Synthese (2022) 200:125
https://doi.org/10.1007/s11229-022-03599-w

ORIGINAL RESEARCH

Sharing with the vulnerable? The Vulnerability Objection and Vanderschraaf's theory of justice as mutual advantage

Lina Eriksson[1]

Received: 23 June 2021 / Accepted: 27 January 2022 / Published online: 11 April 2022

Abstract

The most recent major contribution to the literature on justice as mutual advantage is Peter Vanderschraaf's book *Strategic Justice*. In this book, he develops a theory of justice as convention, where justice is those principles that rational, self-interested agents would choose to solve problems of partially conflicting interest. His theory is thus a kind of theory of justice as mutual advantage. A common criticism of theories of justice as mutual advantage is the Vulnerability Objection: if the principles of justice require that resources are only shared with those that are net-contributors to a cooperative surplus, then those that are not net-contributors to that cooperative surplus (so-called *vulnerable* people) have no claim of justice to any share of the resources. But, the objection states, surely justice cannot exclude people simply because they are vulnerable. Vanderschraaf argues that his theory of justice as convention successfully answers the Vulnerability Objection. However, in this paper, I argue that although Vanderschraaf's theory successfully demonstrates that it can be to the mutual advantage of rational, self-interested people to agree to share equally even when some people contribute more than others, the problem remains of why such people would share with those that can never contribute more to the cooperative surplus than what they would withdraw from it. Vanderschraaf's solution is to weaken the requirement that people actually contribute. But that, I argue, undermines the claim that his theory is a theory of justice as mutual advantage.

Keywords Justice · Mutual advantage · Convention · Vulnerability Objection · Disability

Special Issue Name: T.C. : Indeterminacy and Underdetermination Lead Guest Editor : Dr. Mark Bowker and Dr. Maria Baghramian.

✉ Lina Eriksson
lina.eriksson@gu.se

[1] Gothenburg University, Gothenburg, Sweden

1 Introduction

Understanding justice as a set of rules for mutual advantage has a long history. The basic idea is simple enough: justice is whatever rational, self-interested agents would agree to under such-and-such circumstances, and they will agree to those principles for regulating their co-existence and the distribution of the *cooperative surplus* that are to their mutual advantage. The intuition behind this idea is that if agents A, B and C have, through mutual effort and contributions, together produced a surplus good, then they have a claim of justice to that good, and do not need to share it with others, who did not participate in its production. A person D, who did not contribute to A's, B's and C's cooperation, does thus not have a claim of justice to a share of the gains A, B and C have produced. Of course there are all sorts of qualifications to be made to that claim. But the basic intuition is shared in some way or another by advocates for many different theories of justice, not just theories of justice-as-mutual-advantage (hereafter, I will refer to theories of justice-as-mutual-advantage as theories of JMA). But for JMA-theories, the intuition plays a particularly important role, because according to such theories, justice is whatever rules for distributing the cooperative surplus agents (under such-and-such circumstances) would agree to for their mutual advantage.

However, theories of justice as mutual advantage have never quite become dominant, and one of the main reasons for this is the Vulnerability Objection: if agents A, B and C are all able to benefit each other by cooperating, but agent D cannot contribute anything of value for A, B and C to the collaboration, then A, B and C have no duty of justice to share the cooperative surplus with D. Or, differently put, those people who are ill, severely disabled, or in some other way unable to contribute much to the collaboration, have no standing in terms of justice. But many people think that justice works the other way around; if justice requires anything, it is that resources are shared in such a way that nobody will starve to death because they are ill or disabled. Since theories of justice as mutual advantage seem to claim that the vulnerable are owed nothing in terms of justice, then theories of JMA must be wrong. That, in essence, is the Vulnerability Objection to theories of JMA.

The most recent, and certainly impressive, account of a theory of JMA, is Peter Vanderschraaf's book *Strategic Justice.* In this book, Vanderschraaf sets out to construct a theory of JMA that is not vulnerable to the Vulnerability Objection. More specifically, he aims to construct a theory of justice-as-convention, according to which the principles that rational, self-interested agents agree to for their mutual advantage are principles that solve particular games of partially conflicting interests. These principles are called *conventions*. In this paper, I argue that although much of Vanderschraaf's construction of his theory of justice-as-convention is interesting and convincing in its own right, the theory falls short of meeting the Vulnerability Objection in a satisfactory way.

2 Theories of justice as mutual advantage

At the heart of a theory of JMA is the intuition, described above, that only those that have contributed to a cooperative surplus have a claim to a share of that surplus. Justice

is whatever rules for distributing that surplus that the contributors agree to for their mutual advantage.

Vanderschraaf's book is a defense of theories of JMA, but not only that: he argues that properly defined, theories of JMA are theories of justice as convention. What he defends, more specifically, is the claim that justice should be understood as conventions, developed or created to solve interaction problems of partially conflicting interests in a way that is to the mutual advantage of the agents involved. He writes:

> I have set myself to answer a broad question: Is a satisfactory general justice-as-convention theory possible? My own proposed affirmative answer to this question builds upon the general analysis of convention I give here. I define a convention as a system of strategies that characterize an *equilibrium* solution to a problem of coordination that has a plurality of such equilibrium solutions. Conventions of justice are special cases where the corresponding equilibria solve problems of *conflictual coordination*, in which the interests of the agents involved both coincide to some extent and diverge to some extent. Justice understood this way is *strategic* justice. (Vanderschraaf, 2019, p. xiii)

In the beginning of the book, he discusses five famous dilemmas of interaction. One is Hume's example of the Farmer's Dilemma; a group of neighbouring farmers could all harvest more of their crops if they cooperated by helping each other in turn, but because all of them doubt that their help would be reciprocated when it was their turn to harvest their crops, no help will be given, and all will lose some of their crop as a result. Another such dilemma is the Stag Hunt: if a group of hunters cooperated, they could kill a deer, which would yield more meat for each of them than the hares they could all kill if they hunted separately. But because they don't trust each other to stick to the deer hunt plan, suspecting instead that each would abandon their post to chase after a hare if they happen to see one, none of them is willing to give up his chance of at least killing a hare. Therefore they will all end up hunting hares and will never get a deer. What these and Vanderschraaf's other examples of dilemmas of interaction have in common, is that they are cases in which agents have partially conflicting but also partially aligning interests. Vanderschraaf develops a particular account of convention to describe the kind of cooperative solutions that can develop among the agents facing such dilemmas. These conventions help agents realize a cooperative surplus (for example, harvesting more crop, or getting more meat because of a successful deer hunt), and are therefore to the mutual advantage of these agents. Justice is the conventions that are created to solve such dilemmas of partially conflicting interests. One of the problems with taking justice to be conventions is that conventions are normally understood to be arbitrary in the sense that it does not matter to agents what coordination solution is chosen, as long as one is indeed chosen. But obviously it matters to people in for example the Farmer's Dilemma whether the farmers all cooperate and help harvest as much of their crops as possible, or whether everyone does worse by working separately. But Vanderschraaf argues that the arbitrariness of conventions should be understood in a different way: not as arbitrariness in the sense of indifference, but rather as in a discretionary sense, meaning that agents have multiple options, not that they are indifferent between those options (Vanderschraaf, 2019, p. 45).

Vanderschraaf specifies four necessary conditions of a theory of JMA. He then argues that these conditions "imply that the requirements of a justice as mutual advantage system are conventions" (Vanderschraaf, 2019, p. 272). These four necessary conditions—what Vandershraaf calls 'descriptive assumptions' (Vanderschraaf, 2019, pp. 275–276)—are the following:

M1: Conflicting interests.
M2: Pareto-improvement for contributors.
M3: Negative mutual expectations.

The idea is that when people have partly conflicting interests, they can choose to restrain their own actions in pursuit of their interests a bit, because if everyone does that, each member will in fact be better off than if nobody restrains their pursuit of their interests at all. Further, each person will obey the rules for the reason that if they don't, others will no longer restrain themselves with regard to this person, and she would end up worse off than she would be, would she to obey the rules. When people comply with a set of rules that satisfies M1-M3, they will thus be better off than they would be, if nobody complied. The extra benefit thus generated is called (with Gauthier's terminology) *the cooperative surplus*.

Vanderschraaf also adds

M4: Positive mutual expectations.

This condition states that any given person expects that if she were to comply with the rules that satisfy M1-M3, others would indeed restrain their behaviour towards her in accordance with those rules.

A crucial point is that writers about JMA also seem to take for granted that only those who contribute to the cooperative surplus stand to receive a share of it. Vanderschraaf calls this the *Contribution principle* (Vanderschraaf, 2019, p 279). He writes:

> [The Contribution principle] might look redundant in light of (M3). Indeed, (M3) would be redundant if each member necessarily either follows the requirements of the system or violates them. But to assume this is to overlook an important third possibility. Perhaps some members of society are simply unable to pursue interests at all. Such members might have interests, but they are in no position either to pursue these interests or to limit the pursuit of their interests through their own efforts. Once one admits this possibility, one must at once face the Vulnerability Objection. (Vanderschraaf, 2019, p. 280)

The people who cannot pursue their own interests, or limit such pursuits of their own efforts, are *vulnerable*. They cannot by themselves secure the resources they need for survival and cannot contribute resources to a cooperative surplus by cooperating with others. Because they cannot pursue their own interests, they can also not affect others negatively by failing to restrain their pursuit of their interests. But most people considered vulnerable are not completely unable to pursue or restrain their pursuit of their interests. It is not a matter of all-or-nothing. Rather, to be more precise, they contribute less to the cooperative surplus than what they need to withdraw from that surplus in order to survive. What matters is whether their inclusion in the cooperative venture generates a net gain for others (or at least does not generate a net loss—I leave

aside the question of whether rational, self-interested people would include those that generate neither a net gain nor a net loss).

This description of who is to be considered vulnerable is focused on the contribution of *resources*. But it is not obvious what would count as a contribution. In many of the examples Vanderschraaf gives of the kind of dilemmas that motivate a theory of justice as convention, contributions consist of actions that increase the supply of resources: meat, harvested crops, etc. But other examples focus more on how people can inconvenience each other when they pursue their interests without regard for others' interests. For example, two musicians, faced with the problem that they each need peace and quiet to practice their instruments effectively and so cannot practice at the same time, need to agree on who gets to practice when. In that example, a contribution is better understood as refraining from doing something: one of the musicians in the example contributes by refraining from practicing their instrument, so as to let the other person practice theirs. Similarly, a person can be seen as contributing by refraining from stealing or murdering others—they contribute to others' safety and protection of private property by refraining from pursuing their interests through theft and murder. In the case of refraining from pursuing one's interests in a way that negatively affects others, a vulnerable person is a person who cannot pursue their interests in a way that negatively affects others. They are thus no threat to others. But since they are no threat to others, others have no incentive to include them among those towards whom they in turn restrain themselves. After all, no matter what they do, the vulnerable person cannot do anything bad to them.

We can thus think of vulnerability in this context in two different ways: as inability to contribute to the creation of a resource surplus, or as inability to pose a threat or be a potential inconvenience to others. I will in what comes mostly focus on the contribution to a resource surplus, and thus vulnerability as the inability to produce resources or in other ways contribute to the creation of a resource surplus. But the discussion that follows below applies to both meanings of 'contribution' and of 'vulnerability'.

3 The Vulnerability Objection

As Vanderschraaf points out, theories of JMA usually face what is commonly called *The Vulnerability Objection* (see for example Nussbaum, 2006). The objection goes like this: theories of JMA will fail to recognize that vulnerable people should receive some share of the resources too, and that a theory that ignores the needs of vulnerable people and only distributes resources to those who can produce resources fails to capture our most basic intuitions about justice. But it cannot be right that vulnerable people have no claim on resources simply because they are vulnerable. If the theory implies that vulnerable people are to be excluded because they are vulnerable, then the theory might be a theory of mutual advantage, but it does not bear much resemblance to what we would recognize as justice.

Some, for example Sugden (2021), argue that we do have duties to help the vulnerable, but that these are not duties of justice. However, Vanderschraaf perceives the Vulnerability objection as an objection to theories of JMA, and thus assumes that what our intuitions tell us is wrong with not sharing resources with the vulnerable is that

such failure to share is *unjust*. My interest here is in Vanderschraaf's theory, so in what follows I will assume that our intuition that we should share with the vulnerable is an intuition about justice.

Versions of the Vulnerability objection have been raised in the literature in relation to specific applications of JMA-arguments. For example, as Goodin (1988) notes, arguments that we may distribute resources unequally in a way that privileges our own citizens because citizens in a country cooperate to create a cooperative surplus that they, but not others, have a claim to, suffer from the problem that the line between those that contribute to that cooperative venture and those who do not, does not coincide with the national borders. Many people in other countries have contributed through chains of production, and therefore have a claim on a share of the resources. For the sake of our argument here, however, the more crucial point Goodin makes is that not all people within the borders of ones country have contributed. For example, small children do not contribute, and neither do people with serious health issues, including in particular those with permanent and severe disabilities.

To a large extent because of the Vulnerability objection, JMA theories have not become popular. Vanderschraaf decides to tackle the objection head on: he wants to show that the particular version of JMA that he constructs does not give rise to the Vulnerability objection, and that it therefore is more plausible than other theories of JMA.

4 Vanderschraaf's solution

As noted above, the conditions M3 and M4 above state that you receive a share of the cooperative surplus if you contribute to it and that if you fail to contribute (in the sense of deliberately withholding contributions you could have made), you shall not receive any share of that surplus. But, as Vanderschraaf correctly pointed out, that leaves open the question of whether the vulnerable—who can neither contribute nor fail to contribute (because they cannot contribute and thus also cannot deliberately withhold contributions)—should receive a share of the surplus. Should they be included among those who stand to receive benefits, or not? The Contribution requirement, which is not included among the necessary conditions Vanderschraaf lists for theories of JMA, states that *only* those that contribute/restrain themselves should receive a share of the resources. The question is then whether we should accept the Contribution requirement.

Vanderschraaf argues that the vulnerable should receive a share of the resources, and thus, that the Contribution requirement should be dropped. He develops a game-theoretic model that aims to show that rational, self-interested agents can create a system of mutual advantage that includes sharing with the vulnerable. In effect, this model assumes that only those that can contribute have to contribute. In the model, there are two categories of people: providers and recipients. Further, recipients can be either active or passive recipients. Members of the community interact with each other in an indefinitely repeated game, over discrete time periods. Sometimes the members are active, and sometimes inactive. During the time periods when an agent is active, they can play one of two roles: the role of provider, or the role of active recipient.

Which role they play in their active time periods will vary. When they are *inactive*, they are never providers nor active recipients, but always passive recipients.

When someone is a provider they produce a good, which benefits themselves and others, and which other active members can destroy if they choose to. During the rounds of the game, each agent is matched with another agent, and they are paired with different people each round. In a given round of the game, an inactive member can do nothing at all, so when they are matched with a provider, the provider can decide what portion of the good to keep for themselves, and what portion—if any—to share with the inactive member. But when a provider is matched with an active recipient, the provider has to make an offer of a share of the good to the recipient, and the recipient can either choose to accept this offer, or to reject it, in which case the good is destroyed. These two different scenarios thus correspond to a Dictator game and an Ultimatum game, respectively. The allocation of roles as provider, active and inactive recipient is random.

Because no member of the community is always a provider, but will sometimes be an inactive recipient and sometimes an active recipient, it turns out that the case in which providers always share half of the good with the recipient (whether active or inactive) Pareto-dominates both the case in which the provider is completely greedy and the case in which the provider only shares with active recipients. People thus play the role of inactive recipient often enough that they are better off getting a share of the resources even when playing that role, even if it means that they have to share when they are providers matched with inactive recipients.

Obviously, there is a risk that someone will free-ride when they play the role of provider, by not sharing with an inactive recipient. After all, the inactive recipient cannot retaliate by declining the offer and destroying the good for both of them. But this problem is solved if people punish those who behave this way. The key distinction is that between a guilty and an innocent person. A provider should always offer half of their good to the person they are matched with, if that person is innocent, but should offer none to a guilty person. A guilty person is someone who has offered less than half to an innocent person, and an innocent person is someone who has always offered half when matched with other innocent people. Further, when a person is acting as an active recipient, they should always accept any offer that an innocent provider makes, and reject any offer that a guilty provider makes—even if the offer the guilty provider makes is indeed this time very generous. This way, guilty people who play the role of recipient will be offered no share of the good in question, and when playing the role of provider, their offers will be rejected by active recipients, and the good therefore destroyed.

Vanderschraaf argues that this model shows that people might follow an equilibrium strategy in which resources are shared with those that might never be providers themselves. He writes:

> The presence of inactive members in the system mimics the facts that all humans are vulnerable at certain points in life and that some are vulnerable all their lives. (Vanderschraaf, 2019, p. 291)

In so doing, he parts ways with the traditional assumption of symmetry: he does not assume that all agents deciding on principles for mutual advantage are similar to each other in relevant respects.

5 What does the model show?

It seems perfectly sensible to expect rational people interested in principles of mutual advantage to want to take into account that they will all be inactive recipients at some point, and to make sure that they make more overall than they would if providers only share with active recipients. However, if and when it is possible to identify those who will always be inactive recipients, the reason for sharing with those people disappears. It is true that normally healthy people would be better off in the model if resources are shared with inactive people, than if resources are not shared with inactive people. But the problem is that they could be even better off if they shared resources equally with everyone, except with those that are *permanently* inactive: same size of the pie, but fewer people to share with. Most children would thus be included, as would temporarily ill or injured people, but severely disabled people would not be. Those that are unfortunate enough to become severely disabled later in life through an accident or serious illness might receive resources—that depends on whether rational, self-interested agents would agree to some sort of social insurance scheme (see for example Becker, 2005).

Vanderschraaf notes that it is critical to investigate cases in which the proportion of the time that a person is inactive, and of the time that they are matched with inactive recipients, *varies* between people. These are cases in which not all people are equally active, and in which some people are matched with inactive recipients more often than others are. He discusses the case in which the agent in question is always matched with inactive recipients when being a provider. Would the agent still be better off with a rule that required sharing with everyone, or with a rule that required sharing only with active recipients? It turns out that if the agent themselves is an inactive recipient for a proportion of time greater than 0.2929, they will still benefit from a rule that specifies that providers share equally with everyone.

This is of course an interesting case. But it is not the most crucial case for analyzing whether agents in Vanderschraaf's model would benefit from sharing equally or only with active members of the community. What matters crucially is not so much the proportion of time someone is matched with an inactive recipient, but the proportion of time someone is inactive. More specifically, what's important in answering the Vulnerability objection is what happens when we let the proportion of time that some people are inactive in Vanderschraaf's model approach 1. These are the vulnerable people: would the others share with them?

What proponents of theories of justice-as-convention (or JMA more broadly) must do in order to argue that such theories would include sharing with (all) the vulnerable, is to demonstrate that productive contributors would find it in their interest to share with those that can never, or almost never, contribute a net gain to the cooperative surplus. Vanderschraaf's model shows that given the right assumptions about for example a system of punishment of those that fail to share with innocent recipients, it could be

in the mutual interest of agents to share with the vulnerable. This is no small feat. But *why* would rational, self-interested people who can be *even better off* by excluding permanently and easily identifiable vulnerable people nevertheless choose to share with them? If rational, self-interested people who are not themselves vulnerable decide on rules for their mutual advantage, and they have two options, one that benefits them more by excluding the vulnerable and one that benefits them less by including the vulnerable, why would they choose the latter?

After all, Vanderschraaf's model showed that as long as those that can contribute find themselves in the role of inactive recipient sufficiently often, they will benefit from a system where everyone shares with the (innocent) inactive recipients over one in which providers do not share with (innocent) inactive recipients. But this does not show that they will prefer such a system that includes sharing with those that are always inactive recipients over one that excludes those that are always inactive recipients.

In Vanderschraaf's model, sharing with the vulnerable is enforced through the system of punishment against those that fail to share with the innocent. Given the right punishment system, almost any behaviour can be upheld among rational, self-interested agents—just adjust the punishment system accordingly. But Vanderschraaf does not tell us why rational, self-interested agents should choose to create and uphold a system of punishment that required them to share with the vulnerable.

One could argue that Vanderschraaf has proven that rational, self-interested and net-contributing agents must not necessarily refuse to share with the vulnerable, and that my argument concerns rather what such agents are *likely* to choose.[1] In effect, rational, self-interested agents who are able to contribute to a cooperative surplus, and who are considering whether to enter into an agreement for their mutual advantage, face a choice between two options: (a) exclude the vulnerable, and potentially gain a lot from the agreement, or (b) include the vulnerable, and gain less. Since both options involve benefits for the rational, self-interested agents who are able to contribute, one could consider them both to be appropriate assumptions for a theory of JMA. If so, my argument seems to be just that such agents are more likely to choose the option that excludes the vulnerable than the one that includes them.

But assuming that rational, self-interested agents who are aiming to enter into an agreement for their mutual advantage will choose to include those that it is not to their overall advantage to include, fits badly with the idea behind a theory of JMA. Choosing to share with the vulnerable when they would be better off excluding the vulnerable is not rationally self-interested. In order to argue that these rational egoists would choose something that is not in their rational self-interest, we would either have to ascribe some other motive to them (like altruism), or argue that contrary to what it may seem, it is in fact in the best interest of these rational egoists to share with the vulnerable. Further, sharing with inactive recipients is maintained in Vanderschraaf's model, despite the obvious temptation by rational egoists to refuse to share, by the assumption that the agents punish those that don't share with the innocent. But assuming that rational egoists would agree to punish those that do not share with the vulnerable would require said rational egoists to act in a way that is not rationally egoistic, because they would all be better off with rules that do not require any sharing with the vulnerable.

[1] I am grateful to an anonymous reviewer for raising this point.

Vanderschraaf assumes the agents to be rational and purely self-interested (in line with the tradition of theories of JMA), and he thus does not ascribe any altruistic motives to them. To show that it is in the interest of rational, self-interested people to agree to share with the vulnerable (and to punish those that don't), he therefore has to argue that sharing with the vulnerable is in the interest of these agents after all. Vanderschraaf does not provide that argument. However, others have tried to make such arguments. These suggestions are of two kinds: (a) those we have considered vulnerable do in fact contribute to the cooperative surplus but we just haven't realized that, or (b) we should re-define what it means to contribute in such a way that everyone can contribute. Let us therefore turn to those suggestions to see whether a case can be made for assuming that it is the interest of rational, self-interested people to choose to share with those usually considered vulnerable.

6 Would the contributors choose to exclude the vulnerable?

Excluding the permanently and severely disabled is of course only possible if it is possible to identify who is permanently and severely disabled rather than just temporarily or only mildly disabled. If such identification was not possible, the choice would be between sharing with all and sharing with none, and for reasons discussed above, there are very good reasons to think sharing with none would not be a preferred option (although this of course depends on the proportion of permanently and severely disabled people in the population). But in reality, it is usually rather easy to identify who belongs to the group of permanently and severely disabled people. If rational, self-interested and contributing agents choose to include the permanently and severely disabled, the reason would then be that in fact, nobody is vulnerable in the sense used above. Rather, those that are usually taken to be vulnerable (e.g., the permanently and severely disabled) do contribute, after all.

There are of course many arguments purporting to show that being disabled does not necessarily mean that one cannot be a contributor. For example, disability is context-dependent: a person is not disabled per se, but disabled in a particular society (Smith, 2001). If you cannot walk, you are disabled in a society in which the ability to climb stairs is important, but not in one in which there are lots of ramps. In a society in which high cognitive functioning is usually more important than muscle strength, problems with cognitive functioning can make a person quite severely disabled. But they would not be if their society was different, so that many jobs required physical labour, but did not require so much in terms of processing large amounts of information, finding patterns, problem-solving and decision-making in complex social settings (and that living in such a society did not require such cognitive skills for just managing a household). Whether an impairment is a disability thus depends on the context. We can choose to design our society in such a way that a particular impairment becomes a disability—or not. Would rational, self-interested agents who do not, in their current context, suffer from a disability, choose to redesign their society in such a way as to make more people able to contribute? The not so heartening reply is that that depends on whether the costs of doing so are smaller or greater than the benefits to them of

the extra contributions. A very significant redesign of how things are done in various contexts in a society is likely to be expensive.

Significant investment in rehabilitation facilities, equipment and training is another way in which some people with severe disabilities can become contributors. But again, whether such rehabilitation investments are to the advantage of those who do not need such rehabilitation in order to contribute to the cooperative surplus is an open question in a theory of justice as mutual advantage (Cudd & Eftekhari, 2018). Indeed, whether resources from the cooperative surplus should be used for rehabilitation is, according to a theory of mutual advantage, dependent on whether the benefits to those who do not themselves need rehabilitation to be contributors exceed the costs for them of funding such rehabilitation, and that is likely to vary from case to case.

Other suggestions in the literature have included the idea that permanently and severely disabled people can contribute by being trusting, thereby increasing the general trust in society, which in turn is good for everybody (Silvers & Francis, 2005). But it is doubtful whether the benefit to others of these extra, trusting people would exceed the costs—you do not need to be disabled in order to trust others or contribute to a trusting environment. Further, the permanently and severely disabled people are people like everyone else: some of them will radiate trust and goodwill all around, and others will not.

Another option is for vulnerable people to provide love and care to providers. But a theory that states that justice requires that vulnerable people receive resources only if they can make a provider feel sufficiently loved is not particularly attractive. What if the vulnerable person is not a particularly effusing and loving person? Or if the vulnerable person is caring and loving in general, but is unlucky enough to be surrounded by providers that no sane person would ever love? It seems implausible that whether you should receive a share of the resources necessary for you to stay alive depends on whether you can make somebody else feel loved enough to care about you. The requirement that the vulnerable must make providers feel sufficiently loved in order to receive a share of the resources would also introduce a rather unhealthy power dynamics in close relationships. No good comes from relationships where my life depends on making you feel loved, but you can abuse me at will without losing your share of the resources.

Love between vulnerable people and providers is sometimes instead introduced to serve as the basis of an argument that vulnerable people can have secondary moral standing because they are loved by providers. Indeed, Vanderschraaf correctly notes that this argument is problematic because it introduces partiality into the theory of justice in a way that conflicts with our common intuitions about justice. I agree, but I think the argument is also problematic for another reason, namely that we would still only include some of the vulnerable people. Most parents love their children, so most severely and permanently disabled children might have secondary moral standing through their parents' love for them. But once their parents are gone, it is not obvious that there will be providers who love them. Some will be lucky enough to have partners or friends who love them, but not all will.

Others point out that people might be disabled in some respects, but very much able to contribute to the cooperative surplus in other areas of life (Smith, 2001). Stephen Hawking, for example, was severely physically disabled but extremely talented in

physics, and certainly made great contributions in that field. We should not assume that everyone must contribute in the same way to the cooperative surplus, whether considered disabled or not. But even so, there are those who unfortunately lack the capacity to contribute much in any area. The point that nobody contributes in every way, and that a lot of people contribute in some (valuable) way, is a healthy reminder that impaired people should not all be put in the same category and that able-bodied people should not think that all disabled people are unable to contribute something of great value and advantage to others. But nevertheless, even if some who are disabled in some respects are great contributors in others, what such a strategy of justifying inclusion amounts to, is the inclusion of some more people, but not all. After all, the general claim would still be that you are included among those whom justice applies to only if others find it worthwhile to include you. Your inclusion has to benefit them more than it costs them.

Finally, we could define what it means to contribute to a cooperative surplus in such a way that contributions require very little in terms of active engagement. That would possibly allow all permanently and severely disabled people to count as contributors. One example is defining contributions as refraining from breaking the law. Sangiovanni (2007), for example, argues that what grounds the obligations of distributive justice within a state is that citizens contribute to upholding that state and its institutions through paying taxes, complying with the law, etc. Some of that argument is clearly about contributing to the cooperative surplus of resources through paying taxes that secure public goods, and this is not something some permanently and severely disabled people will be able to do, and neither will some of them be politically active and support and maintain the political system that way. But part of the argument might also be interpreted as being about complying with the law. The argument would be: 'No matter how disabled you are, you can refrain from breaking the law'. The problem is that if a person P is so severely disabled that there is not much damage they can do even if they really try, other rational, self-interested people would not agree to share resources with them in order for them to refrain from breaking laws [a point that Vanderschraaf himself makes (Vanderschraaf, 2019, p. 284)]. Whether it is worth the cost to pay off the severely disabled people with resources in order for them to refrain from breaking laws, depends on how much damage they can do if they set out to break those laws. Further, perhaps the cooperative surplus should be defined not in terms of extra resources, but only in terms of the peace that results when everyone complies with the laws. What a contributor can expect, is that because she doesn't behave criminally to others, they will not behave criminally to her. But for those unable to work, it is not enough that others do not assault them or steal from them. They need resources, and they cannot get those resources on their own. As long as we take a theory of justice to concern, among other things, the distribution of resources, this strategy will thus not help.

Another example along similar lines is Sisson's and DeNicolo's (2014) suggestion to define contribution as leaving others to pursue what gain they can achieve by their own potential in peace. They explicitly note that their suggestion does not mean that anyone is entitled to a share of resources, and this, of course, can be very problematic for severely disabled people who need such a share in order to survive. (Sisson and DeNicolo then go on to develop other arguments for why severely disabled people

should still receive some, although not an equal, share, but these reasons do not seem to stem directly from the JMA-account.)

There thus does not seem to be a strong argument for why rational, self-interested agents who can contribute to a cooperative surplus would choose to share resources with the permanently vulnerable. We can acknowledge that people contribute in different ways, but some people will always be contributing less than what their share of the surplus would be, and therefore including them would not overall be beneficial to others. Indeed, the relevant distinction is not really between those that can contribute, and those that cannot contribute at all. Rather, it is between those whose contributions are large enough that the absolute size of the pie share per person increases, on the one hand, and those whose contributions are so small that the absolute size of others' shares decreases. It all comes down to a rather unappealing marginal cost-and-benefit-analysis. Alternatively, we can re-define contributions in such a way that the permanently vulnerable can contribute after all, but unless the permanently vulnerable's contributions are valuable enough to others to motivate sharing resources with them, the vulnerable will not get a share of those resources. Or we can make sure the permanently vulnerable are included by re-defining contributions in such a way that they can contribute and by stating that contributors do not receive any actual resources. That way, providers have no reason to exclude them, but on the other hand, the permanently vulnerable do not get the resources they need. None of these solutions is great as a response to the problem that rational, self-interested agents lack reason to share resources with the permanently vulnerable, and thus we lack an account of why such agents would choose to drop the Contribution requirement.

We can of course choose to drop the Contribution requirement regardless. But that would affect the nature of Vandershraaf's theory. In what follows, I discuss these implications.

7 Does dropping the contribution requirement mean importing normative assumptions?

The Contribution requirement is not an arbitrary assumption that we can drop at will, rather, it is intimately connected with the notion of JMA, and that if we drop the Contribution requirement, we need to do so on the basis of an argument that is in line with the ideas of rationality, self-interest, and mutual advantage. Otherwise, the resulting theory is not really a theory of JMA. But as we have seen, including those that cannot contribute does not advantage those that can and do contribute (see also Sugden, 2021). The pie stays the same size, but now it is to be shared among more people. Those who can and do contribute would be better off if they shared the pie only among themselves.

Thus, we cannot motivate why rational, self-interested agents (who are not themselves permanently vulnerable) would find it in their interest to share the surplus they create with the permanently vulnerable. We could still decide to drop the Contribution requirement from our theory of JMA because including it has unpalatable results. But if we do, we must change something about that initial bargaining situation in order for rational, self-interested people to reach a conclusion that requires sharing resources

with people who do not actually contribute to the cooperative surplus. That change will be normatively motivated. Vanderschraaf is, however, in general not in favour of using normatively motivated assumptions as the basis for a theory of JMA.

His view becomes clear in his discussion of the problem of too many possible equilibria. Many theories based on game theory suffer from the problem that there are too many possible equilibria, and the set needs to be narrowed down somehow. Vanderschraaf criticizes others (like Binmore and Gauthier, see Vanderschraaf, 2019, pp. 309–311) for making normatively motivated assumptions about the bargaining situation, thereby incorporating elements of what he calls a 'different theory of justice' into their theories of JMA. For example, he writes that:

> Despite Gauthier's dogged efforts to show otherwise, his baseline contains at the outset too much of the moral structure he aspires to develop as the output of the final agreement. If Gauthier sticks to assuming that parties will follow the agreement only on account of mutual advantage, then he should not be entitled to suppose that these parties would be willing to begin from a baseline where their actions are already significantly constrained by property rights. Such property rights really should be regarded as part, and a very important part at that, of the cooperative surplus. Recognizing that Gauthier would resist my opinion I believe that Gauthier has effectively incorporated into his theory fundamental natural rights the way Locke conceives of such rights, so that his is no longer a purely justice as mutual advantage theory. (Vanderschraaf, 2019, p. 211).

He follows up by stating:

Plainly, the justice as mutual advantage theorist needs criteria that on the one hand do real work in identifying acceptable sets of equilibria and on the other hand do not smuggle in elements from some fundamentally different theory of justice. (Vanderschraaf, 2019, p. 311)

In order to narrow down the number of possible equilibria, Vanderschraaf instead explicitly chooses a condition—the Baseline Consistency condition—that is taken to be technical rather than normative, so as to avoid importing normative considerations into the decision process. (This condition states that only those agreements are allowed that would not need to be renegotiated when the size of the cooperative surplus changed. However, it is a matter of disagreement whether such a condition is normatively neutral—theories of justice that are based on a notion of thresholds, for example stating that questions of justice only arise when people are below a certain level of welfare, conflict with the Baseline condition).

The problem is that by replacing the Contribution requirement with the weaker requirement that only requires agents to contribute if they can, Vanderschraaf imports a normative consideration into the very set-up of the theory. Further, that normative consideration is strongly associated with other theories of justice, but is alien to theories of JMA because it in effect requires rational, self-interested agents to ignore that it is not in their interest to share the cooperative surplus with the permanently vulnerable. Vanderschraaf claims that Gauthier's theory is not a pure theory of JMA because it incorporates private property rights, but the same problem thus applies to Vanderschraaf's theory: it, too, incorporates a normative assumption into the set-up,

and that normative assumption is not really in line with the ideas of rationality, self-interest and mutual advantage. Vanderschraaf thus both distances himself from those who construct their initial bargaining situations on the basis of explicit normative considerations, and joins their group. It is no longer clear just what kind of project it is he takes himself to be doing.

A first possibility for a theory of mutual advantage is to hold that whatever results the decision process generates are just: whatever rational, self-interested agents would agree to constitutes justice. This position is sometimes referred to as 'the Contractarian view'. But this does not seem to be the position Vanderschraaf holds, since he explicitly claims to be using a reflective equilibrium method, and drops the Contribution requirement in order to arrive at results that fit better with an independent standard of justice. This move is of course in line with the way Rawls used the reflective equilibrium method, but then again, Rawls was explicitly engaged in a normative project.

A second possibility is that the theory aims to show that rational, self-interested agents could agree to just principles, that is, that justice could arise out of the negotiations among such agents. This seems to be Vanderschraaf's position, although his use of a reflective equilibrium suggests that he is also willing to adjust his intuitions about justice a bit in the light of the outcome of the bargaining process. But this possible aim of the theory is undermined by the argument that rational, self-interested agents, who are capable of playing the role of provider most of the time, would not find it in their interest to share with those that cannot contribute much, and would thus not be motivated to drop the Contribution principle.

A third possibility is that the theory should be understood as explaining why it is that we have the rules for sharing that we do, or perhaps why we have the intuitions about sharing that we do. Vanderschraaf himself notes (2021) that his project is predominantly explanatory. As he does not elaborate on this point, it is unclear how such an explanatory project relates to the second possibility above. But it does seem to me that the explanatory route might be promising. If we go sufficiently long back in history to be able to speak about evolutionary tendencies, it is reasonably easy to explain why we generally hold that we should share with the vulnerable. We have evolved to have the kind of intuitions about justice and tendency for sharing that we have, because overall, it is to our mutual advantage to share resources. From an evolutionary perspective, distinguishing between those that are permanently vulnerable, those that are disabled in some way but can contribute significantly in another way and those that are temporarily ill or injured is not worth the costs of attention to clues, keeping track etc. On this interpretation, the fact that some people cannot contribute enough to cover the costs they generate is not a problem.

A fourth option is to bite the bullet and accept that the principles for sharing that results from a mutual advantage bargaining situation will not require equal sharing, or anything like sharing enough to keep everybody alive. Vanderschraaf himself explicitly states that he has shown that rational, self-interested agents can agree to share equally, not that they will agree to it. His, like so many other game theoretic analyses of various phenomena, suffers from the problem of multiple equilibria. His requirement of Baseline consistency narrows down the set of possible equilibria, and he shows that given his assumptions, the resulting set will include the equal sharing-equilibrium. But the set also includes other equilibria. If he were a contractarian, and thus held that any

outcome resulting from the specified bargaining process must be just by definition, all of these other equilibria must be considered just too, despite not involving sharing equally. It would be interesting to know a bit more about these other equilibria, and in particular to know more about just how unequal they are. Does Vanderschraaf hold that these are just too? Or would he be prepared to subject them to consideration along the lines of a reflective equilibrium, and change his assumptions further in order to generate a result that is more in line with our considered judgments about justice? The answer would tell us more about the kind of project Vanderschraaf takes himself to be engaged in, and thus just how problematic the Vulnerability Objection is for it.

8 Conclusion

Vanderschraaf's project is impressive. I think he has achieved something important, but we probably disagree about what it is that he has achieved. If all it takes to qualify for a share of the cooperative surplus is that you contribute if you can, then the permanently vulnerable can indeed be included among those that stand to gain a share of the cooperative surplus as a matter of justice. But then, the underlying idea seems to be that everyone should receive a share of the resources as a matter of justice, unless they actually refuse to be part of the project. In that theory construction, considerations of mutual advantage plays no, or at least a quite small, role. I think this theory is interesting and in many ways promising.

However, although at times it seems Vanderschraaf is perfectly happy with adjusting the decision process so that it generates a more normatively appealing result, at other times, he seems to criticize others for having included normatively motivated assumptions in their versions of bargaining between rational, self-interested agents. Further, he takes great care to narrow down the set of possible equilibria of his model by using a purportedly non-normative condition (Baseline Consistency). It is thus not quite clear to me what Vanderschraaf thinks about basing his theory on normatively motivated assumptions in order to generate a normatively appealing result. It is also not quite clear whether his project is a normative one, aimed at specifying a theory of justice we should embrace, or whether it is an explanatory project, aimed at explaining what kind of agreement rational egoists would agree to under such-and-such conditions.

But even without dropping the Contribution requirement, his model constitutes a partial answer to the Vulnerability Objection. Theories of justice as mutual advantage are often based on the assumption that all agents are quite similar to each other. But in his model, Vanderschraaf explicitly allows that people are active contributors to varying extent, and that some will be contributors more than others. He has then shown that it can be in the interest of rational egoists to share, even despite the fact that people's degree of contribution varies. But he has not satisfactorily solved the hardest part of the Vulnerability Objection, that is, why rational egoists would agree to share resources with those that never, or almost never, contribute more than what they withdraw from the cooperative surplus.

Acknowledgements I am grateful to Göran Duus-Otterström and Lars Eriksson for helpful comments on drafts of this paper.

Funding Open access funding provided by University of Gothenburg. Not applicable.

Data availability Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest Not applicable.

References

Becker, L. C. (2005). Reciprocity, justice, and disability. *Ethics, 116*(1), 9–39.

Cudd, A. & Eftekhari, S. (2018). Contractarianism. In E. N. Zalta (Eds.), *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition). Retrieved on June 20, 2021, from https://plato.stanford.edu/archives/sum2018/entries/contractarianism/.

Goodin, R. (1988). What is so special about our fellow countrymen? *Ethics, 98*(4), 663–686.

Nussbaum, M. C. (2006). *Frontiers of justice: Disability, nationality, species membership*. The Belknap Press of Harvard University Press.

Sangiovanni, A. (2007). Global justice, reciprocity, and the state. *Philosophy & Public Affairs, 35*(1), 3–39.

Silvers, A., & Francis, L. P. (2005). Justice through trust: Disability and the "outlier problem" in social contract theory. *Ethics, 116*(1), 40–77.

Sisson, M., & DeNicolo, M. (2014). Minimal mutual advantage: How the social contract can do justice to the disabled. *European Journal of Political Theory, 14*(2), 161–179.

Smith, S. (2001). The social construction of talent: A defence of justice as reciprocity. *Journal of Political Philosophy, 9*(1), 19–37.

Sugden, R. (2021). Hume's theory of justice and Vanderschraaf's vulnerablity objection. *Philosophical Studies, 178*, 1719–1729.

Vanderschraaf, P. (2019). *Strategic justice—Convention and problems of balancing divergent interests*. Oxford University Press.

Vanderschraaf, P. (2021). Reply to critics. *Philosophical Studies, 178*, 1741–1756.