Nash proved that this bargaining solution satisfies, among other things, *Pareto* and *Independence of Irrelevant Alternatives* (IIA). A bargaining solution satisfies Pareto if it always selects an outcome that is Pareto efficient, meaning it is not possible to make both bargainers better off by selecting an alternative element in the feasible set. IIA is a bit more difficult to describe. Consider the following: *A* and *B* are two bargaining problems sharing the same disagreement point and the feasible set of *B* is a subset of the feasible set of *A*. Let *c* be the division

[8] We in the course of this paper consider the simplest case involving just two bargainers.

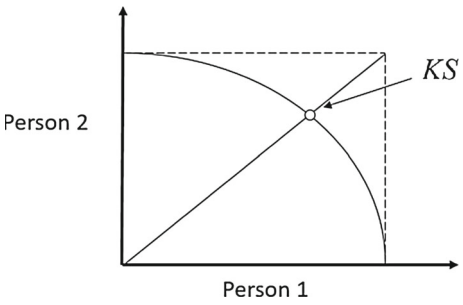

Fig. 2 Simple two-person bargaining problem. K–S solution is found by determining where the line formed by connecting the disagreement point with the ideal point intersects the pareto frontier

selected by a bargaining solution when applied to bargaining problem *A*. If *c* is also an element of the feasible set of *B*, the bargaining solution satisfies IIA if, when applied to bargaining problem *B*, it also recommends *c*. In other words, if an arrangement *c* is selected by the Nash bargaining solution, then *c* should still be chosen even if elements of the feasible set other than *c* are removed from the bargaining problem.

IIA is considered to be somewhat controversial,[9] and for this reason various alternative bargaining solutions not underwritten by IIA have been suggested. One such solution is the famous Kalai–Smorodinsky bargaining solution (K–S). Briefly, K–S replaces IIA with the axiom *Monotonicity*, which states that if the set of feasible alternatives is expanded, no agent should receive less than what they would have received under the original, more restricted feasible set. Graphically, the K–S bargaining solution can be identified by connecting the disagreement point with the 'ideal point' (the often infeasible arrangement in which neither agent makes a concession to their counterpart). The K–S solution lies at the intersection of this line segment and the bargaining frontier (see Fig. 2). Note that Gauthier's famous minimax relative concessions solution is formally equivalent to the K–S solution for situations involving only two bargainers.

Finally, we also consider both the egalitarian and utilitarian bargaining solutions within this framework. The egalitarian solution satisfies both *IIA* and a version of *Monotonicity* and stipulates that the solution should be one which equalizes the gain both agents receive above the disagreement point. The utilitarian bargaining solution satisfies *IIA* and picks the outcome which maximizes the joint gain the agents receive above the disagreement point.

Before we go on to consider the dynamics of these four popular bargaining solutions, a quick word as to how we interpret the elements of the feasible set. Typically, elements of the feasible set are interpreted as corresponding to utility pairs specifying the utility both agents involved in the bargaining process receive. We instead interpret the elements of the feasible set as the level of resources bargainers are able to secure.[10] We deviate from the assumption that bargaining outcomes represent utilities because some bargaining solutions are committed to the possibility of meaningful interper-

[9] See, however, Thrasher (2014) for an argument as to why the seemingly innocuous 'symmetry' axiom may be problematic.

[10] We also assume, as Alexander and Skyrms (1999) do, that the resource in dispute is homogeneous and divisible.

sonal utility comparisons while other solutions deny such comparisons are possible. If our goal is to explore the stability and dynamical properties of both kinds of bargaining solutions (as we do in Sects. 4, 5), a common currency must be agreed to. Thus for the remainder of this paper we consider 'resource' versions of these four famous bargaining solutions, whereby the disagreement point and elements of the feasible set refer to the (possible) resource holdings of the two bargainers.

Finally, note that while these solutions do not satisfy the same axioms, they all speak to a concern raised in the previous section regarding extant work on the social contract. The above bargaining solutions are compelling (in part) because they can be fruitfully applied in a wide variety of distinct bargaining scenarios. As mentioned, prior work on the dynamics of divisional norms dealt instead with rigid bargaining solutions that could only fruitfully be utilized in the context of a particular strategic scenario (such as 'demand 5' in the mini-Nash demand game). We explore the dynamics of these more general divisional norms in the following sections.

3 Dynamics of divisional norms

Now that we've laid out a number of popular bargaining solutions, we can perhaps say something about their respective stability properties. First, however, we must specify what, exactly, the underlying strategic interaction of interest is. To state this precisely we introduce the notion of a *bargaining bundle.* A bargaining bundle is simply a set of bargaining problems accompanied by a probability distribution determining how likely a particular bargaining problem is to arise. Agents then employ their preferred bargaining solution (i.e., their 'strategy') to navigate the bargaining problem (selected from the bargaining bundle) they are presented with. If the bargaining solutions employed by two agents deliver the same recommendation, bargainers receive the corresponding payoff. If bargaining solutions conflict, no agreement is made and the disagreement point obtains. Hence, the strategic scenario of interest (i.e., the 'game') is the bargaining bundle and the 'strategies' agents have at their disposal are the various bargaining solutions discussed in the previous section. As we will observe in Sect. 4, quite a lot can be said about the stability properties of these various strategies even when we do not specify a particular bargaining bundle.

With this in hand, we now turn our attention to the notion of an evolutionarily stable strategy (ESS). A strategy is said to be an ESS if, very bluntly, a population composed entirely of those employing this strategy cannot be successfully invaded by a 'mutant' utilizing an alternative strategy. This notion of stability is commonly drawn on in both the biological and social sciences and is employed with the aim of better understanding the endpoints of a dynamic process.[11] More precisely, we say a strategy (s) is an evolutionarily stable strategy if for all alternative strategies (s'), the following two conditions are satisfied:

(i) $\pi(s, s) \geq \pi(s', s)$, and
(ii) if $\pi(s, s) = \pi(s', s)$ then $\pi(s, s') > \pi(s', s')$,

[11] For a nuanced discussion on the limitations of ESS methodology see, Huttegger and Zollman (2013).

where $\pi(s, s')$ refers to the payoff an agent playing strategy s receives against a counterpart utilizing strategy s'. In other words, natives must at least do as well as invaders when paired against natives. If they do equally well, then natives must do better than invaders when pitted against an invader strategy. These two conditions jointly ensure that a small handful of invaders will fare worse than the natives, thereby preventing their expansion. This stability concept has been used quite extensively to investigate a number of philosophical issues, from the conventionality of meaning to the social and moral emotions.

A weaker notion of stability, so-called neutral stability, is almost identical to evolutionary stability except the inequality of condition (ii) is changed from a strict inequality to a weak inequality (i.e., the consequent of the conditional in (ii) is now $\pi(s, s') \geq \pi(s', s')$). The difference between these two stability concepts is straightforward: if a strategy is an ESS then all potential invaders will quickly be jettisoned from the community (since natives always outperform small handfuls of invaders). If, on the other hand, a strategy is said to be an NSS (neutrally stable strategy) it is possible for invaders and natives to coexist. Neither does better than the other.

Returning to our bargaining scenario, it is easy to show that all strategies (i.e., bargaining solutions) are *at least* neutrally stable for all bargaining bundles. Without loss of generality, consider the egalitarian bargaining solution. A community of egalitarians can resist invasion if the recommendations of the intruding bargaining solution does not coincide with that of the egalitarian solution. In this case, the disagreement point will be reached, meaning the invader will do worse against egalitarian natives than the natives do against themselves, satisfying condition (i) and (ii).

If, however, the intruding bargaining solution makes the same recommendations as the egalitarian bargaining solution for all bargaining problems in the bargaining bundle, there will be no difference in payoff between invader and native, meaning the egalitarian strategy is neutrally stable. Since this is the best an invading strategy can do against a population of egalitarians, the egalitarian bargaining solution is *at worst* neutrally stable.

Note that this holds for *all* bargaining solutions considered in the previous section. Since all strategies are at least neutrally stable for all bargaining bundles, this evolutionary analysis on its own provides us with little insight as to which of these divisional norms is likely to go to fixation.[12] The 'spinelessness' of the ESS and NSS stability concepts is due to the fact that when two different bargaining solutions make conflicting recommendations the disagreement point is reached. If instead the disagreement point could somehow be avoided, a dynamic analysis may have real bite.

One intuitive suggestion is to posit some fairness norm the agents appeal to when their bargaining solutions provide conflicting recommendations. This could be something like 'flip a coin to determine which bargaining solution to adopt' or somehow 'split the difference' between the two recommendations. Yet this is not satisfactory as it presupposes the very thing we're interested in providing an explanation of: the

[12] This is due to the fact that the norm which ends up being widely adopted will in large part simply hinge on the initial composition of the population. If egalitarians are for whatever reason initially prevalent then they are likely to go to fixation since more of their interactions result in the avoidance of the disagreement point.

emergence of divisional norms. Instead, we consider in the next section a so-called 'metabargaining' approach in which individuals, when confronted by disagreement, transform the feasible set in a way which tracks their shared commitments and then reapply their preferred bargaining norm. As we will show, this metabargaining procedure has some desirable properties. In what follows, we spell out this metabargaining procedure in more detail and discuss the evolutionary stability of our four bargaining solutions under this procedure. We then turn to Gauthier's work on bargaining and the social contract and, in Sect. 5, suggest this is another means by which we can construct a dynamic account of divisional norms.

4 Metabargaining

As mentioned, when two individuals use different bargaining solutions offering conflicting recommendations, the disagreement point is reached. We now consider a two-step procedure which can allow agents to avoid this sub-optimal outcome. This procedure is inspired by a similar approach taken by van Damme (1986).[13] The first step involves altering the feasible set of the initial bargaining problem (in a way to be made precise and argued for below). Second, after the feasible set is altered, the bargainers then reapply their bargaining solution to this new bargaining problem. If disagreement persists, this two-step procedure is repeated. Taken together, this procedure often enables individuals to avoid the disagreement point, but by no means guarantees the two bargainers will converge on a solution. One way to conceive of this procedure (and metabargaining generally) is as follows: an agent's preferred bargaining solution can be thought to reflect what the individual takes a fair division to be. Metabargaining, then, involves continued bargaining over a modification of the feasible set, where outcomes both bargainers deem to be patently unfair are jettisoned from the feasible set. We spell out the details of this below.[14]

Consider a bargaining problem with bargainers Bob and Rob. Bob appeals to bargaining solution S_1 and Rob appeals to bargaining solution S_2. S_1 recommends Bob receive three units of the resource while Rob receive five units. S_2, on the other hand, recommends four for both agents. The two bargaining solutions provide conflicting recommendations. Yet note that while the bargaining solutions offer distinct recommendations as to how to divide goods between Bob and Rob, there are nonetheless apparent points of agreement. For one, on both solutions Bob receives *no more than four units of the resource*. Additionally, both solutions ensure Rob should receive no more than five units of the contested resource. If, as previously suggested, individuals take their solution to in part reflect what they take a fair division to be, neither will

[13] While van Damme first introduced the notion of metabargaining in this 1986 paper, it should be noted that van Dame's primary interest was not the dynamics of divisional norms.

[14] A reader may at this point wonder whether there are connections between the so-called Nash program and metabargaining. There is a sense in which these two approaches are very similar. The Nash program is occupied with outlining procedures that underly the Nash bargaining solution. Metabargaining can be thought of as exploring which metabargaining procedures result in the Nash bargaining solution. However, there is no strong tie between the Nash program and this paper, as our interest is primarily in the dynamics of divisional norms given some underlying metabargaining procedure.

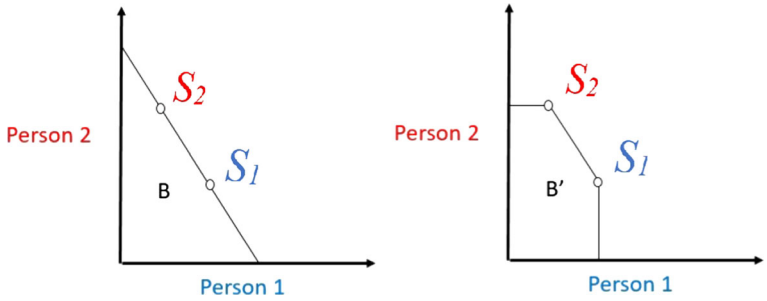

Fig. 3 Illustration of the metabargaining process. Bargaining solution of agent one (S_1) does not coincide with the recommendation of agent 2's bargaining solution (S_2). As a result the feasible set is truncated to yield B'

think it appropriate for someone to receive more than the maximum allowance specified by either bargaining solution. In other words, although agents may not agree on an exact solution, they will agree certain divisions are patently unfair.

We alter the feasible set of the bargaining problem to reflect this. In particular, we modify the feasible set so it no longer contains any outcomes in which Bob receives more than 4 units of the good or Rob receives more than five units of the good.[15] Once this alteration to the bargaining problem is made, the agents then reapply their bargaining solutions. If in this new bargaining problem the solutions coincide, then the game terminates and the agents receive the level of resources dictated by the bargaining solution. If the solutions diverge, then this two-step process of modifying the feasible set and applying the bargaining solution is repeated. In a nutshell, the above procedure captures the fact that both agents are interested in solving the problem of division and do so in a way which appeals to points of agreement between them.

Put more generally, when agents employ bargaining solutions that yield conflicting recommendations, the agents modify the underlying bargaining problem in the following fashion. Let person 1's bargaining solution make the following recommendation (π_1^1, π_1^2), where π_1^1 and π_1^2 specify person 1 and person 2's payoff, respectively. Likewise, the divisional norm of person 2 makes the recommendation (π_2^1, π_2^2). We modify the feasible set by excluding any outcome which bestows upon person 1 a payoff above $max(\pi_1^1, \pi_2^1)$ or person two a payoff greater than $max(\pi_1^2, \pi_2^2)$. Both parties then continue to bargain with this more restrictive feasible set (see Fig. 3).

This metabargaining procedure has several attractive properties. For one, it ensures that any changes made to the underlying bargaining problem are in some sense *endorsed by* both agents. Both bargainers would find it unfair if either one of them received more than the maximum they would enjoy under either bargaining solution, and for this reason, such divisions are no longer taken to be real possibilities. Additionally, the metabargaining procedure allows agents to avoid the disagreement point

[15] We can also modify the feasible set to reflect the fact that neither solution recommends that Bob receive *less* than 3 units and Rob *less* than four units. This is possible but complicates the dynamic analysis conducted in Sect. 4.1. For this reason we stick with this simpler metabargaining procedure but note that this alternate metabargaining procedure is well worth investigating and may yield different qualitative results than those uncovered in this paper.

in a variety of cases (as seen in the next section). Yet this is not to say that the metabargaining procedure will always allow bargainers to resolve their differences. Consider for a moment the interaction between agents utilizing two different bargaining solutions, both satisfying IIA. In this case, if there is initial disagreement between the two bargaining solutions the metabargaining procedure will not prevent a breakdown in bargaining. Truncating the feasible set in the way outlined above will not change the recommendation of a bargaining solution satisfying IIA, allowing for long-standing and unresolvable disagreement between the agents. Recall that a solution satisfies IIA if contracting the feasible set in a way which retains the original recommendation made by the bargaining solution ensures that the bargaining solution, when applied to the new feasible set, makes the same recommendation. Thus, initial disagreement between two solutions satisfying IIA cannot be resolved since neither bargaining solution is willing to 'budge.' In this case, the agents simply reach the disagreement point.

4.1 The stability of the utilitarian bargaining solution

We can now explore the dynamics of the divisional norms discussed in Sect. 2. Whenever two individuals using different divisional norms disagree, they resolve their dispute by appealing to the metabargaining procedure outlined above. In what follows, we discuss our central result: that the utilitarian bargaining solution has certain stability properties alternative norms of division lack.

Recall the notion of a bargaining bundle: a set of bargaining problems and a probability distribution determining how likely the various bargaining problems are. How a particular strategy fares depends in part on what strategy it is paired with as well as the underlying bargaining bundle. In this way, bargaining bundles can be thought of as specifying the game played by the bargainers. With this in hand we can now state the main result of this section:

(i) for all bargaining bundles, the utilitarian bargaining solution will always be at least neutrally stable, and
(ii) the utilitarian bargaining solution (of the four solutions considered here) is the only solution which is at least neutrally stable for all bargaining bundles.

One can think of (i) as an existence claim—there is some bargaining solution (namely, the utilitarian bargaining solution) with the property of always being an NSS for all bargaining bundles—while (ii) is best conceived of as a (partial) uniqueness claim—there is no other bargaining solution (among the four popular solutions considered in this paper) with this stability property. These two claims together seem to provide the utilitarian bargaining solution with a privileged status. The utilitarian bargaining solution is stable across a wide range of cases, and uniquely so. For the remainder of this section we provide brief informal proofs of claim (i) and (ii).

First, a few clarifications. Let B be an arbitrary bargaining bundle (a set of distinct bargaining problems accompanied by a probability distribution determining the likelihood of each bargaining problem). We assume two things about bargaining problems. First is that the feasible set just contains outcomes where both agents do as well or better than they would do compared to the disagreement point. Second, we assume the

slope of the frontier of the feasible set is *strictly negative*.[16] This latter claim entails all points on the frontier are pareto-optimal.[17] In other words, once we are at the frontier, it is not possible to make someone better off without making their partner worse off. This captures the fact that in bargaining contexts, the gains of one individual are often at the expense of her counterpart.

Furthermore, an agent using strategy s_i receives a payoff of $\pi(s_i, s_j)$ when paired against an agent using strategy s_j. Importantly, we assume that in each bargaining problem the two agents are equally likely to be 'Person 1' or 'Person 2' (see Fig. 1). Thus $\pi(s_i, s_j)$ is an *expectation*, taking into account both the strategies employed by the agents as well as uncertainty regarding which bargainer the focal agent is.[18] Recall from the previous section that our metabargaining procedure truncates the feasible set by ruling out divisions providing either agent more than the maximum resource level they would receive under the two bargaining solutions. As mentioned in Sect. 4, bargaining solutions satisfying IIA are 'obstinate': they select the same division in both the original as well as truncated bargaining problem. We now consider a community of agents that utilize the utilitarain solution when faced with an arbitrary bargaining bundle. It is straightforward to show that the utilitarian bargaining solution is always a neutrally stable strategy.

First, consider an invading solution that does satisfy IIA. This obstinate invader does just as well as the natives if their bargaining solution makes the same recommendation as the utilitarian bargaining solution for all divisional problems in the bargaining bundle. In this case, the invaders do as well, but no better, than the native utilitarians and thus the natives are neutrally stable with respect to the invaders. If the invader's solution does not coincide with the utilitarian solution, then the disagreement point is reached (since both invader and native are obstinate), meaning the utilitarian natives can resist invasion.

Now we consider solutions that do not satisfy IIA. This is slightly more complicated since some bargaining solutions not satisfying IIA, such as the K–S solution discussed in Sect. 2, are *concessive*. Concessive bargaining solutions are those which always (via the metabargaining procedure) converge on the division recommended by their obstinate counterpart. As Fig. 4 illustrates, when paired with a utilitarian, the K–S bargainer will eventually converge on the utilitarian bargaining solution via the metabargaining procedure.[19] Thus a K–S invader paired with a native will do as well

[16] We further assume, as is commonly done in the bargaining literature, that the feasible set is compact and convex.

[17] This also excludes the case in which a portion of the frontier is a horizontal or vertical line since the slope would be either zero or undefined and thus ruled out by our restriction that the slope always be strictly negative.

[18] In other words, we presuppose that there is no correlation between the bargaining solution an agent decides to utilize and the agent's bargaining position. We grant that this is not always a reasonable assumption to make. For instance, it may be the case that those who advocate for utilitarian arrangements tend to have quite a bargaining advantage.

[19] For the K–S solution to converge on the recommendation of its obstinate counterpart, the bargaining problem must be such that the recommendations made by both bargaining solutions are on the pareto frontier and the ideal point is not a member of the feasible set. If these conditions fail to hold, then the K–S solution will not converge and the disagreement point will be reached.

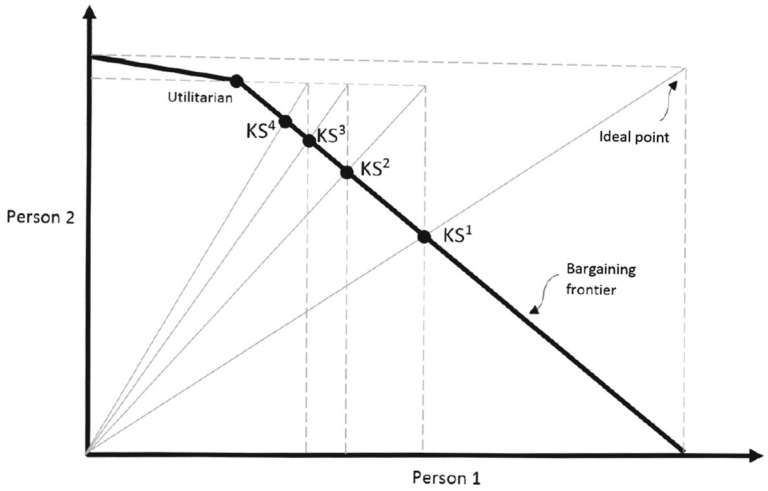

Fig. 4 Illustration of the metabargaining process involving a utilitarian and K–S bargainer. KS_1 refers to the K–S solution of the original bargaining problem. The feasible set is then modified as specified by the metabargaining procedure and KS_2 denotes the K–S solution of the resulting bargaining problem

as natives do against themselves (since the utilitarian bargaining solution is reached in both cases). However, agents utilizing the K–S solution may do worse than utilitarians when paired against fellow invaders. In this case, they receive the K–S solution while natives garner the utilitarian division against concessive K–S invaders. Since the average payoff associated with the K–S solution cannot exceed that of the utilitarian solution, this means K–S invaders can at best do as well as native utilitarians.

What of those bargaining solutions that do not satisfy IIA but are also not concessive? In this case there is no guarantee the non-concessive bargainer will converge on the division advocated by their steadfast counterpart via the metabargaining procedure. If the non-concessive solution does not eventually converge on the division recommended by the obstinate bargaining solution then the disagreement point is reached and the invading solution will be weeded out of the community. If the non-concessive solution by some fortuitous accident converges on the recommendation of the native, then just like the concessive IIA bargainer, they can at best do as well as the natives. Taking all of these cases together, it is clear that the utilitarian bargaining solution will always be neutrally stable, and in many cases, will be evolutionarily stable.

Now consider the 'uniqueness' claim, which states that none of the other bargaining solutions considered in Sect. 2 will be neutrally stable for all bargaining bundles. It is easy to see why this is the case for the K–S solution. A population of agents utilizing this concessive bargaining solution can easily be invaded by a utilitarian. Utilitarians when paired with a native receive their part of the utilitarian solution. If the division recommended by the native K–S bargaining solution coincides with that of the utilitarian solution, then the natives are neutrally stable with respect to the invaders. If there is disagreement between the two solutions, the utilitarian solution is attained and thus the invaders do better (on average) against natives than natives do

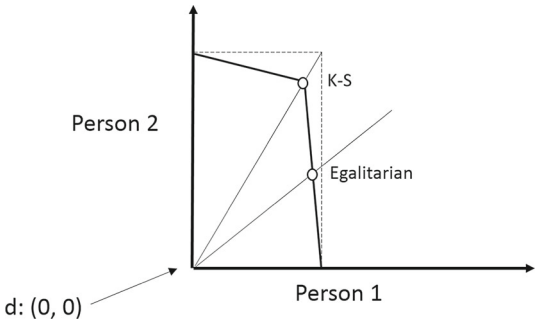

Fig. 5 K–S and egalitarian solutions

against their own. Ergo, the K–S solution is neither neutrally nor evolutionarily stable for all bargaining bundles.

We now consider the remaining egalitarian and Nash solutions. Both are susceptible to invasion from the K–S bargaining solution. Consider a population of egalitarians faced with a bargaining bundle that contains only one bargaining problem, featured in Fig. 5. Since the egalitarian solution satisfies IIA, a K–S invader, when paired with a native, will simply get the egalitarian bargaining solution. Thus invaders do no better than natives against natives. How well do invaders and natives fare when paired with an invader? As it turns out, invaders, when paired with themselves, outperform natives forced to interact with invaders. This can be read off of Fig. 5. When a K–S invader interacts with another K–S invader, the outcome provides an extraordinarily high average payoff. Compare this to the egalitarian solution (which is attained when a native pairs with an invader), which provides a rather mediocre average payoff. Thus K–S invaders can gain a foothold in the egalitarian community, driving the natives to extinction.

It can similarly be shown that K–S can successfully invade a population of Nash bargainers. Thus we've established that the Egalitarian, Nash and K–S bargaining solutions all fail to be neutrally stable for all bargaining bundles. There are scenarios in which these solutions can easily be undercut by the introduction of an alternative bargaining solution. This is not a shortcoming of the utilitarian bargaining solution, as the utilitarian solution is at least neutrally stable for all bargaining bundles.

Together these results reveal a very interesting dynamic which highlights the importance of the utilitarian bargaining solution. Note that not only is the utilitarain bargaining solution robust in the sense that it is resistant to invasion, the stability properties of the remaining bargaining solutions are such that it is possible a community not initially at the utilitarian bargaining solution will gravitate towards it. Consider, for instance, a population of egalitarians (a similar story applies for the Nash bargaining solution). In this case, the utilitarian bargaining solution cannot invade, as the utilitarian will receive the disagreement point whenever paired with a native egalitarian. Yet this does not mean the egalitarian community is stable—K–S invaders can thrive for a variety of different bargaining bundles. This in turn opens the door to utilitarians, who gain a foothold in the community once it is populated by those using the K–S bargaining solution. A path exists from egalitarian to utilitarian bargaining solutions.

5 Gauthier and the stability of the just person

We now consider an alternative to the metabargaining approach discussed in the previous section that can similarly allow us to study the dynamics of these more sophisticated bargaining norms. In particular, we draw inspiration from David Gauthier's landmark work on bargaining and the social contract (Gauthier 1986). In *Morals by Agreement*, Gauthier advocates for the so-called minimax relative concessions bargaining solution as the rational way of dividing the benefits and burdens of social cooperation.[20] Yet Gauthier is also moved by concerns of stability and contends that in addition to being the rational way to divide resources, the minimax relative concessions solution, when widely adopted by members of a community, is also stable. In what follows we question this latter claim.

Gauthier characterizes the *just person* as one who endeavors to act on the basis of the 'principle of minimax relative concession with those of her fellows whom she believes to be similarly disposed' (157). It is only when the agent has reason to believe her counterpart will not adhere to Gauthier's preferred divisional norm is she allowed to deviate from the minimax relative concessions recommendation and instead act so as to 'maximize her own utility.' In this sense Gauthier's just agents are *conditionally* just—they abide by the minimax bargaining solution if and only if others are likely to follow suit.[21] This caveat is important for it in part immunizes Gauthier's just agent from exploitation at the hands of the egoist, who, unlike the just agent, is keen to exploit whenever it is to her advantage.[22] The just agent will not fall prey to the egoist—she won't naively cooperate with the egoist but instead protect herself by acting so as to maximize her payoff—and much of Gauthier's *Morals by Agreement* is an attempt to convince the reader that the pesky egoist will not unsettle a community of conditionally just agents following the minimax bargaining solution. In the course of addressing the egoist Gauthier develops one of the more controversial aspects of his book, the heterodox theory of choice known as 'constrained maximization.'

Yet the egoist is not the only, or even most pressing, threat to Gauthier's community of conditional minimax relative concession bargainers. Consider, for instance, an agent who cares about fairness and justice (in the sense that she is not looking to exploit her peers), but disagrees with Gauthier on the specifics. Instead of endorsing the minimax relative concessions solution, she favours the utilitarian bargaining solution and aims

[20] Recall that the minimax relative concessions solution is identical (in the two-person case) to the K–S solution of Sect. 2.

[21] Gauthier contends a just individual will adhere to the following condition, namely, that 'each person will select a fair optimizing response to the choice he expects the others to make, provided such a response is available to him; otherwise, his choice must be a utility maximizing response. A just person is disposed to interact with others on the basis of [this condition]' (157). In other words, a just agent will do their part to bring about desirable and fair outcomes if they have assurance others are doing their part to promote justice. When the agent in question does not have reason to believe others are similarly compelled, they will instead act on the basis of self-interest.

[22] For instance, Gauthier's just agent is allowed to break contracts and simply behave in a fashion that maximizes her self-interest if she has no reason to believe her counterpart will abide by the terms of an agreement. Thus Gauthier's just agents will not be preyed upon by the egoist in situations resembling Hume's farmer's dilemma (which is essentially a sequential prisoner's dilemma).

to bring about this distribution regardless of the behavior of her counterpart. In other words, when confronted by a divisional problem this individual demands the amount of the resource consistent with the utilitarian bargaining solution, whether it is to her advantage or not. What occurs when Gauthier's just agent meets this fervent utilitarian? Gauthier's bargainer, realizing her counterpart is not similarly disposed to endorse the minimax relative concessions solution, will opt to simply 'maximize expected utility,' which, in this context, will entail her endorsing the utilitarian bargaining solution favoured by her counterpart (so as to avoid the disagreement point).[23] As a result, invading utilitarians will receive a higher payoff against minimax bargainers than minimax bargainers do against themselves as the utilitarian solution on average provides a higher payoff than the minimax relative concessions solution. Furthermore, Gauthier's community is susceptible to invasion by other agents utilizing alternative unconditional bargaining strategies. Unconditional Nash bargainers, for instance, can penetrate the population if for a given bargaining bundle the Nash bargaining solution offers a higher average payoff than the minimax relative concessions bargaining solution. Ergo, Gauthier's community of minimax relative concession bargainers is not stable. Further, note that a community of unconditional utilitarian agents is impervious to the introduction of a few conditional minimax bargainers.[24] Utilitarians can invade Gauthier's community, but Gauthier's just agents cannot upset a utilitarian population.[25]

Of course one obvious response to this is that the deck has been unfairly stacked against Gauthier. While Gauthier requires his agents condition their behavior on those of their counterpart, we have allowed the utilitarian to stick to her guns. She unconditionally advocates for utilitarian arrangements, and for this reason is able to get her way when faced with the more concessive just agent considered by Gauthier. Does the story change if we instead consider a *conditional* utilitarian who, like Gauthier's just agent, maximizes expected utility if her fellow is not 'similarly disposed' to opt for the utilitarian bargaining solution?

This is a difficult question to cleanly answer, as what it means for an agent to 'maximize' in this context hinges on her beliefs regarding the behavior of her counterpart. We gesture at a solution in what follows. Consider a small minority of conditional utilitarians in a community of conditional minimax bargainers. Over time, both types of bargainers update their beliefs on the basis of past experience regarding how individuals employing the alternate bargaining solution will behave with them. Agents

[23] Gauthier also says at certain points that his minimax just agent will cooperate with those who approximate the minimax relative concessions bargaining solution. This seems to indicate that the minimax just agent may often agree to utilitarian arrangements not out of self-interest, but instead because the utilitarian bargaining solution often reasonably approximates the minimax relative concessions solution.

[24] Conditional minimax bargainers will do as well as utilitarians when placed to interact with utilitarian bargainers. Yet when paired with minimax bargainers, utilitarians outperform the invaders. Thus utilitarians can resist invasion from Gauthier's just agents.

[25] This observation is similar to one made by Binmore (1990). Binmore contends that conditional utilitarians (those who settle on the utilitarian bargaining solution with in-group members but accommodate out-group members by opting for their counterpart's favorite bargaining solution) will be able to invade a population of unconditional bargainers (bargainers who all unconditionally utilize the same bargaining solution). Our observation complements this, as we contend that *unconditional* utilitarians will be able to invade a community inhabited by *conditional* bargainers.

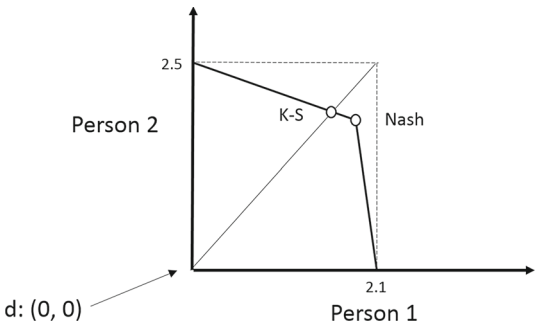

Fig. 6 Nash and K–S (minimax relative concessions) solutions. Here the Nash solution guarantees 2 units of resource for both agents. The K–S solution, on the other hand, yields an average payoff of 1.91

then best respond to their counterpart given their beliefs regarding their counterpart's likely behavior.[26] A similar process has been formally studied in the context of bargaining games in a famous paper by Young (1993).[27] Young finds that under similar conditions play moves to the Nash bargaining solution. This provides us with some reason to think divisions between utilitarian and minimax bargainers will likewise settle on the Nash solution. If so, then for many bargaining bundles, conditional utilitarians will be able to successfully infiltrate a population of minimax bargainers. When paired with natives, utilitarian invaders will receive the Nash bargaining solution, which, as illustrated in Fig. 6, often results in a higher average payoff than the minimax relative concessions solution. In this case, invaders do better against natives than natives do against themselves, meaning conditional minimax bargainers are not evolutionarily stable.[28] Note further that a population of conditional utilitarians will be impervious to invasion due to the fact that the average payoff associated with the utilitarian bargaining solution is at least as high as the average payoff associated with the Nash bargaining solution. To be precise, the conditional utilitarian strategy is (once again) an NSS for all bargaining bundles. Invading strategies can at best do as well as utilitarian incumbents, and for many bargaining bundles the conditional utilitarian strategy will be an ESS.

Gauthier was right to be concerned about the stability of his social contract, but he was overly focused on the egoist. As we have illustrated, diversity in the form of alternative bargaining solutions can unseat the minimax relative concessions solution.

[26] In other words, when utilitarians meet fellow utilitarians they settle on the utilitarian division, but when they are tasked to interact with minimax bargainers, they act on the basis of their beliefs regarding the likely behavior of their counterpart (and vice versa).

[27] In particular, Young studies a process whereby individuals interact and update their beliefs regarding the likely behavior of others based on a shared history of past play. He finds that this process results in agents converging on a generalization of the Nash bargaining solution. This result is somewhat robust and still obtains even when agents are more sophisticated best responders (Saez-Marti and Weibull 1999)

[28] Note that this means conditional utilitarians are not the only strategy that can destabilize the population. Any conditional bargaining strategy could potentially infiltrate the population so long as the average payoff associated with the Nash bargaining solution is greater than the average payoff associated with the minimax relative concessions solution.

6 Conclusion

Overall, this paper makes three substantive contributions. First, we provide two means of tackling a longstanding and somewhat unnoticed problem in the literature on the evolution of the social contract.[29] As mentioned, prior work tends to investigate the simplest of bargaining games. While illuminating, previous explorations are somewhat limited in scope and applicability. We aim instead to develop a means of investigating the dynamics of sophisticated, all-purpose divisional norms that can be applied in a variety of strategic settings.

Bargaining solutions from economics and philosophy provide guidance across a wide range of strategic scenarios and are thus good candidates to consider when attempting to better understand the dynamics of divisional norms. Yet disagreement between solutions poses a difficult problem and threatens to render our standard stability concepts useless. Furthermore, resolving disagreement by appealing to an intuitive principle such as 'split the difference' seems to presuppose what we aim to provide an account of: the emergence of divisional norms. To handle cases of disagreement, we instead consider a procedure where bargainers amend the set of possible outcomes and then reapply their favoured bargaining norm. In particular, the feasible set is modified in a fashion which reflects points of agreement between conflicting divisional norms. We do not contend this procedure is obviously the most compelling or natural way to proceed, as there may be other attractive procedures which avoid the disagreement point.[30] Instead, what we have provided is a proof of possibility, demonstrating one way in which the dynamic and static traditions in social contract can be fruitfully brought together, as well as an invitation to others to address these issues in the study of the evolution of the social contract.

Second, we find that the metabargaining process of Sect. 4 as well as the conditional bargaining strategies considered in Sect. 5 both indicate the utilitarian bargaining solution has a special status. The utilitarian solution is able to resist invasion across a wide-range of bargaining problems. In particular, the utilitarian bargaining solution is a neutrally stable strategy for *all* bargaining bundles. This cannot be said for any of the other bargaining solutions considered in this paper. This suggests the utilitarian bargaining solution may be generally favored in those cases involving procedures or mechanisms that allow those subscribing to different bargaining norms to nonetheless avoid the disagreement point.[31] Yet these findings should be taken with some cau-

[29] It is worth noting that this problem has not gone *completely* unnoticed. See, for instance, Binmore (2005). However, Binmore's analysis differs from ours substantially since he both invokes cultural evolutionary theory and the veil of ignorance to argue for certain principles of justice. Our contribution is purely naturalistic as it doesn't appeal to the device of the veil.

[30] See, for instance, Trockel (2002) and Naeve-Steinweg (2002, 2004) who explore alternative metabargaining procedures.

[31] We can see why this may be so. When some mechanism or procedure exists that allow those with differing norms to avoid the disagreement point, the utilitarian solution should always be an NSS since utilitarians still always do at least as well (if not better) against fellow utilitarians than invaders do against utilitarians. Yet it is clear that other bargaining solutions may not be an NSS as they could be invaded by the utilitarian. This, of course, crucially depends on the details of the mechanism or procedure in question.

tion since our models are rather exploratory in nature and further work is needed to determine the true significance of the utilitarian bargaining solution.

Finally, we have suggested that certain social contract theorists have erred by overlooking the importance of moral diversity. Gauthier's minimax relative concessions solution, for instance, can be undercut by the introduction of individuals beholden to different divisional norms. While the purely self-interested certainly do threaten to destabilize the status quo, our results compel those in the social contract tradition to expand their focus beyond the egoist.[32]

References

Alexander, J. M. (2007). *The structural evolution of morality*. Cambridge: Cambridge University Press.

Alexander, J. M., & Skyrms, B. (1999). Bargaining with neighbors: Is justice contagious? *Journal of Philosophy*, *96*, 588–598.

Barrett, J. (2014). The evolution, appropriation, and composition of rules. *Synthese, 195*(2), 623–636.

Bednar, J., & Page, S. (2007). Can game(s) theory explain culture? The emergence of cultural behavior within multiple games. *Rationality and Society*, *19*, 65–97.

Bicchieri, C. (2006). *The grammar of society: the nature and dynamics of social norms*. Cambridge: Cambridge University Press.

Binmore, K. (1990). Evolution and utilitarianism: Social contract III. *Constitutional Political Economy*, *1*(2), 1–26.

Binmore, K. (2005). *Natural justice*. Oxford: Oxford University Press.

Braithwaite, R. B. (1955). *Theory of games as a tool for the moral philosopher*. Cambridge: Cambridge University Press.

Bruner, J. (2015). Diversity, tolerance and the social contract. *Politics, Philosophy and Economics*, *14*(4), 429–448.

Bruner, J. (2017). Minority (dis)advantage in population games. *Synthese*,. https://doi.org/10.1007/s11229-017-1487-8.

Bruner, J., & O'Connor, C. (2017). Power, bargaining and collaboration. In T. Boyer-Kassem, C. Mayo-Wilson, & M. Weisberg (Eds.), *Scientific collaboration and collective knowledge*. Oxford: Oxford University Press.

Gaus, G. (2010). *The order of public reason: A theory of freedom and morality in a diverse and bounded world*. Cambridge: Cambridge University Press.

Gauthier, D. (1986). *Morals by agreement*. Oxford: Oxford University Press.

Huttegger, S., & Smead, R. (2011). Efficient social contracts and group selection. *Biology and Philosophy*, *26*, 517–531.

Huttegger, S., & Zollman, K. (2013). Methodology in biological game theory. *The British Journal for the Philosophy of Science*, *64*, 637–658.

Moehler, M. (2018). *Minimal morality: A multi-level social contract theory*. Oxford: Oxford University Press.

Naeve-Steinweg, E. (2002). Mechanisms supporting the Kalai–Smorodinsky solution. *Mathematical Social Sciences*, *44*(1), 25–36.

Naeve-Steinweg, E. (2004). The averaging mechanism. *Games and Economic Behavior*, *46*(2), 410–424.

Nash, J. (1950). The bargaining problem. *Econometrica*, *18*, 155–162.

Saez-Marti, M., & Weibull, J. (1999). Clever agents in Young's evolutionary bargaining model. *Journal of Economic Theory*, *86*(2), 268–279.

Skyrms, B. (1996). *The evolution of the social contract*. Cambridge: Cambridge University Press.

[32] I thank Brian Skyrms, John Thrasher, Simon Huttegger, Keith Dowding, Ryan Muldoon, Jeff Barrett, Rudolf Schuessler, Hannah Rubin, Aydin Mohsen, and the participants of the MSPT Seminar as well as MSPT Works In Progress Group at ANU, the Evolutionary Approaches to Moral Behavior Workshop at Monash University, the Social Dynamics Research Seminar at UC Irvine, the 9th Decisions, Games and Logic Conference at the University of Michigan, Ann Arbor and the PPE Society Meeting in New Orleans. Special thanks goes to RJ Leland who helped me carefully think through many of these issues.

Skyrms, B. (2013). Natural social contracts. *Biological Theory*, *8*, 179–184.
Sugden, R. (1986). *The economics of rights, cooperation and welfare*. London: Palgrave Press.
Thrasher, J. (2014). Uniqueness and symmetry in bargaining theories of justice. *Philosophical Studies*, *167*, 683–699.
Trockel, W. (2002). A universal meta bargaining implementation of the Nash solution. *Social Choice and Welfare*, *19*(3), 581–586.
van Damme, E. (1986). The Nash bargaining solution is optimal. *Journal of Economic Theory*, *38*, 78–100.
Vanderschraaf, P. (forthcoming). *Strategic justice*. Oxford University Press.
Vanderschraaf, P. (2016). In a weakly dominated strategy is strength: Evolution of optimality in stag hunt augmented with a punishment option. *Philosophy of Science*, *83*(1), 29–59.
Wagner, E. (2012). Evolving to divide the fruits of cooperation. *Philosophy of Science*, *79*(1), 81–94.
Young, P. (1993). An evolutionary model of bargaining. *Journal of Economic Theory*, *59*(1), 145–168.
Zollman, K. (2009). Explaining fairness in complex environments. *Politics, Philosophy and Economics*, *7*(1), 81–97.