Reply to Critics

Abstract: I reply to commentaries by Justin Bruner, Robert Sugden and Gerald Gaus. My response to Bruner focuses on conventions of bargaining problems and arguments for characterizing the just conventions of these problems as monotone path solutions. My response to Sugden focuses on how the laws of humanity present in Hume's discussion of vulnerable individuals might be incorporated into my own proposed account of justice as mutual advantage. My response to Gaus focuses on whether or not my account of justice as mutual advantage can incorporate deep differences in values across subgroups of a larger society.

Keywords: justice as mutual advantage, convention, monotone path solution, vulnerability objection, laws of humanity, value domain

§1. Introduction

I am delighted to have this opportunity to respond to this panel of critics, who here present many penetrating questions and arguments and from whose work I have learned so much during my career. Below I will respond in turn to what I regard the most significant issues Bruner, Sugden and Gaus each raise, and conclude with a short discussion of some general unsettled issues.

§2. Bargaining Conventions

The Nash bargaining problem is an invaluable tool for modeling a social contract because its nonagreement point corresponds to a baseline state and its feasible set corresponds to alternative social contracts. For example, the nonagreement point maps to a State of Nature and the feasible set maps to alternative social regimes of Hobbes' and Gauthier's theories. Points of the feasible set that are Pareto superior to the nonagreement point correspond to social contracts where parties receive shares of what Gauthier terms a *cooperative surplus* they produce by

following the terms of these contracts that they cannot achieve at the State of Nature (1986 p. 130, p. 141). I identify a just social contract with a satisfactory solution of a bargaining problem.

But what counts as a satisfactory solution? Seven decades after John Nash's foundational work, attempts to answer this question sustain a lively research program. Nash set much of the agenda of this program himself by proposing both strategic and axiomatic analyses of the bargaining problem he argued should be complementary (1950, 1953). Thomas Schelling (1960) inaugurated another approach that identifies solutions in terms of learning and focal points. This approach draws upon both the results of bargaining experiments and concepts from evolutionary game theory. My own discussion of the bargaining problem stresses the learning-focal point approach and to a lesser extent the axiomatic approach. I think the strategic approach, and a good part of the axiomatic approach, effectively ask for too much by aspiring to identity a uniquely correct solution for every Nash bargaining problem. Real individuals confronted with real bargaining problems may consider many distinct outcomes as possible solutions. In some of the simplest bargaining problems, where claimants vie for shares of some divisible good like a cake, each outcome where all the good is distributed and each claimant gets some positive share is a strict equilibrium. Each of these strict equilibria characterizes what I call a *basic convention* that is incumbent when the claimants have common knowledge that they follow this equilibrium rather than any of the others (2019 pp. 81-82). Even a cake division problem with only two claimants has infinitely many different bargaining conventions. I set myself to answer the question: Which subset of the full set of available bargaining conventions can characterize just social contracts?

In *Strategic Justice* I propose this answer: The conventions that satisfy a *baseline consistency* condition such that if the feasible set expands, all fare at least as well if the new

solution point is determined using the old solution point as nonagreement point as they fare if the new solution is determined from the original nonagreement point (2019 p. 312). Although Bruner reads me as defending the egalitarian solution specifically, the egalitarian solution is typically one of many monotone path solutions. I argue for my proposed answer from two directions. From the axiomatic side, I argue for the monotone path solution social contracts on the grounds that these social contracts do not need to be renegotiated from scratch in the face of changing background conditions. The family of *monotone path solutions* of the bargaining problem are baseline consistent. These solutions are *decomposable* (Kalai 1977) and can be reached either straight from the nonagreement point or in stages from successive intermediate solution points. The Kalai-Smorodinsky solution and the Nash solution can both fail to be decomposable, and Bruner shows this by example for the Nash solution.[1]

In fact the monotone path solutions have other stability properties I discuss little or not at all in *Strategic Justice* that lend more support to the axiomatic case for monotone path solutions. Bruner discusses the effects first analyzed by Myerson (1981) of the timing of preferences over options before and after uncertainties are resolved. Myerson shows that for a quite general class of bargaining problems, only a weighted utilitarian and the egalitarian solutions are such that these sorts of timing effects never cause conflicts over when to apply the solution. Bruner illustrates Myerson's result by example, where each side would strictly prefer to apply the egalitarian solution ex ante, before payoff uncertainties are resolved, while they would disagree over whether to apply the Nash solution ex ante or ex post. I agree with Bruner that this timing effect property gives further support to the egalitarian solution in particular. Bruner also observes that I could have appealed more directly to a subgroup consistency property, such that all do at least as well if the solution is reapplied across subgroups.[2] The monotone path solutions, but not

the Kalai-Smorodinsky solution, satisfy such subgroup consistency. The monotone path solutions, but not the Nash solution, are also *population monotone*, that is, no parties in a given set lose while others gain merely because new claimants arrive or original claimants depart.[3]

Bruner rightly challenges the other side of my overall argument from learning and focal points. In *Strategic Justice* I explore the idea that a community may come to regard certain solutions of the bargaining problem as focal on account of inductive learning. I apply a very simple model of inductive learning, *weighted fictititous play*, to several well-known bargaining problems. In these problems, inductive learners can reach many different bargaining conventions, as one would expect. And in each of these problems, the distribution of the learned conventions in these examples is centered at or very nearly at the egalitarian solution. I conclude that the egalitarian solution may have some general attracting power for inductive learners, and that this in turn could lead communities to regard the egalitarian solution as a focal convention. This conclusion is admittedly tentative, since I analyze a small number of examples. Bruner applies a different learning rule, the replicator dynamic, to a much larger set of bargaining problems. In his more systematic study, Bruner finds that conventions at or near the Nash solution tend to emerge most frequently. Bruner's findings look like powerful evidence against the learning-focal point side of my argument.

Bruner is certainly right to call for more study of learning applied to bargaining games, particularly bargaining games with asymmetric payoff structure. There is much open research territory regarding asymmetric bargaining games. This is partly because the best-known models of learning in games are only easy to apply to games with symmetric payoff structure or where each agent has but a handful of pure strategies, and asymmetric bargaining games satisfy neither of these conditions. Still, I think one can explain the discrepancy between Bruner's and my own

findings. All of the feasible sets of Bruner's asymmetric bargaining games are *comprehensive*, so that if any point $\boldsymbol{y}$ that weakly Pareto dominates the nonagreement point is itself weakly dominated by a point $\boldsymbol{x}$ in the feasible set, then $\boldsymbol{y}$ is also in the feasible set.[4] Most of the bargaining problems I analyze do not have comprehensive feasible sets. To illustrate, Figure 1 depicts the feasible set of the Braithwaite bargaining problem.

[Figure 1. Braithwaite Problem Feasible Set]

This feasible set is the set of convex combinations of the payoff vectors of the celebrated conflictual coordination game Braithwaite presented in his 1954 Cambridge lecture (1994). The Figure 2 payoff matrix summarizes Braithwaite's game.

[Figure 2. Braithwaite Game]

The Figure 1 feasible set is not comprehensive. In particular, none of the points where Matthew receives his best payoff of the (M,G) equilibrium but Luke receives less that what he receives at (M,G) are in this feasible set. Similarly none of the points where Luke receives his (G,M) equilibrium payoff and Matthew receives less than his (G,M) payoff are in this feasible set. Figure 3 depicts an extension of the Figure 1 feasible set that is now comprehensive.

[Figure 3. Expanded Braithwaite Feasible Set]

All of the aforementioned points where either Luke or Matthew receives his ideal payoff are in the Figure 3 comprehensive set. In a different recent study Bruner (2020) applied the replicator dynamic to a series of noncomprehensive bargaining sets generated from conflictual coordination games and found that the orbits of this dynamic tend to converge to points at or near the Kalai-Smorodinsky solution. Bruner's findings led me to try a similar comparison of

both versions of the Braithwaite Problem. I applied weighted fictitious play to a bargaining game that generates the Figure 3 comprehensive set, and the resulting distribution of emerging bargaining conventions was concentrated near the Nash solution, which is consistent with the results of Bruner's replicator dynamic studies. The same learning dynamic applied to the original Braithwaite bargaining game produces a distribution of bargaining conventions concentrated near the egalitarian solution (2019 p. 174). I think this difference is not so surprising given that in the original Braithwaite game, the "endpoint" equilibria (G, M) and (M, G) where one claimant gets all of the good at stake and the other none are both strict, so that each can "pull" the learning process somewhat away from the other "endpoint" and towards the "middle". In the game that generates the comprehensive Figure 3 set, neither of the "endpoint" equilibria are strict, and there are whole sets of strategy profiles where Luke gains his highest payoff and where Matthew gains his highest payoff. In this game Matthew's optimal outcome set is larger than Luke's, so it is no surprise that inductive learning drifts towards the more lopsided Nash solution.

The difference in these two cases may have an interesting substantive interpretation. In the original Braithwaite bargaining game, each agent has a strict preference for the strict equilibrium least favorable to himself over the nonagreement point. In the bargaining game of the expanded comprehensive set, neither agent has any such strict preference. The available social contracts reflected by comprehensive sets may reflect situations where parties care nothing for the prospects of their partners and each can achieve a most preferred outcome that leaves the others no better off than they are at the State of Nature. The noncomprehensive set social contracts may reflect situations where all want to achieve a social contract of true mutual advantage where all fare better than they would fare at the State of Nature.

§3.Humanity, Justice and the Vulnerable

Examples like Thucydides' Melian dialog and Hume's race of weaker rational creatures elicit a strong general response from many philosophers I summarize here: Far from being cases where justice is irrelevant, situations where differing parties have great differences in relative power are cases where justice is *most* relevant.[5] These philosophers echo the Melian representatives, who claim that a plea for justice should serve one in times of distress (*Peloponnesian War* V:90). Some argue justice as mutual advantage should be rejected because it withholds its benefits from some who are in distress, namely the vulnerable. As Gaus observes, this *Vulnerability Objection* is an important case of the *agent domain problem*, the worry that a purportedly just social system might define the community who benefit from this system so that that certain despised or relatively weak members of society are denied benefits. In *Strategic Justice* I argue that the Vulnerability Objection rests upon the assumption that justice as mutual advantage necessarily incorporates a *contribution requirement* of the form reflected in the aphorism attributed both to Saint Paul and to Vladimir Lenin: 'He who does not work, neither shall he eat'.[6] According to this requirement the vulnerable, who I define as members of society who simply cannot contribute to the cooperative surplus,[7] would be ineligible for any of the benefits of justice. I address the Vulnerability Objection by relaxing the contribution requirement so that it binds only those who can contribute. This reflects more nearly the actual text of Saint Paul's epistle: 'if anyone was unwilling to work, neither should that one eat.' I argue that given this more relaxed requirement a justice as mutual advantage system can extend its benefits to the vulnerable.

Sugden questions both my rationale for this proposed solution to the problem of the vulnerable and the need for this sort of solution. Sugden suggests that I can turn to Hume for an

alternate solution, one he regards as more natural than my own proposed solution. As Sugden observes, Hume expressly states that the laws of humanity would oblige humans to treat the weaker rational creatures gently. Hume follows this his fictitious example of the weaker creatures by observing, 'The great superiority of the civilized EUROPEANS above barbarous INDIANS, tempted us to imagine ourselves on the same footing with regard to them, and made us throw off all restraints of justice, and even of humanity, in our treatment of them (*Enquiry* 3.2:19).' Sugden takes Hume to think that the Europeans' actions were certainly inhumane. And he observes that Hume occasionally discusses other cases of definite moral obligation that do not appear to lie within the scope of justice, such as the duty a rich man has to make some provision for those in need (*Treatise* 3.2.1:14). Sugden argues that I can extend this idea to the vulnerable in general. Sugden's Hume-inspired response to the Vulnerability Objection is quite elegant: Let humanity do the work of providing benefits for the vulnerable, and leave justice for securing benefits for those who can contribute to the cooperative surplus.

Sugden's proposed solution, which I think he would say is simply Hume's solution, would appear to make the project of defending justice as mutual advantage substantially easier. Sugden sees me as casting an unnecessarily wide a net by trying to rig justice as mutual advantage to extend benefits to the vulnerable. Why should I not instead follow Hume and help myself to the laws of humanity in my discussion of the vulnerable? I certainly think Sugden's proposal merits further exploration. I also think Hume's hostile critics tend to overlook his discussions of humanity, and that more generally the relationship between justice and different other-regarding virtues such as benevolence and humanity is underexplored territory in contemporary philosophy.

Nevertheless, I hesitate to accept Sugden's proposal outright. I think I should make it clear that while I draw much inspiration from Hume, I am not trying to develop an overall argument that is faithful to Hume in all respects. Rather, I am trying to see how far viewing justice as a system of conventions goes in answering the questions Plato and his successors raise: What is justice? Why be just? Sugden views me as adopting a somewhat Platonic stance Hume would find alien. I think Hume very much wants to answer these questions, though he certainly does not think of justice as some sort of timeless, ideal system of the sort Plato envisioned. Hume appreciates that justice is an exceedingly complex phenomenon one that upon reflection we should find exceedingly surprising. I think I am in agreement with Sugden in regarding Hume's project as predominantly explanatory, and I would say the same about my own *Strategic Justice* project. Like Hume, I want to address the question of the vulnerable. Sugden cites my claim that a principle that the vulnerable are owed benefits of justice is part of what Sidgwick calls 'the Morality of Common Sense' or what Rawls calls a *considered judgment* of morality (1981 pp. 214-215, 1971 pp. 20). Sugden questions my confidence in this claim, or at least part of this claim. As I read Sugden, he allows that a belief that the vulnerable are owed some of society's benefits might be a considered judgment, but that I am too quick to suppose this judgment includes the belief that such benefits are benefits of justice. After all, Hume appeals only to laws of humanity when he discusses obligations to the weaker rational creatures. I think Sugden is proposing that one can view all of the duties with respect to the vulnerable as falling under the umbrella of humanity.

Again I hesitate to embrace Sugden's proposal, despite its considerable explanatory potential. For I doubt Hume's laws of humanity can secure anything resembling the rights of a system of justice. Sugden thinks that humanity can ground some definite general duties towards

the vulnerable, and gives as an example the requirement in Hume's Britain that every parish make provision for certain poor residents. Here Sugden and I simply disagree. Indeed, Sugden gives me grounds for doubting his position in his own example, for he notes that the parishes were legally required to make these provisions for poor residents. This suggests that these duties, which were indeed general and evidently specific, in fact fell in the scope of justice. I think this is no accident, since one fundamental way humanity differs from justice is the former concerns alleviating suffering while the latter concerns rights. Hume certainly appreciates this distinction. Regarding the weaker rational creatures, he says only that the laws of humanity require that humans 'give gentle *usage* to these creatures (*Enquiry* 3.2:18, my emphasis).' The laws of humanity might forbid gratuitously cruel treatment of these creatures, but Hume does not rule out their enslavement and makes no claim that these laws would guarantee that such creatures receive anything they might need for mere survival. Even killing such creatures might not violate the laws of humanity if one kills them painlessly.[8] Moreover, what humanity requires and what justice requires are not so plainly distinct, and where they may overlap justice is thought to have clear priority. For example, homicide and torture are generally regarded as both unjust and inhumane, and I think no one tries to ground prohibitions against homicide and torture only or even primarily on the basis of humanity. I think it is telling that the Melians make their plea to the Athenians on the basis of justice, without even mentioning duties of humanity. I also think it is telling that in charging some of the defendants of the Nuremberg Trials with crimes against humanity, no one took this charge as an accusation only of treating certain people cruelly.

Part of my reluctance to employ humanity as Sugden recommends is based upon my current view that humanity is rooted in a parent virtue, namely, benevolence. Here I think I am in agreement with Hume. I think the most natural way to read Hume on humanity, a quality he in

fact seldom discusses directly, is that he regards humanity as benevolence extended beyond one's nearest and dearest. As Hume argues so eloquently, benevolence tends to be both indeterminate and partial whereas justice is both determinate and impartial. One has considerable leeway in choosing how and to whom one expresses benevolence. In the example of the wealthy man who is obliged to give some of his excess wealth to the poor, which Hume gives in the context of his discussion of benevolence, Hume does not say which of the poor the rich man should aid or how much aid he should provide them. Hume also argues that each individual act of benevolence tends to produce immediate good whereas individual acts of justice might fail to produce immediate good, yet be necessary to maintain the integrity of the whole system that produces good. For these reasons Hume maintains the duties of benevolence are natural while those of justice are artificial (*Treatise* 3.2.2,6, *Enquiry* Appendix III:5). Viewing humanity as extended benevolence would help to explain both its origins and why humanity plays what I think is a rather limited role in social life. People who have become accustomed to extending their benevolence to those to whom they have closer ties might start extending their benevolence more generally and often in a weaker manner. I suspect that while most people have some tendency to act humanely, perhaps the only definite obligation of humanity they recognize is to refrain from gratuitously causing suffering.

However, Sugden perhaps unconsciously suggests a different possibility when he argues that a natural obligation to make provision for vulnerable individuals could emerge as an outgrowth of a duty of justice to reciprocate the aid others have extended in situations like the repeated Mutual Aid game. This would make justice the parent virtue of humanity. And this possibility is one more reason Sugden's intriguing recommendation merits further exploration.

§4. Value Domain Questions

Gaus rightly observes that while I discuss agent domain questions at some length, especially regarding the vulnerable, I do not address some important parallel value domain questions. In particular Gaus raises questions of evolving political commitments and other-regarding preferences. I will respond first to Gaus' political commitments challenge. Figure 4 summarizes a generalization of Gaus' Tory-Laborite Battle of the Sexes game where each side can support either Tory rule (T) or Laborite rule (L), and they have government exactly when they follow the same pure strategy.

[Figure 4. Tory-Laborite Game]

So long as $\alpha_i \geq 2$ for $i = 1, 2$, the Tories and the Laborites both fare better by following some arrangement where they alternate between (T,T) and (L,L) than they fare at a no government outcome. Such an alternation schedule is a correlated equilibrium convention so long as they are at both (T,T) and (L,L) at least some of the time and thus always have *some* government.[9] But each side strives for perpetual rule, so that in the end it will avoid having to compromise its own principles to any degree. Now to the core of Gaus' challenge: Suppose the two sides are pushed into adopting a Liberal Democratic constitution requiring them to follow a compromise scheme. To make Gaus' example even more specific, suppose this constitution institutes a correlated equilibrium $\boldsymbol{f}$ according to which they alternate between (T,T) and (L,L) for half the time each. If $\alpha_1 = \alpha_2 = 3$ as in Gaus' version of this game, then at $\boldsymbol{f}$ each side achieves an overall expected payoff of $\frac{23}{2}$, far less than its ideal payoff of 20. Gaus proposes that the Liberal Democratic constitution remains in force long enough for the descendants of the Tories and

Laborites who saw this constitution enacted to replace their ancestors, and that these young Tories and Laborites now regard the incumbent regime as the best regime. For the equilibrium $\boldsymbol{f}$, this is the case when $\alpha_1 = \alpha_2 \geq 20$. These young Tories and Laborites have had their own payoffs formed according to some process like the normative expectations formation process Sugden proposes (1998, 1986 pp. 214-218), where Sugden draws inspiration both from Lewis' analysis of convention as a coordination equilibrium (1969 p. 42) and from Hume's claim that community members come to regard the rules of an incumbent convention as having normative force (*Treatise* 3.2.2:24, 3.2.7:10). Gaus concludes that such a presumably happy ending cannot count as a just political regime according to my account of justice as mutual advantage, since now no one concedes anything for the sake of others as my account requires.

I think Gaus' politics example and a variant on it serve to clarify some of the limits of justice as mutual advantage, although perhaps these limits pose a serious practical problem only from one side of these limits. Gaus is right to conclude that the regime of his happy-ending example is not a just regime according to my account. For the young Laborites and young Tories are no longer in the circumstances of justice as I, or Hume before me, define these circumstances. The young Laborites and young Tories have found their way to a state similar to the imaginary utopia Hume describes where the preferences parties have over alternative outcomes are so closely aligned that justice is not necessary (*Enquiry* 3.1:6). Furthermore, I would not dismiss such an exit from the circumstances of justice as simply too unlikely to be taken seriously, despite the experiences of political parties past and present. Something like this happy ending may have occurred in the 20th century among many of the adherents of different religious groups as the culmination of a variety of loosely connected forces, including a somewhat organized ecumenical movement. In recent years, many people of faith have joyfully

joined forces with people outside their original religious communities, working and worshipping together in ways that their predecessors might have thought unimaginable.

But some of the other recent experiences of many religious people illustrate how justice as mutual advantage is limited from another side. Gaus can turn the screw in his example and have me consider a variation on the Tory-Laborite game where $\alpha_i < 2$ for at least one side. Suppose that $\alpha_1 = 3$ as in Gaus' own game and $\alpha_2 = 0$. This reflects a situation where the Tories would grudgingly accept Laborite rule in order to avoid the chaos of no government, but the Laborites would prefer such chaos over any positive chance of Tory rule. In this game L is strictly dominant for the Laborites, and the only equilibrium is (L, L) where the Tories submit completely. If the Laborites are so committed to their own principles, then the two sides fall outside the circumstances of justice. Moreover, if now $\alpha_1 = \alpha_2 = 0$, so that both sides prefer no government over anything less than complete rule by their own side, then the only equilibrium is (T, L), which bears an obvious similarity to Hobbes' State of Nature. This situation resembles the predicament of many belonging to religious groups both during and in the aftermath of 20th century ecumenicism, where some opposing factions are prepared to suffer the consequences of schism rather than make any concessions on questions they regard as simply nonnegotiable. Justice as mutual advantage cannot regulate the interactions of parties this intractably opposed. Indeed, I cannot see how any account of justice could regulate such parties without at least one side thinking itself wrongly oppressed. This variation on Gaus' example shows that justice as mutual advantage presupposes that even deeply opposed parties are prepared to reach *some* modus vivendi rather than go to a state resembling Hobbes' war of all against all. And that is indeed a serious practical limitation.

Gaus uses another Battle of the Sexes game to illustrate his other-regarding preferences challenge. In this game, Alf is an egoist whose ideal outcome is $\left(PR_2\left(p\right), PR_2\left(p\right)\right)$ where he receives all and Betty receives none, and Betty is other-regarding so that her ideal outcome is $\left(PR_1\left(p\right), PR_1\left(p\right)\right)$ where each receives half. Gaus observes that in a bargaining problem constructed from this game, the standard solution concepts would assign Alf three-fourths and Betty only one-fourth. Betty is apparently victimized to a significant extent because she is other-regarding while Alf is selfish. Gaus observes that even so staunch a defender of justice as mutual advantage as Gauthier might regard such an outcome as exploitative. One can note that Gaus' striking example might seem problematic only from a third party perspective.[10] From the point of view of someone like Hume's judicious spectator (*Treatise* 3.3.1:14) or Smith's better known impartial spectator (*Theory of Moral Sentiments* III.2:31, 3:3), such a lopsided distribution might appear plainly wrong, precisely because Betty loses for having developed an estimable quality that Alf failed to develop. But from an internal point of view, perhaps there is no problem. Given their preferences, the distribution where Alf receives three-fourths and Betty receives one-fourth produces the payoff vector $\left(\frac{5}{2}, \frac{5}{2}\right)$ characterized by their yielding equally to the other's most favorable equilibrium. This situation differs from that of cases where some parties are vulnerable, since in such latter cases the vulnerable might find distributions leaving them with little or none just as objectionable as might some third parties.

Much of the force of Gaus' example rests upon background assumptions that in Berry and Alf's culture, all of the women and none of the men are other-regarding and their choices are limited to $PR_1\left(p\right)$ and $PR_2\left(p\right)$. The asymmetry of Gaus' example can be greatly reduced by extending his game in some plausible directions. Figure 5 summarizes an extension of Gaus'

game where Alf and Betty now each have a new pure strategy $PR_3(p)$ that characterizes a social regime where Alf now receives nothing while Betty receives all.

[Figure 5. Extended Alf-Betty Game]

Since Betty is other regarding, her payoff at the $(PR_3(p), PR_3(p))$ outcome is $\delta < 3$. $(PR_1(p), PR_1(p))$ is still Betty's ideal outcome, but just by making the $PR_3(p)$-regime a possibility, the resulting bargaining problem is now such that the egalitarian solution has Alf receive two-thirds and Betty receives one-third. Betty can gain an even larger share if she and Alf are representatives of more mixed populations. Figure 6 summarizes a variation on the Figure 6 game where Betty is replaced with Claudia, an egoist like Alf.

[Figure 6. Alf-Claudia Game]

In the bargaining problem constructed from the Figure 6 game, the egalitarian solution has Claudia and Alf both receive half. Now suppose that the subpopulations of women and men are both mixed, so that there are other-regarders and egoists in both subpopulations. One can summarize an interaction between members of these populations as a Bayesian game that combines the payoffs of the Figure 5 and Figure 6 games. In this Bayesian game agents like Betty can get more than they would get if the populations are homogeneous. For example, if only half of the women are other-regarding and one-fourth of the men are other-regarding, then in the resulting bargaining problem where Alf and Betty are matched Betty now gets $\frac{3}{7} \approx 0.429$ while Alf gets $\frac{4}{7} \approx 0.571$. In other mixed populations where the proportion of other-regarding women is lower of the proportion of other-regarding men is higher, the resulting egalitarian solution comes even closer to equal division.

§5. Unsettled Issues

The general defense I present of social contracts characterized by monotone path solutions is a set of stability arguments. Again my approach is predominantly explanatory. The leading idea of this set of arguments, which I should have stated more explicitly in *Strategic Justice*, is that we come to regard the monotone path solution social contracts as just because these social contracts are stable in ways that other social contracts, including Nash and Kalai-Smorodinsky solution contracts, are not stable. The excellent stability properties of the monotone path solutions come at a price. These solutions presuppose that utilities are interpersonally comparable in some manner. Nash required his own solution to satisfy scale invariance, which effectively short-circuits the question of interpersonally comparable utilities. Nash regarded the idea of interpersonally comparable utilities as meaningless, as do many philosophers and social scientists. I belong to another camp who follow John Harsanyi in thinking that in at least some contexts, interpersonally comparable utilities are unavoidable and that philosophers and social scientists need to develop rigorous logical foundations for such utilities.[11] But as Bruner rightly notes, I do not say a great deal regarding this general project. In *Strategic Justice* my approach resembles Braithwaite's, in that I use some scalings proposed for interpersonal utilities in certain examples without trying to go farther. I regard the question of interpersonally comparable utilities as one of the great unsolved problems of social science. While I think there has been significant progress in recent years, particularly in Ken Binmore's landmark work (1994, 1998), we remain far from a complete solution. And to the extent the conventionalist project rests upon interpersonally comparable utilities, this project remains incomplete.

Sugden's and Gaus' comments expose another and closely related way my analysis is incomplete. In *Strategic Justice* I devote much discussion to how different conventions might

evolve in a community without exploring the origins of community members' preferences that are the ultimate basis of the payoffs of these conventions. I generally take the preferences of agents as given, and work with payoffs reflecting these preferences. Of course, an agent's preferences do not emerge from nowhere, and in particular Gaus and Sugden point out in different ways that preferences could possibly interact with the behaviors that support alternative conventions. As noted above, Hume thought that over time community members tend to attach normative significance to incumbent conventions. One way to understand Hume's claim is that when interacting agents follow a convention, their compliance reinforces their reciprocal expectations of future compliance, which in turn drives their preferences to evolve so that they come to regard the rules of their convention as part of the body of the *right* rules for interacting. Sugden is one of the few to propose a systematic account of this type of preference evolution with his theory of normative expectations. I did not explore the interaction between preferences and actions in *Strategic Justice* because I tried to present an account of justice as mutual advantage that is maximally flexible with respect to the preferences of the individuals regulated by a system of justice. The sources of an individual's preferences are many and sometimes conflicting. Such sources might include, though are not limited to: personal experimentation, family upbringing, political or religious indoctrination, and acquired tastes of activities such as consumption. I admittedly shy from exploring the complicated story of preference formation in more depth. But I would agree with Gaus and Sugden that the acquired tastes for following the rules of an incumbent convention may be an important part of this story. I conjecture that future work on the evolution of preferences will in fact strengthen the case for justice as mutual advantage. Sugden's work on normative preferences is a large step in this direction.

Lastly, and again relatedly, I think the relationship between justice as mutual advantage and the composition of a community a justice as mutual advantage system regulates requires further exploration. As Sugden and Gaus both recognize in somewhat different ways, agent domain and value domain problems become more likely as subpopulations become more homogeneous and their members more easily identifiable. As Gaus frames his example, the asymmetric distribution between Alf and Betty looks exploitative because they are representatives of a culture where all of the women, but none of the men, have been "nurtured" to be other-regarding. Sidgwick gives a similar example in *The Methods of Ethics*, claiming 'we should consider a law unjust which compelled only red-haired men to serve in the army, even though it were applied with the strictest impartiality to all red-haired men (1981 p. 267).' Why do we think such a law, which would be based upon an outcome of a natural lottery, would be unjust? And why do we think that some other selection method for conscription such as actually drawing lots might be compatible with justice? We think Sidgwick's fictitious law would be unjust because this law singles out the members of a distinct and easily identifiable subpopulation before the fact, assigning them demanding and dangerous service on the basis of some trait characteristic of this subpopulation not needed for the provision of such service. An artificial lottery assigns chances for this service in a somewhat egalitarian way across a wider segment of the population, and actual service is required only after the fact of the lottery. Gaus rightly observes that I exploit this idea in my discussion of the vulnerable, since I argue that everyone has some chance of falling into vulnerability at least some of the time, so that a system of justice that extends benefits to the vulnerable works partly as would an insurance scheme. In my response to Gaus above, I argue that value domain problems might be mitigated if various subpopulations are more mixed with respect to values than they are in Gaus' example. Perhaps

some heterogeneity among the members of various subcommunities together with some uncertainty among community members regarding subgroup membership are both necessary and sufficient conditions for preventing the emergence of domain asymmetries. But establishing this general claim, and in particular identifying how heterogeneity and uncertainly are needed to forestall domain asymmetries, is the stuff of future work.

REFERENCES

Barry, Brian. 1989. *Theories of Justice*. Berkeley: University of California Press.

Binmore, Ken. 1994. *Game Theory and the Social Contract Volume I: Playing Fair*. Cambridge, Massachusetts: MIT Press.

Binmore, Ken. 1998. *Game Theory and the Social Contract Volume II: Just Playing*. Cambridge, Massachusetts: MIT Press.

Braithwaite, Richard. (1955) 1994. *Theory of Games as a Tool for the Moral Philosopher*. Bristol: Thoemmes Press.

Bruner, Justin. 2020. 'Nash, bargaining and evolution'. unpublished manuscript.

Buchanan, Allen. 1990. 'Justice as reciprocity versus subject-centered justice'. *Philosophy and Public Affairs* 19: 227–53.

Harrison, Jonathan. 1981. *Hume's Theory of Justice*. Oxford: Oxford University Press.

Harsanyi, John. 1955. 'Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility'. *Journal of Political Economy* 63: 309-321.

Harsanyi, John. 1977. *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*. Cambridge: Cambridge University Press.

Hobbes, Thomas. (1651) 1994. *Leviathan*, ed. Edwin Curley. Indianapolis: Hackett Publishing Company.

Hume, David. (1740) 2000. *A Treatise of Human Nature*, ed. David Fate Norton and Mary J. Norton. Oxford: Oxford University Press.

Hume, David. (1777) 1998. *An Enquiry Concerning the Principles of Morals: A Critical Edition*, ed. Tom Beauchamp. Oxford: Clarendon Press.

Gauthier, David. 1986. *Morals by Agreement*. Oxford: Clarendon Press.

Kalai, Ehud. 1977. 'Proportional Solutions to Bargaining Situations: Interpersonal Utility Comparisons.' *Econometrica* 44:1623-1630.

Lenin, V. I. (1918) 2014. *State and Revolution*, annotated by Todd Chretien. Chicago: Haymarket Books.

Lewis, David. 1969. *Convention: A Philosophical Study*. Cambridge, Massachusetts: Harvard University Press.

Myerson, Roger. 1981. 'Utilitarianism, Egalitarianism, and the Timing Effect in Social Choice Problems.' *Econometrica* 49: 883-897.

Nash, John. 1950. 'The Bargaining Problem.' *Econometrica* 18: 155-162.

Nash, John. 1953. 'Two-Person Cooperative Games.' *Econometrica* 21: 128-140.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, Massachusetts: Harvard University Press.

Schelling, Thomas. 1960. *The Strategy of Conflict*. Cambridge, Massachusetts: Harvard University Press.

Sidgwick, Henry. (1907) 1981. *The Methods of Ethics*, 7th ed. Indianapolis: Hackett Publishing Company.

Smith, Adam, (1759) 1982. *The Theory of Moral Sentiments*. Raphael, D. D. and MacFie, A. L. (Eds.). Indianapolis: Liberty Fund.

Sugden, Robert. (1986) 2004. *The Economics of Rights, Co-operation and Welfare*, 2nd ed. Houndsmills, Basingstoke, Hampshirer and New York: Palgrave MacMillan.

Sugden, Robert. 1998. 'Normative expectations: the simultaneous evolution of institutions and norms', in *Economics, Values and Organization*, ed. Avner Ben-Ner and Louis Putterman. Cambridge: Cambridge University Press.

Thomson, William and Lensberg, Terje. 1989. *Axiomatic Theory of Bargaining With a Variable Number of Agents*, Cambridge: Cambridge University Press.

Thucydides. 1993. *On Justice, Power and Human Nature: Selections from* ***The History of the Peloponnesian War***, ed. and trans. Paul Woodruff. Indianapolis: Hackett Publishing Company.

Vanderschraaf, Peter. 2020. 'Stability Challenges for Moehler's Second-Level Social Contract'. *Analytic Philosophy* 61: 70-86.

---

[1] Bruner's example is similar to the example of this phenomenon I give on pp. 312-318 of *Strategic Justice*.

[2] This property is sometimes referred to as *weak stability* (Thomson and Lensberg 1989 p. 142).

[3] I discuss subgroup consistency and population monotonicity in more depth in a later essay (2020).

[4] A point $\boldsymbol{x} = (x_1,...,x_n)$ weakly Pareto dominates another point $\boldsymbol{y} = (y_1,...,y_n)$ if $x_i \geq y_i$ for each $i \in \{1,...,n\}$ and $x_i > y_i$ for at least one $i \in \{1,...,n\}$.

[5] Barry (1989 pp. 163) and Buchanan (1991 p. 232) give longer and far more eloquent statements of this general response.

[6] Saint Paul gives his version of this aphorism in 2 Thessalonians 3:10. Lenin gives his in *State and Revolution*, Chapter 5, Section 3.

[7] I do not assume that the members of society are necessarily all humans, but I argue that a community regulated by any system of justice is likely to limit membership to humans for reasons of salience (2019 pp. 294-298).

[8] Harrison makes much the same point in his discussion of Hume's weaker creatures example (1981, p. 277).

[9] If $\alpha_i > 2$, $i = 1, 2$, then (T, T) and (L, L) are themselves strict correlated equilibria and hence also characterize conventions.

[10] Brian Skyrms pointed this out in discussion of an earlier version of Gaus' essay at a workshop hosted by Chapman University in November 2019.

[11] See Harsanyi (1955, 1977 pp. 48-83). However, unlike me, Harsanyi is reluctant to presuppose interpersonal utility comparisons for the purpose of resolving bargaining problems (1977, pp. 192-195).

Figure 1. Braithwaite Problem Feasible Set

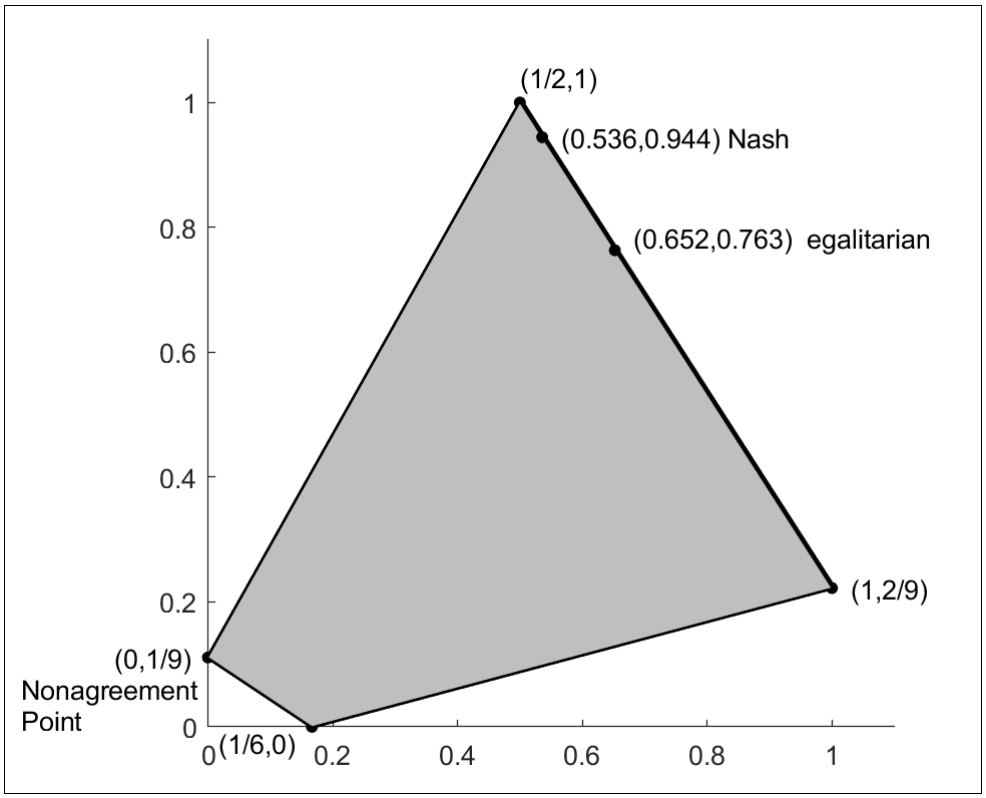

Figure 2. Braithwaite Basis Game

		Luke: M	Luke: G
Matthew	M	$\left(\frac{1}{6}, 0\right)$	$\left(\frac{1}{2}, 1\right)$
	G	$\left(1, \frac{2}{9}\right)$	$\left(0, \frac{1}{9}\right)$

$M =$ claim none, $G =$ claim all

Figure 3. Expanded Braithwaite Feasible Set

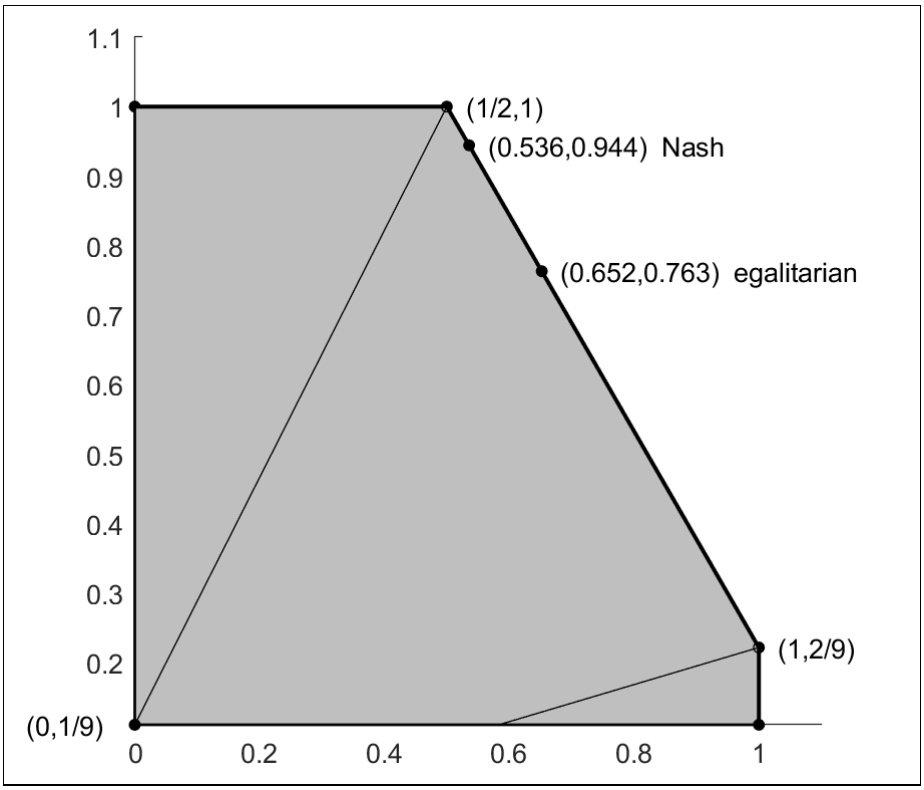

Figure 4. Tory-Laborite Game

		Laborites: T	Laborites: L
Tories	T	$(20, \alpha_2)$	$(2, 2)$
	L	$(1, 1)$	$(\alpha_1, 20)$

$T =$ support Tory rule, $L =$ support Laborite rule
$D =$ support Liberal Democratic rule

Figure 5. Extended Alf-Betty Game

		Betty		
		$PR_3(p)$	$PR_1(p)$	$PR_2(p)$
Alf	$PR_3(p)$	$(2, \delta)$	$(1,1)$	$(1,1)$
	$PR_1(p)$	$(1,1)$	$\left(\frac{5}{2}, 3\right)$	$(1,1)$
	$PR_2(p)$	$(1,1)$	$(1,1)$	$(3,2)$

$PR_1(p) =$ each receives half
$PR_2(p) =$ Alf receives all
$PR_3(p) =$ Betty receives all, $\delta < 3$

Figure 6. Alf-Claudia Game

		Claudia		
		$PR_3(p)$	$PR_1(p)$	$PR_2(p)$
Alf	$PR_3(p)$	$(2,3)$	$(1,1)$	$(1,1)$
	$PR_1(p)$	$(1,1)$	$\left(\frac{5}{2}, \frac{5}{2}\right)$	$(1,1)$
	$PR_2(p)$	$(1,1)$	$(1,1)$	$(3,2)$

$PR_1(p) =$ each receives half
$PR_2(p) =$ Alf receives all
$PR_3(p) =$ Claudia receives all