Is mutual advantage a general theory of justice? More domain worries

Gerald Gaus

Philosophical Studies
An International Journal for Philosophy
in the Analytic Tradition

ISSN 0031-8116

Philos Stud
DOI 10.1007/s11098-020-01501-3

Springer

Philos Stud
https://doi.org/10.1007/s11098-020-01501-3

Is mutual advantage a general theory of justice? More domain worries

Gerald Gaus[1]

1 Vanderschraaf's fundamental contribution

Peter Vanderschraaf's *Strategic Justice* is the culmination of one of the major philosophical projects of the second half of the twentieth century. The basic thought is compelling: perhaps social morality can be either derived from, or reconciled with, the best strategy for each rational agent to promote her own interests in cooperative contexts. Morality would then track self-interest. Until the appearance of *Strategic Justice,* 1986 was the apex of this approach, with the publication of David Gauthier's *Morals by Agreement*, Jean Hampton's *Hobbes and the Social Contract Tradition*, Gregory Kavka's *Hobbesian Moral and Political Theory* and Robert Sugden's *The Economics of Rights, Cooperation and Welfare.* Vanderschraaf's book is by far the most sophisticated, scholarly and thoughtful version of this comprehensive project.[1] Vanderschraaf has a better grasp of the philosophical and game theoretic background than any previous exponent, and builds on this to yield some stunning innovations and insights. *Strategic Justice* is a tremendous achievement—it requires hard work and careful study, but the rewards are great.

[1] Michael Moehler's recent *Minimal Morality* (2018) presents a non-comprehensive statement of this project. See Sect. 2 below.

✉ Gerald Gaus
jerrygaus@gmail.com

[1] Moral Science & Political Economy, Philosophy, University of Arizona, Social Science Bldg. Rm 217, Tucson, AZ 85721-0027, USA

Published online: 27 June 2020

2 The domain and range of justice as mutual advantage

A decade of philosophical interest and investigation followed the remarkable year of 1986, until the project seemed to depart the philosophical stage as abruptly as it had (re)appeared. Much—in my view far too much—of this commentary was focused on Gauthier's rather unique solution to the bargaining problem.[2] More important are deep worries concerning the specification of the *domain* of the agents whose agreements generate justice, and so the *range* of the resulting principles (conventions, etc.). The familiar *Agent Domain Worry* focuses on restricting the domain of agents who enter into agreements/conventions to those able to engage in mutually beneficial interactions. If the domain of agents who devise morality through mutual benefit is restricted to such cooperators, it seems the range of the resulting moral outputs would also be restricted to them—they alone would be the subjects of morality. Gauthier seemed willing to embrace this implication: "Animals, the unborn, the congenitally handicapped and defective, fall beyond the range of morality tied to mutuality."[3] That is, given that the inputs (domain) of mutual advantage is limited to cooperators, the *range* of the resulting moral principles do not cover non-cooperators. One of Vanderschraaf's contributions to the mutual advantage project is to show how the vulnerable and unproductive might be covered in the range of justice: "the Vulnerability Objection is not the 'silver bullet' that liquidates justice as mutual advantage after all" (p. 282).[4]

I shall largely focus here on a different concern—the *Value Domain Worry*. This worry is that justice as mutual advantage only applies to a limited domain of value interactions: it applies only to a subset of conflictual interactions that we need justice to resolve. If so, we might say, justice as mutual advantage is, at best, a partial theory of justice: it cannot account for the full *range* of our judgments of justice, or even the full range of social conflicts that we need justice to resolve.[5] In his recent *Minimal Morality*, Moehler presents a case for mutual advantage as applying only in the restricted domain of prudential agents, and so deals only with certain "cases of conflict."[6] In non-conflict cases, other understandings of morality are applicable. I shall argue here that Moehler's intuition is correct: mutual advantage is not an account of the full range of justice. I shall also argue that, somewhat disturbingly, it is precisely its most powerful tools—game theory and bargaining theory—that obscure this.

[2] This was, of course, "minimax relative concession," a close cousin of the Kali-Smorodinsky bargaining theory. See Gauthier (1986) chap. V. As is well-known, Gauthier later abandoned this approach and, indeed, eventually (Gauthier (2013)) abandoned bargaining theory altogether.

[3] Gauthier (1986), p. 268.

[4] All parenthetical page references in the body of the text refer to Vanderschraaf (2019).

[5] This in itself is not a deep criticism, as I am skeptical that there is any fully comprehensive theory of justice. See Gaus (2011), pp. 551–557. However, my worry is that, at best, justice as mutual advantage is restricted to a relatively narrow range of conflicts, as I argue in Gaus (2019).

[6] Moehler (2018), p. 15.

3 A condition under which mutual advantage obtains

An important innovation of *Strategic Justice* is a new analysis of the circumstances of justice (pp. 110ff), which is then developed into an analysis of the circumstances under which justice as mutual advantage applies (pp. 275ff). For the sake of brevity, I will focus on just one of Vanderschraaf's five conditions, since I think it leads to the puzzles I wish to explore.

To begin, assume that all individuals have some values, interests, aims or concerns, which we will call *V*, so *V*i is person *i*'s *V(alue)*. I assume that *V*i can yield something like *i*'s ordering of outcomes over some feasible set $\{O\}$.[7] So $V_i(O_x) \succ V_i(O_y)$ is to be read as "person *i* ranks O_x above O_y in terms of her *V*."[8] Call p the population subject to a set of rules R_p. We can partition p into $p^{\mathbf{C}}$ (the contributing population) and $p^{\mathbf{V}}$ (the vulnerable population)—this latter group are not active participants in production, acting upon, or enforcing R_p. The condition on which I shall focus is:

> **M1** *Conflict (The Need for Compromise).* In R_p, a system of rules in population *p,* $V_i(R_p)$ is person *i's* value satisfaction *(*V_i) under universal conformity to R_p. **M1** requires that if R_p is to qualify as justice under mutual advantage, each *i* in $p^{\mathbf{C}}$ (the cooperating population) restrain the pursuit of V_i to some extent in order to advance for some person *j (*in *p),* V_j to some extent.[9] **M1** thus requires that for R_p to qualify as a case of justice as mutual advantage, it must be that for all persons *i* in $p^{\mathbf{C}}$, $V_i(\text{MAX}) \succ V_i({R_p}^C)$, where the former is the maximum value that V_i can achieve in $\{O\}$.

4 The agent domain worry

Before moving on to the Value Domain Worry, it is useful to point out how our reformulation of **M1** helps brings out the Agent Domain worry. Unlike Vanderschraaf, I have included a population parameter to make it clear that any claim about how high V_i is must be indexed to some population. I, at least, cannot see how that can be avoided. I have tried to be careful in my specifications to capture Vanderschraaf's subtle indication that we should not define p simply in terms of contributors: he partitions p into those who are capable of pursuing their values (p^C) and those who are not (p^V), providing the basis for his later argument that there is a case for including in the range of justice in p those in p^V (thus dodging the "silver bullet"). Vanderschraaf's ingenious argument is that in all populations people transition from membership in p^V to p^C and perhaps back again, so we can view a

[7] For simplicity's sake, let us assume that V_i yields a cardinal score for every member in {O}, and has a maximum value over {O}.

[8] Or, alternatively "person *i* prefers O_x to O_y on the basis of her *V*."

[9] "(M1) Conflicting interests …. requires each Party capable of pursuing interests to restrain pursuit of her own interests to some extent in order to advance the interests of other parties to some extent" (pp. 275, 276).

population as a series of iterated games over periods in which people in p switch from contributors to non-contributors. Moreover, contributor Alf can have reasons to include non-contributor Betty in the range of justice if Charlie will punish Alf for not doing so (p. 290). Nevertheless, as far as I can see, a version of the Agent Domain Worry remains as long as there is, persisting over time, a proper subset of p, p' such that for person i, $V_i(R_{p'}) \succ V_i(R_p)$.[10] If i and likeminded others can leave the p population for the p' (or build the best wall ever wall around p'), I cannot see why justice as mutual advantage says they should not. In this case person i does better through cooperation in the smaller p' population. This is not a problem about minimal or core coalition controlling the distribution of a fixed pie, or about consistency of distributive principles over subparts of the population[11]: it is about i's optimal population for producing the largest per capita social surplus. I do not see any argument for a maximally large p, so in this sense an Agent Domain Worry persists.

5 The value domain where V = interests

Throughout the mutual advantage literature the idea of a person's "interests" is repeatedly invoked: indeed, the term "divergent interests" occurs in the subtitle of *Strategic Justice.* This is, of course, not a well-defined notion, but it is most at home in talk about self-interest, and in division problems where individuals have some good or cost to divide among them. If we adopt this narrow, but I think core, use of "interests" then I believe **M1** is well motivated. Consider, for example, a cake division problem among p claiming parties. According to **M1** person i cannot demand the entire cake, for that would mean that others receive no advantage; so it seems clearly right that a rule of mutually advantageous division must give i less than his MAX, so $V_i(\text{MAX}) \succ V_i(R_p)$. Here some sort of bargaining solution is sensible. In resource division problems, or where there is a conflict about the benefits and costs of joint activity such that the parties have partially divergent self-interests (such as in Braithwaite's game between the neighboring musicians, Luke and Matthew (pp. 99ff)), bargaining solutions make sense. These interactions have salience in the justice as mutual advantage literature, since they manifestly fit its model of conflict and are susceptible to bargaining solutions.

Moehler, I believe, appreciates that the intuitions behind justice as mutual advantage and the appropriateness of the tools it employs—such his own invocation of a modified Nash bargaining solution—are most compelling when focused on conflicts among "*Homo prudens*," who greatly values her life, autonomy and the satisfaction of her basic needs.[12] If we accept that justice as mutual advantage is tied to conflicts between members of the species *Homo prudens*, two options present themselves. *First*, one might follow what often seems Hobbes's lead, and hold that

[10] Some other conditions are needed, but these are easily met.

[11] See Young (Young 1984), pp. 121, 122.

[12] Moehler (2018), p. 113.

Homo prudens is the best general model of *Homo ethicus*: since we basically pursue our self-interest, morality is really about how to reconcile these self-interests. In this case, justice as mutual advantage remains a general theory of justice. Alas—this seems simply false: though we are self-interested, we are many other things as well, and many of the problems of social life and morality arise from these un-self-interested features, as I will show below. The *second* alternative is Moehler's view: accept that justice as mutual advantage is at home in disputes about self-interest, but then accept that this shows that it only applies to a partial range of moral problems. It is then not a general or comprehensive theory of justice.

6 The value domain where V = other-regarding values

Vanderschraaf is crystal clear that he does not think that justice as mutual advantage is restricted to conflicts about self-interest. He rightly dismisses the all-too-common idea that game theory, because it relies on *Homo economicus*, is restricted to self-interested interactions: "both serious interests in ends other than material gain and passionate concern for the welfare of others are fully compatible with the notion of *homo economicus*, properly understood" (p. 5). More importantly, in his analysis Vanderschraaf is explicit that "In addition to their own selfish concerns, members [of population p] might wish to advance the conflicting agendas of various communities represented in society, such as political movements or religions. Even a society *whose members' interests are purely altruistic* can satisfy (**M1**) …." (p. 276, emphasis added).

On one reading, this sentence claims that only if all members *are purely altruistic* (or purely selfish, or purely ideological) does justice as mutual advantage apply. Justice as mutual advantage would thus suppose what might be called *type-homogeneity*: while purely self-interested people might have different aims, they are all of the purely selfish type. And so too with pure altruists. If this is the correct interpretation, justice as mutual advantage simply cannot be a comprehensive theory of justice, especially in a diverse society. Humans are of all types, from the relentlessly selfish to the incredibly altruistic, but most are somewhere in-between, and in-between in different ways. One of the consistent findings of experimental economics is that people are of different types, with varying degrees of pro-social values.

To be a candidate for a comprehensive theory of justice relevant to resolving conflicts in a diverse society, then, justice as mutual advantage must be able to handle conflicts between different types. It must be consistent with *type-heterogeneity*. On this alternative reading Vanderschraaf is claiming that justice as mutual advantage applies to all conflicts (properly described as expressing the circumstances of justice) among all V_i types, such as a conflict between an egoist and a more pro-social or fair-minded agent. This comprehensiveness claim goes well beyond Gauthier's. While Gauthier stressed that justice as mutual advantage is not restricted to self-interest (it includes interests *of* the self, not simply interests *in*

the self),[13] he saw his "non-tuism" assumption as precluding a concern for others. As he put it, morals by agreement presupposes "mutual unconcern." Gauthier seems to argue for this partly on moral grounds: feminist thought, he says, has taught us how allowing altruistic preference can justify exploitation.[14]

To see this better, recall the basic "Battle of the Sexes" game," which Vanderschraaf takes as a quintessential instance of the circumstances of justice.

Here Alf and Betty have cardinal preferences (3 = best) over different proposed rule systems (PR_1, PR_2)[15] and I assume that in this case the only difference in these rule systems is how they divide up a fixed amount of resource X, so their most preferred rule system is selected by their V function over this matter. Suppose in this case $PR_1 = V_{\text{ALF}}(\text{MAX})$, and $PR_2 = V_{\text{BETTY}}(\text{MAX})$. Now, inspired by Gauthier's feminist claim, suppose that Betty's V_{B} function is such that V_{B} (MAX) occurs when the resource (X) is divided half to her and half to Alf; putting her share second, V_{B} (MAX) = (½X, ½X). Alf, on the other hand, has a purely self-interested value function such that V_{A} (MAX) occurs at (X, 0).[16]

Gauthier, I think, would already see this as exploitative, since Betty's maximum includes sharing with Alf but his does not include sharing with her. And many philosophers seem to agree. There is such deep suspicion that pro-social preferences for the good of others (more often ascribed to women) are the result of "adaptive preferences" or ideologically-induced self-sacrifice, that many adopt Gauthier's view that they should be excluded from the justification of morality, treating them as if they were illegitimate. The start of analyzing human morality is to be the (counterfactual) assumption that we do not care about others. I find this as implausible as it is unpalatable. Humans value the good of others in diverse ways. Vanderschraaf agrees: his innovative case for including the vulnerable in justice as mutual advantage presupposes that some in p^{C} may care for those in p^{V} (p. 286). Note, then, that in his own analysis Vanderschraaf supposes different types—some who care about the vulnerable and some who do not.

While I do not see Betty's valuing most highly a 50/50 split as inherently an indication of exploitation, **M1** *requires* that on the acceptable rule system R_{p}, Betty *cannot* claim even her half, for it *must* be that $V_{\text{B}}(\text{MAX}) \succ V_{\text{B}}(R_{\text{p}})$. Given some additional plausible assumptions about the shape of their utility functions, if there is an "egalitarian bargain" over the *utilities* in Fig. 1, it can be shown that she must move off of her preferred distribution towards Alf's, and so now she *must* get *less* than ½X! It is easy to generate a "fair" bargain where Alf gets three-quarters of X. This does look awfully odd, and now perhaps we do have a sort of exploitation: someone who has already taken into account the welfare of others is *required* to concede even more. Only if we are captivated by the numbers in the utility functions in Fig. 1 would we think that somehow a mutually advantageous morality *must*—conceptually—give

[13] Gauthier (1986), pp. 7, 11, 87.

[14] Though he also suggests that it is inconsistent with true rational endorsement (1986), p. 11.

[15] Recall that we said V_{i} ranks options in the feasible set.

[16] We can assume that the other elements of PR_1 give Betty reasons to coordinate on it. Alternatively assume V_{A} (MAX) = (.99X, .01X).

Fig. 1 Battle of the sexes game

		Betty: $PR_1(p)$	Betty: $PR_2(p)$
Alf: $PR_1(p)$		2, 3	1, 1
Alf: $PR_2(p)$		1, 1	3, 2

Betty less than her first choice. Note that if we focus on self-interests (how much X each gets), even if Betty received $V_B(\text{MAX})$, they still equally divide X. So X-wise $V_B(\text{MAX})$ specifies a mutually advantageous outcome where no one gets their maximal share, whereas utility-wise Betty gets everything she wants. **M1** thus strikes me as an unreasonable constraint on a just resolution in *this* domain—which is to say that justice does not *require* a bargain. On my view Betty's preferred *PR*2 is a perfectly acceptable solution.

7 The value domain where V = ideological value

Recall that according to Vanderschraaf justice as mutual advantage applies to "conflicting agendas of various communities represented in society, such as political movements or religions." So it is supposed to yield just resolutions of ideological disputes. Thus far I have focused on asymmetric cases, where the agents are of different types. Let us now assume they are of the same type—here they are both ideologues, one a true-blue British Tory and the other a Labourite. Recall our basic interaction, now slightly modified in the Battle of Ideologies (Fig. 2).

Each loathes living under the rules of the other, yet agrees with Vanderschraaf that even such a life would (barely) be better than the free-for-all of the state of nature. If they cannot coordinate, each prefers upholding thier own rules to the odd case where they uphold the other's rules. Vanderschraaf claims that **M1** applies to conflicts between such political or ideological utility functions, so the just resolution *cannot* be either Tory or Labour rules, since that would give one party their MAX. Again, given some plausible assumptions, an egalitarian bargain might end up identifying something in the middle as the fair and just resolution—say the Liberal Democrat's manifesto! But about the only thing they agree on is that the Lib Dems do not uniquely specify fair or just rules. The problem is that each already has a very firm notion of justice and fairness: *that* is what their dispute is about. But given that, how can they submit to a bargain about what justice is? They already each believe they know.

Vanderschraaf's account seems to assume that justice as mutual advantage can be a meta-theory of justice: justice about disputes about justice. But once one of the parties accepts that the Lib-Dem manifesto identifies justice, it seems they have given up on their ideology (it becomes what is sometimes called a "mere political preference"). If the Tory position is correct, the Lib Dems specify an *unjust*

Fig. 2 Battle of ideologies game

		Labour	
		Tory Rules	*Labour Rules*
Tory	*Tory Rules*	3 20	2 2
	Labour Rules	1 1	20 3

position. The idea that they could be backed into it via justice as mutual advantage must surely seem disconcerting to them.

Suppose, however, many of our Tories and Labourites are backed into it. Many accept the Liberal Democrat manifesto as specifying the just resolution, and come to live under it. Assume now the nightmare outcome for both: having lived under Liberal rules, many of their children become fervent Liberals! We might think that justice as mutual advantage has been vindicated in the long-term, finding a stable compromise solution. Quite the contrary: now that some people think being a Liberal is the best ideology (suppose that previously not even the Lib Dems thought that), **M1** automatically precludes Liberalism as an acceptable specification of justice, for now Liberals get their V_{LIBERAL} (MAX), and **M1** prohibits that. Despite Vaderschraaf's important demonstrations of various dynamic stability properties of his proposed bargaining solution, in this case it is inherently unstable: as soon as the bargain becomes the most favored outcome *for anyone* it is disqualified as genuinely of mutual advantage.

8 The wonders and worries of numbers

I count myself as an advocate of formal modeling, and an enthusiastic fan of *Strategic Justice*. By applying sophisticated game theoretical tools—including models of dynamic interactions—Vanderschraaf illuminates one issue after another. He has, for example, forever changed my thinking about Hobbes's state of nature and his problem of forming the Leviathan. And Vanderschraaf's case for including the vulnerable in a system of mutual advantage is a major advance in the theory. The other side, though, of enthusiasm for formal models is care about the range of their applications. The elegance of game theory with its abstract utilities and its clear mathematics can make us forget that sometimes the numbers refer to self-interests, sometimes the value of helping others and yet other times valuations of fairness or ideological systems. I have tried to show here how resolutions that make perfect sense for one domain of conflict can be unacceptable for others. That any complete and consistent value-based ordering of options can be represented by a

utility function by no means implies that everything important about a value-based ordering, or about a person's judgment, is captured by that utility function.[17]

Acknowledgements A version of this paper was presented at the symposium on Vanderschraaf's *Strategic Justice* at the Smith Institute for Political Economy and Philosophy at Chapman University. My thanks to John Thrasher for organizing the symposium, and to all the participants—especially Peter Vanderschraaf—for their comments and insights.

References

Gaus, G. (2011). *The order of public reason*. Cambridge: Cambridge University Press.
Gaus, G. (2019). Moral conflict and prudential agreement: Michael Moehler's minimal morality. *Analysis, 79*(January), 106–115.
Gauthier, D. (1986). *Morals by agreement*. Oxford: Oxford University Press.
Gauthier, D. (2013). Twenty-five on. *Ethics, 123*(July), 601–624.
Moehler, M. (2018). *Minimal morality*. Oxford: Oxford University Press.
Sen, A. (1977). On weights and measures: informational constraints in social welfare analysis. *Econometrica, 45*(October), 1539–1572.
Vanderschraaf, P. (2019). *Strategic justice*. Oxford: Oxford University Press.
Young, H. P. (1984). *Equity: In theory and practice*. Princeton: Princeton University Press.

[17] The view I am arguing against, which is assumed by many analyses to be a necessary truth, is what Sen (1977) calls "welfarism," a view that is informationally restrictive, and in many ways controversial.