7-6-2024

Strategic Justice, Conventions, and Game Theory: Introduction to a *Synthese* Topical Collection

Michael Moehler
Virginia Tech

John Thrasher
Chapman University, thrasheriv@chapman.edu

Recommended Citation

Moehler, M., Thrasher, J. Strategic justice, conventions, and game theory: introduction to a *Synthese* topical collection. *Synthese* 204, 28 (2024). https://doi.org/10.1007/s11229-024-04680-2

Strategic Justice, Conventions, and Game Theory: Introduction to a *Synthese* Topical Collection

Synthese (2024) 204:28
https://doi.org/10.1007/s11229-024-04680-2

ORIGINAL RESEARCH

Strategic justice, conventions, and game theory: introduction to a *Synthese* topical collection

Michael Moehler[1] · **John Thrasher**[2]

Received: 2 June 2024 / Accepted: 14 June 2024 / Published online: 6 July 2024

Abstract
Evolutionary, game-theoretic approaches to justice and the social contract have become increasingly popular in contemporary moral and political philosophy. (Vanderschraaf, Strategic justice: convention and problems of balancing divergent interests, Oxford University Press, 2019) theory of strategic justice represents the most recent contribution to this tradition and, in many ways, can be viewed as a culmination of it. This article discusses some of the central features of Vanderschraaf's theory and relates them to the contributions in this collection. Some of the contributions directly address Vanderschraaf's work, while others explore related topics in game theory, bargaining theory, formal philosophy, rationality, equality, justice, and the theory of conventions. This collection aims to bridge a gap between disjoint but closely related literature spanning a wide range of disciplines. The contributions allow readers to systematically engage with the topic of strategic justice, advance dialogue, and more easily follow this rich and expanding field of study.

Keywords Convention · Justice · Vanderschraaf · Game theory

1 Conventionalism

Evolutionary, game-theoretic approaches to justice and the social contract have been increasingly popular in contemporary moral and political philosophy. They have been defended by Robert Sugden (1986), Ken Binmore (1994, 1998, 2005), Brian Skyrms (1996, 2004), Jason Alexander (2007), and Cailin O'Connor (2019), to mention just

✉ John Thrasher
thrasheriv@chapman.edu

Michael Moehler
moehler@vt.edu

[1] Philosophy, Politics, and Economics, Virginia Tech, Blacksburg, USA

[2] Philosophy Department / Smith Institute for Political Economy & Philosophy, Chapman University, Orange, USA

a few. Peter Vanderschraaf's theory of conventional justice, developed in *Strategic Justice: Convention and Problems of Balancing Divergent Interests* (2019), represents the most recent contribution to this tradition and, in many ways, can be viewed as its culmination.

In modern philosophy, conventionalism originates with Hume's (1739/1740) moral and political theory, and many contemporary accounts of conventionalism share some of its core features.[1] According to Hume, human beings are morally sensible by nature despite being guided by instrumental rationality and self-interest. They possess 'natural virtues' rooted in the weak sentiment of benevolence that makes human beings sensitive to the needs of others. This sensitivity alone, however, is insufficient for establishing a well-functioning society. To this end, 'artificial virtues' manifest themselves in agents' dispositions to comply with conventions of justice, particularly private property rights, to regulate social interaction.

According to Hume, for two reasons, conventions are not the result of contractual promises. First, Hume argues that promising itself is a kind of convention, and second, on historical grounds, Hume (1742) rejects the idea that actual agreement among agents on a social contract has ever occurred. Instead, according to Hume, conventions evolve over time from a combination of self-interest and the understanding that, under moderate scarcity of resources and rough natural equality, reciprocal behavior is likely to be mutually beneficial. As Hume (1739/1740: Book 3, Part 2, Sect. 2) puts it, "two men, who pull the oars of a boat, do it by an agreement or convention, tho' they have never given promises to each other." Conventions serve as coordination devices that allow agents to gradually leave the state of nature to the extent that they learn to trust each other and develop mutual expectations.

The emergence of trust among agents solves the assurance problem and renders Hobbes's (1651) absolute sovereign redundant for establishing social order and political authority. Nevertheless, once society is established, the problem of compliance arises, necessitating an explanation and justification for the continued adherence of agents to conventions. The problem of compliance arises especially because existing conventions are the product of ongoing coordination upon which individual agents typically have unequal influence, if they have any influence at all, particularly for past interactions.

As such, the existing conventions may not always be in the interest of all current members of society, even if all current members of society consider some system of conventions to be better than none. Although the current system of conventions may be strictly Pareto-superior to the state of nature, some members of society may prefer other feasible systems of conventions that would allow them to benefit more than the current system, especially if historical injustices occurred. This feature of strategic justice may create instabilities that must be addressed to maintain social order and peaceful interaction.

[1] For a discussion of Hume's moral and political theory, see Moehler (2018, 2020). Conventionalism has a long tradition in pre-modern philosophy, especially in work influenced by Epicurus and his Roman popularizer Lucretius. In many ways, Hume and others in the early modern period are rediscovering Epicurean thought. Epicurus is also likely the first philosopher to develop a conventionalist mutual-advantage theory of justice (Thrasher 2013).

Hume (1739/1740, Book 3, Part 1, Sect. 2) explains the continued adherence of agents to existing conventions by stressing that agents typically will realize that their private good is intimately linked with the public good, especially expressed in the form of a functioning social order. Moreover, Hume (1739/1740, Book 3, Part 3, Sect. 1) maintains that agents' "sympathy" with each other motivates compliance with established conventions.[2] Finally, Hume argues that over time, agents will start to value the existing conventions intrinsically. They will follow the conventions not merely for instrumental reasons but will internalize their demands by developing a moral sense that corresponds to and approves of the conventions. To solve the is-ought problem, Hume (1739/1740, Book 3, Part 3, Sect. 6) assumes that agents' moral sense will approve of itself.

In game-theoretic terms, Hume argues that agents will develop commitment power that binds them to follow the established conventions, even if doing so does not benefit them in each instance. This internalization process allows Hume to combine justice with self-interest and, in theory, solve the problem of compliance. Suppose sufficiently many or all members of society develop a moral sense that approves the demands of the existing conventions and constrains agents' behavior. In that case, adhering to the conventions is most beneficial for agents in the short and long term. In theory, Humean conventions are self-enforcing. In practice, however, Hume is aware that agents often will be shortsighted. As such, he argues that conventions must be enforced institutionally to uphold justice and social order to benefit all members of society.

2 Vanderschraaf's theory of justice

Following these broad features of Humean conventionalism and building upon his previous work, Vanderschraaf (2019) presents a systematic analysis of conventionalism and its notion of strategic justice, combining rigorous game-theoretic analysis, innovative use of (social) scientific methods and normative analysis in the context of the social contract. In his book, Vanderschraaf develops a new theory of justice (justice as convention) that, despite a mutual advantage approach, considers the most vulnerable members of society and defends an egalitarian bargaining solution as a principle of justice.

The central claim of Vanderschraaf's book is that justice is conventional, and conventions can be understood precisely and game-theoretically. Vanderschraaf (2019, p. xi) notes that the thesis that justice is conventional is not new, and "versions of this thesis were proposed by some of Plato's Sophist predecessors." As mentioned in the previous section, Hume and others took up the idea in the early modern period, and the notion of convention was further developed by twentieth century philosophers and economists, most notably by Thomas Schelling (1960) and David Lewis (1969).

Vanderschraaf's main contribution begins with questions regarding the relationship between justice and conventions. These questions include puzzles about how to characterize conventions, the background conditions of justice, and how *justice as*

[2] Hume's friend Adam Smith also has a similar account of the origin and stability of conventions and norms, though his account uses sympathy in in a way that makes his interestingly different from Hume's (Hankins and Thrasher 2022).

convention relates to the larger idea of *justice as mutual advantage*. Each of these considerations entails several sub-questions, including how to model anarchy dynamically, apply bargaining theory to justice, and protect the vulnerable in a resolutely conventionalist theory of justice. Although Vanderschraaf's book is far too ambitious in scope and dense in argumentation to précis in a way that does justice to it, in this section, we will provide a brief overview of Vanderschraaf's main innovations.

As Vanderschraaf (2019, p. xii) notes in the preface, most philosophers throughout history have thought that justice as convention was "plainly wrongheaded." He admits that he initially agreed with this assessment. Nevertheless, justice as convention has remained a "persistent irritant" throughout the history of moral and political philosophy. To explain the persistent appeal of this tradition, Vanderschraaf (ibid.) notes that any theory of justice must answer what he calls the *content* and the *motivation* question. The content question concerns the substance of justice: what does justice demand, and how can agents know these demands? The motivation question concerns the reasons for agents to follow the demands of justice once understood.

The staying power of justice as convention, as Vanderschraaf notes, is that it has a clear answer to the motivation question and a method for answering the content question that is unavailable to most other theories. However, the cost of this approach, as Vanderschraaf (ibid.) expresses it, is to risk "making justice somewhat less exalted than some think it should be." The key to generating a plausible, and at least somewhat exalted, theory of justice as convention requires using the notion of convention to solve the motivation problem and characterizing convention in such a manner that it also solves the content question in a way that is recognizably a theory of justice.

For starters, a conventional theory of justice must rely on conventions. However, not just any conventions will suffice in the context of justice. This is a thorny problem because the conventionalist must steer between the Scylla of using a non-conventional notion of justice to narrow the scope of acceptable conventions and the Charybdis of indeterminacy. Vanderschraaf tacks carefully and with high precision through these narrow and dangerous waters.

Vanderschraaf's approach begins with a definition of convention. Conventions are arbitrary in that they could have been otherwise than they are—we could drive on the left rather than the right or speak French rather than English. Vanderschraaf notices that two distinct senses of arbitrariness apply in the context of conventions: the indifference sense and the discretionary sense. Solutions to problems of pure coordination tend to be indifference-sense arbitrary, while solutions to problems of conflictual coordination, which concern justice, tend to be discretionary-sense arbitrary. Indifference arbitrariness suggests that the parties involved are indifferent between the equilibrium solutions of the game. In contrast, discretionary arbitrariness means that the parties are not indifferent between the relevant conventional equilibria.[3] All coordination games will have discretionary arbitrary solutions, but not all such games will have indifference arbitrary solutions. That is, although all conventions are arbitrary, it is not true

[3] It is perhaps natural to think that indifference arbitrariness is a property of solutions to pure coordination games, while discretionary sense arbitrariness is a property of solutions to impure coordination games, but this is false. As Vanderschaaf (2019) shows in Chapter 2 of his book, impure coordination games will have indifference and discretionary sense solutions when correlated equilibria and mixed strategies are included.

that agents "never care which of the available conventions they follow in the end" or that conventions are "orthogonal to justice," as Vanderschraaf (2019, p. 67) stresses.

One of Vanderschraaf's core innovations is to link conventions to the notion of correlated equilibrium. In the work of David Lewis (1969) and Christina Bicchieri (2005), conventions (or norms) are Nash equilibria of the underlying games. By contrast, Vanderschraaf (2019, p. 69) argues that conventions may, "in fact, regulate situations where none of the corresponding strict Nash equilibria are coordination equilibria." He argues that conventions should be understood as a "correlated equilibrium," a supergame of the original base game wherein agents coordinate their strategies based on some external mechanism.

This reasoning relates to Vanderschraaf's (2019, p. 71) use of an "equilibrium-in-conjectures" approach to understanding why agents might employ mixed strategies. Each agent's Nash mixed strategies are probabilistically independent of one another and drawn solely from their common knowledge of the game structure and payoffs. However, if one is concerned with conflictual "coordination games," as Vanderschraaf (2019, p. 33) suggests in the context of justice, which rely not only on predictions of future behavior but also on evidence from past behavior (e.g., involving reciprocity), a wider conception of conventions is required that relies on correlated, prospective, and retrospective assessments by the players.

For Vanderschraaf, in the context of a game, an equilibrium is a convention if it is (i) a correlated equilibrium of the game, (ii) there is more than one such correlated equilibrium (discretionary sense arbitrariness), and (iii) both of these characteristics are common knowledge to the agents.[4] As Vanderschraaf (2019, p. 82) stresses, with "this definition it is possible for a convention to be characterized by equilibria of indefinitely repeated games even where the profile of actions this convention prescribes in a given interaction period is not an equilibrium of the base game."

Although there is much more important detail in Vanderschraaf's construal of convention, it should be clear that Vanderschraaf's move to a conjectural, epistemic justification and model of the rationality of agents, along with his embrace of correlated equilibrium as the solution concept relevant for conventions, is a substantial advance over other accounts of conventions. Its chief advantage is its generality and precision. As Vanderschraaf (2019, p. 84) writes, "[t]he game-theoretic definition of convention given [in this book] is designed to capture all of the possible conflictual coordination conventions." It is "…designed for analyzing the conventions of justice" (ibid.).

However, to understand the background context of justice and determine actual conventions of justice, more information is needed about what Hume (1739/40) called the "circumstances of justice." One can think of the circumstances of justice as Vanderschraaf's analog of the state of nature of classical social contract theory. Vanderschraaf shows that previous models of the state of nature, in particular, Rawls's (1971) structural conditions in the "original position" and Hume's model of the circumstances of justice, all have limitations. Consequently, Vanderschraaf (2019, p. 116) proposes an alternative called the "Generic Circumstances of Justice." These circumstances determine that:

[4] This brief characterization considerably simplifies Vanderschraaf's (2019) presentation.

> parties have the right background conditions for justice when (i) they have available to them a variety of conventions over which their preferences differ to some extent, (ii) they can by working together generate a cooperative surplus characterized by some of these conventions, but (iii) each is also vulnerable to being taken advantage of by others who aim for outcomes better for themselves that result in their fellow parties suffering relative losses.

According to Vanderschraaf (2019, p. 116), these conditions are important because they "effectively set certain formal constraints upon norms of justice." They serve as the formal constraints on the possible conventions that can make up the substance of justice. However, even with these formal constraints, justice still has considerable indeterminacy. Not only must the norms of justice be conventions consistent with the circumstances of justice, but they also need to be seen as just by the members of society.

In this context, as Vanderschraaf (2019) does in Chapter 4 of his book, one might ask whether justice, as a formal set of norms, is essential to the state of nature. As Hobbes (1651) did, should one assume that the state of nature will inevitably devolve into a state of war? If one accepts Vanderschraaf's assumptions in this part of his book, his sophisticated analysis of dynamic anarchy shows that Hobbes was right. Peace is too fragile in the state of nature and war is the natural state of affairs. The way out for Vanderschraaf, as for Hobbes and Locke, is to establish norms of justice.

To this end, Vanderschraaf considers fair division in a bargaining problem as a canonical equilibrium selection problem and model of justice. Here, Vanderschraaf follows contemporary contractarian thinkers like Rawls (1958, 1971), Gauthier (1986), Gaus (1990), Skyrms (1996, 2004), Binmore (1994, 2005), Moehler (2018), Muldoon (2016), and Bruner (2015, 2020), among others.

Following Skyrms, Vanderschraaf includes both evolutionary and learning dynamics in his bargaining model. In so doing, he reaches the conclusion that the egalitarian bargaining solution favored by Braithwaite (1955) and Raiffa (1953) is the most likely to emerge. Controversially, this requires some account of the possibility of interpersonal utility comparisons and the acceptance of the "Baseline Consistency criterion" that, according to Vanderschraaf (2019, p. 312), allows agents to make "seamless" transitions if the cooperative surplus contracts or expands over time.

If one can follow Vanderschraaf down this road, one arrives at an interesting convergence of Aristotle, natural law, Hume, and game theory. Vanderschraaf (2019, p. 188) writes:

> Rejecting positive offers perceived as too lopsided in Ultimatum games and punishing defectors at a personal cost in public good games are explainable as products of an evolved tendency in our species to treat others fairly and to punish those who fail to do far beyond laboratory settings. A general requirement to "play fair" is deeply rooted in the natural law tradition from antiquity. A number of the great figures in both the classical and the modern natural law traditions maintain that requirements of the natural law follow from some version of the Golden Rule. These natural law requirements correspond to the principle "Do unto others as you would have them do unto you," since they require one to act for the benefit of others, possibly at some personal cost.

According to Vanderschraaf, fairness in terms of roughly equal division is the baseline norm of justice that is most likely to develop and survive over time. If this demonstration is successful, Vanderschraaf has, as he suggests, found a way to merge the seemingly opposed traditions of natural law and conventionalism.

Even though (at least some of) the norms of justice are what one might call "natural conventions," the Hobbesian concern about anarchy persists. Vanderschraaf agrees with Hume and Hobbes that a government is needed to establish and secure justice. In Chapter 6 of his book, Vanderschraaf (2019) develops a complex argument intended to show that there is a Humean, conventional solution to the Hobbesian problem of selecting a sovereign and establishing a government. Vanderschraaf argues that a Humean governing convention between the rulers and the ruled can be stable over time and that there is good reason (though not decisive) for the ruled to select a democratic over a non-democratic sovereign.

This chapter is one of the richest in the book and resists simplification. While many may find it too abstract to provide an account of what actual political systems should look like, Vanderschraaf's attempt to solve some of the thorniest problems in political theory while using a conventionalist method gets high marks for its degree of sophistication. As with the previous accounts of Hobbes and Hume, from which he draws, many readers will likely object to some of the details and/or characterizations he provides. Nevertheless, as with Vanderschraaf's predecessors, there is much to be learned from what is here, whatever one's disagreements.

Up to this point, Vanderschraaf has argued that justice is a convention characterized by fairness and amenable to stabilization by a conventional government. In so doing, he has constructed and defended a specific version of what Brian Barry (1989) called a "justice as mutual advantage" theory. In the final two chapters of his book, Vanderschraaf defends justice as mutual advantage, arguing that it is a legitimate theory of justice and that complying with the demands of such a theory is compatible with rationality. The great advantage of Vanderschraaf's conventionalist mutual advantage theory is that it can easily answer the content and motivation questions about justice. Vanderschraaf (2019, p. 321) argues that the content of justice is defined by "[c]onventions that share out the benefits and the burdens of life in society," and agents are motivated to follow those conventions because doing so serves their "own interests" on condition that they expect "others to obey as well."

In the book's concluding chapter, Vanderschraaf addresses several challenges to mutual advantage theories. One that might seem to afflict his own theory is the "indeterminacy problem," namely, that there may be too many possible conventions of justice to justify any particular one. As Vanderschraaf notes, however, the solutions to this problem used by Rawls (1971), Gauthier (1986), Binmore (1994, 1998, 2005), and other contractarian thinkers rely on incorporating elements of non-mutual-advantage theories, which weakens the power and plausibility of these theories.

3 Contributions

This collection brings together philosophers, economists, political scientists, and game theorists who engage with themes from Vanderschraaf's book and his work more generally. Some of the contributions directly address Vanderschraaf's work, while others explore related topics in game theory, bargaining theory, formal philosophy, rationality, equality, justice, and the theory of conventions. Beyond addressing Vanderschraaf's work and its many facets, this collection aims to bridge a gap between disjoint but closely related literatures that span a wide range of disciplines. The contributions allow readers to systematically engage with the topic of strategic justice, advance dialogue, and more easily follow this rich and expanding field of study.

Brian Skyrms engages with the concept of "correlated equilibrium," which, as mentioned in the previous section, is central to Vanderschraaf's view of conventions. Skyrms discusses a generalization of Vanderschraaf's concept of correlated conventions that he terms "Quasi-Conventions." He introduces the idea of "coarse correlated equilibria" and explores how such equilibria can lead to improved payoffs. In doing so, he addresses the question of learnability through simple uncoupled learning dynamics, surveying laboratory experiments to support his argument. The generalization proposed by Skyrms introduces the notion of "strains of commitment," which he suggests can be viewed from different perspectives. Skyrms concludes that while the strains of commitment may prevent the generalization from serving as a stand-alone definition of convention, Quasi-Conventions can still be important modules within larger, true conventions in certain settings.

Chad Van Schoelandt, building closely on Vanderschraaf's work, engages with a general criticism of mutual advantage approaches to justice regarding the treatment of the vulnerable. Critics often argue that mutual advantage approaches to justice cannot adequately protect the vulnerable. Van Schoelandt shows that this "vulnerability objection" can be answered in principle, similar to the line of reasoning defended by Vanderschraaf. Van Schoelandt suggests that while it may not be guaranteed and too much to ask from a conception of justice in this tradition, it is possible for a mutual advantage approach to protect the vulnerable. Finally, Van Schoelandt emphasizes the diversity of possible vulnerabilities and suggests that the social contract tradition offers a variety of strategies for addressing them. He suggests that understanding this diversity can inform approaches to justice that incorporate the protection of the vulnerable.

Sahar Heydari Fard discusses the concept of strategic injustice, considered as a system of formal and informal rules and conventions that result in profoundly unfair outcomes for particular social groups. She identifies the necessary conditions under which such injustices occur and proposes methods for eliminating them. To this end, Heydari Fard expands upon Vanderschraaf's analysis of the circumstances of justice by incorporating "asymmetric conflictual coordination games" that represent fairness issues within a dynamic social network. Heydari Fard explains how network dynamics affect the emergence and stabilization of exploitative behavior and unfair conventions, even if attempts are made to restrain them. She argues that such unfair conventions are often resilient to uncoordinated individual behavior, suggesting that maintaining rough equality becomes a coordination problem. Finally, Heydari Fard suggests that restructuring the social relations network, akin to a social movement, is necessary to resolve such coordination problems effectively.

Mario Juarez-Garcia and *Alexander Schaefer* engage with an aspect of social contract theory closely connected with Vanderschraaf's game-theoretic analysis of Hobbes's and Hume's state of nature, extending the discussion to Rousseau. While game theory has been extensively used to support Hobbes's idea of a natural state of war, it has been largely neglected to portray Rousseau's peaceful state of nature. Juarez-Garcia and Schaefer's contribution formalizes Rousseau's critique of Hobbes. It identifies flaws in Hobbes's assumptions, such as the absence of an exit option and an unrealistic view of human nature. By integrating Rousseau's criticisms into a game-theoretic model, the contribution explores some relevant implications for Vanderschraaf's discussion of political authority.

Hannah Rubin discusses the role that honesty in communication plays in shaping conventions, in particular in support of those in need. While previous research suggests that punishment encourages cooperation within an evolutionary setting, including Vanderschraaf's discussion of such dynamics in the context of the stag hunt, Rubin defends a more nuanced view. She argues that the effectiveness of punishment depends on what behavior is being punished. Punishing those falsely claiming they need resources may hinder cooperation because the costs associated with deceit in cooperative efforts may outweigh the benefits, making cooperation less attractive. This contribution draws a differentiated picture regarding the enforcement of conventions and thus adds complexity with regard to solving the problem of compliance through deterrence in practice.

Jeppe von Platz returns to the vulnerability objection against mutual advantage approaches to justice. Von Platz directly engages with Vanderschraaf's proposed solution to the objection that justice as mutual advantage neglects those most in need of protection. However, von Platz takes a more critical stance than Van Schoelandt. To address the vulnerability objection in the context of his theory, Vanderschraaf (2019, p. 287) introduces the "Indefinitely Repeated Provider-Recipient Game," showing that, in certain scenarios, justice as mutual advantage can encompass concern for the vulnerable. Von Platz argues that this response does not fully address the problem raised by the vulnerability objection, which maintains that justice should provide equal basic concern for all regardless of vulnerability.

Lina Eriksson continues the discussion of the vulnerability objection, but she takes it in a different direction. In reference to the "Indefinitely Repeated Provider-Recipient Game," Vanderschraaf argues not only that his theory satisfactorily addresses the vulnerability objection but also that it adequately captures the objection. Eriksson disagrees, suggesting that while Vanderschraaf's theory may show why rational agents may share resources equally even if some members of society contribute more than others, the theory fails to explain why rational agents would share with those who can never contribute more than what they take. Eriksson suggests that Vanderschraaf's solution weakens the requirement for agents to actually contribute, consequently diminishing the theory's claim to be based on mutual advantage.

Kaushik Basu shifts the discussion to consider the connection between justice, conventions, and political leadership to secure a stable social order and avoid social oppression. Basu argues that conventions and leaders are essential for maintaining justice and social order. However, he suggests that these pillars can sometimes fail, presenting two new games, "Greta's Dilemma" and the "Incarceration Game," to illustrate such potential failure. By highlighting these issues, Basu's contribution stresses the need to reconsider collective behavior and design conventions that restrict the power of leaders in advance, using moral intention as a guide. Basu's view, linking the discussion to Vanderschraaf's view on political authority and revolution, aligns more closely with Hume's view of political leadership than Hobbes's demand for an absolute sovereign with unlimited and undivided power.

Chris Melenovsky engages with Vanderschraaf's introduction of the "Baseline Consistency" criterion mentioned in the previous section, which is central for Vanderschraaf's defense of the egalitarian bargaining solution. The criterion requires considering how well individuals fare under existing conventions compared to other hypothetical social conditions. Melenovsky argues that such comparisons are not feasible because different social conditions typically will lead to different preferences for individuals. Thus, there is no direct way to assess individual welfare across different social conditions. To apply the Baseline Consistency criterion, Vanderschraaf's theory would require an interpersonally valid standard for welfare comparisons, which is inconsistent with the general assumptions of the framework adopted by Vanderschraaf's theory. Abandoning the Baseline Consistency criterion leads to the problem of multiple equilibria and, thus, potentially to no agreement on conventions of justice or the justification of clearly objectionable conventions.

Michael Moehler continues the discussion of Vanderschraaf's rationale for the Baseline Consistency criterion, focusing on Vanderschraaf's defense of the egalitarian bargaining solution as a principle of justice and its normative implications. Moehler highlights that the Baseline Consistency criterion may actually conflict with central features of conventionalism as a coherent position in social contract theory. Moreover, the criterion limits the applicability of Vanderschraaf's theory of justice to societies where members de facto possess an egalitarian sense of justice, resembling Hume's proposed solution to the is-ought gap, which requires agents to approve of their own moral sense, as mentioned in the first section. Such limitation is problematic in the face of moral diversity and also affects Vanderschraaf's theory of political authority.

Despite these limitations, Moehler and probably all other authors and contributors to this collection will agree that Vanderschraaf's theory of justice represents a significant contribution to social contract theory that merits further discussion.

Acknowledgements We are grateful to all those who submitted articles to this topical collection and the anonymous reviewers for their excellent work. We would also like to thank Wiebe van der Hoek for his support throughout the editing and production process. We would also like to thank the Smith Institute for sponsoring the original conference on Peter Vanderschraaf's book, which ultimately led to this collection. This publication was made possible partly through the support of Grant 62405 from the John Templeton Foundation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation. Finally, our gratitude goes to Peter Vanderschraaf for making this collection possible through his work and intellectual engagement with us and other contributors to this collection.

Funding Open access funding provided by SCELC, Statewide California Electronic Library Consortium.

Declarations

Competing interest The authors are not aware of any conflict of interest.

References

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

Barry, B. (1989). *Theories of justice*. University of California Press.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Binmore, K. (1994). *Game theory and the social contract* (Vol. 1). MIT Press.

Binmore, K. (1998). *Game theory and the social contract* (Vol. 2). MIT Press.

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

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

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

Bruner, J. (2020). Bargaining and the dynamics of divisional norms. *Synthese, 197*(1), 407–425.

Gaus, G. (1990). *Value and justification: The foundations of liberal theory*. Cambridge University Press.

Gauthier, D. (1986). *Morals by agreement*. Clarendon Press.

Hankins, K., & Thrasher, J. (2022). Smithian Sympathy and the Emergence of Norms. *Philosophy and Phenomenological Research, 105*(3), 638–656.

Hobbes, T. (1651). *Leviathan*. Edited by Richard Tuck. 1996 ed. Cambridge University Press.

Hume, D. (1739/1740). *A treatise of human nature*. Edited by David Norton and Mary Norton. 2000 ed. Oxford University Press.

Hume, D. (1742). *Of the original contract*. Edited by Eugene Miller. Liberty Fund.

Lewis, D. (1969). *Convention: A philosophical study*. Harvard University Press.

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

Moehler, M. (2020). *Contractarianism*. Cambridge University Press.

Muldoon, R. (2016). *Social contract theory for a diverse world: Beyond tolerance*. Routledge.

O'Connor, C. (2019). *The origins of unfairness: Social categories and cultural evolution*. Oxford University Press.

Raiffa, H. (1953). Arbitration schemes for generalized two person games. In H. Kuhn & A. Tucker (Eds.), *Contributions to the theory of games II* (pp. 361–387). Princeton.

Rawls, J. (1958). Justics as fairness. *Philosophical Review, 67*(2), 164–194.

Rawls, J. (1971). *A theory of justice. 1999* (revised). Harvard University Press.

Schelling, T. (1960). *The strategy of conflict*. Harvard University Press.

Skyrms, B. (1996). *Evolution of the social contract*. Cambridge University Press.

Skyrms, B. (2004). *The stag hunt and the evolution of social structure*. Cambridge University Press.

Sugden, R. (1986). *The economics of rights, co-operation and welfare, 2004* (2nd ed.). Palgrave Macmillan.

Thrasher, J. (2013). Reconciling justice and pleasure in epicurean contractarianism. *Ethical Theory and Moral Practice, 16*(2), 423–436.

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

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.