Philos Stud (2014) 167:431–451
DOI 10.1007/s11098-013-0106-x

The scope of instrumental morality

Michael Moehler

Published online: 12 February 2013

Abstract In *The Order of Public Reason* (2011a), Gerald Gaus rejects the instrumental approach to morality as a viable account of social morality. Gaus' rejection of the instrumental approach to morality, and his own moral theory, raise important foundational questions concerning the adequate scope of instrumental morality. In this article, I address some of these questions and I argue that Gaus' rejection of the instrumental approach to morality stems primarily from a common but inadequate application of this approach. The scope of instrumental morality, and especially the scope of pure moral instrumentalism, is limited. The purely instrumental approach to morality can be applied fruitfully to moral philosophy only in situations of extreme pluralism in which moral reasoning is reduced to instrumental reasoning, because the members of a society do not share, as assumed by traditional moral theories, a consensus on moral ideals as a basis for the derivation of social moral rules, but only an end that they aim to reach. Based on this understanding, I develop a comprehensive two-level contractarian theory that integrates traditional morality with instrumental morality. I argue that this theory, if implemented, is most promising for securing mutually beneficial peaceful long-term cooperation in deeply pluralistic societies, as compared to cooperation in a non-moralized state of nature.

Keywords Evolutionary morality · Reasonableness · (Moral) pluralism · Instrumental rationality · (Pure) moral instrumentalism · Two-level contractarian theory

M. Moehler (✉)
Department of Philosophy, Virginia Tech, 229 Major Williams Hall, Blacksburg, VA 24061, USA
e-mail: moehler@vt.edu

1 Introduction

The instrumental approach to morality enjoys a long tradition in moral philosophy and is linked most closely with Hobbes' moral theory.[1] For Hobbes, morality is a means by which rational self-interested agents can leave the state of nature and cooperate peacefully with one another in society. Morality allows rational agents best to fulfill their interests in a world of natural equality, scarce resources, and conflicting ends. In contemporary moral and political philosophy, the instrumental approach to morality, considered as an independent moral approach, has been defended most prominently by David Gauthier in *Morals by Agreement* (1986).

In *The Order of Public Reason* (2011a), Gerald Gaus rejects the instrumental approach to morality as a viable account of social morality. Gaus' rejection of the instrumental approach to morality, and his own moral theory, raise important foundational questions concerning the adequate scope of instrumental morality. In this article, I address some of these questions and I argue that Gaus' rejection of the instrumental approach to morality stems primarily from a common but inadequate application of this approach. The scope of instrumental morality, and especially the scope of pure moral instrumentalism, is limited. The purely instrumental approach to morality can be applied fruitfully to moral philosophy only in situations of extreme pluralism in which moral reasoning is reduced to instrumental reasoning, because the members of a society do not share, as assumed by traditional moral theories, a consensus on moral ideals as a basis for the derivation of social moral rules, but only an end that they aim to reach.

Hobbes recognizes this feature of the instrumental approach to morality and applies the approach accordingly.[2] In Hobbes' state of nature, agents face, *de facto*, a situation of extreme pluralism due to their divergent moral and non-moral ideals and potentially conflicting interpretations of the laws of nature. In this situation of extreme pluralism, the instrumental approach to morality is the only viable account of social morality, if agents share at least one end, such as the aim of securing peaceful long-term cooperation, which demands, according to Hobbes, the institution of an absolute sovereign.[3] If, by contrast, agents share a substantial moral basis, such as a shared moral sense, as a starting point for the derivation of social moral rules, then the purely instrumental approach to morality loses its justificatory power and it cannot offer a convincing account of moral motivation, because the approach would then operate in the domain of traditional morality. The scope of pure moral instrumentalism is limited.

Gaus' criticism of the instrumental approach to morality and the basic features of his own moral theory that aims to capture the existing phenomenon of social morality show that Gaus' moral theory falls into the domain of traditional morality.

[1] See, in particular, Hobbes (1651).

[2] See Hobbes (1651, Part I, Chaps. XIII–XVI, and Part II, Chap. XVII).

[3] For a detailed, although non-orthodox, interpretation of Hobbes' moral and political theory that assumes a less extreme form of moral pluralism in Hobbes' state of nature and attributes to Hobbes a more substantial moral theory than pure moral instrumentalism, see Lloyd (1992, pp. 254–270, and 2009). For further discussion of the differences between orthodox and non-orthodox interpretations of Hobbes' moral and political theory, see Gaus (2013).

As such, it is not surprising that the purely instrumental approach to morality fails for the purpose of Gaus' project, and, for this reason, Gaus' criticism cannot generally discredit the instrumental approach to morality. Instead, if the instrumental approach to morality is applied adequately, then it can play a significant role in moral philosophy. In order to support this claim, I develop a comprehensive two-level contractarian theory that integrates traditional morality with instrumental morality. I argue that this theory, if implemented, is most promising for securing mutually beneficial peaceful long-term cooperation in deeply pluralistic societies, as compared to cooperation in a non-moralized state of nature.

The article is organized as follows. In Sect. 2, I discuss the core features of the instrumental approach to morality. In Sect. 3, I address Gaus' criticism of the purely instrumental approach to morality. In Sect. 4, I lay out the basic features of Gaus' moral theory in order to show that Gaus' moral theory falls into the domain of traditional morality and that its scope is limited to reasonably pluralistic societies. In Sect. 5, I develop a comprehensive two-level contractarian theory that combines traditional morality with instrumental morality and is valid for deeply pluralistic societies. In Sect. 6, I conclude with some methodological considerations concerning the adequate scope and normative authority of pure instrumental morality.

2 The instrumental approach to morality

Although there is no single instrumental approach to morality, but rather a variety of such approaches, instrumental approaches to morality share certain core features. Most significantly, they typically rely on instrumental rationality as a core concept of reasoning. Instrumental rationality assumes, in the relevant context, that agents are guided by hypothetical imperatives of the form: 'If, and only if, an agent believes that x is the best means for achieving y, and an agent aims to achieve y after considering all costs and benefits of x, then an agent should do x'.[4] The precise normative and epistemic demands that are imposed on agents' formation of interests, such as consistency requirements and consideration of empirical facts, vary among different conceptions of instrumental rationality. The more demanding these standards are, the more normatively demanding a specific conception of instrumental rationality is, because agents' formation and ordering of interests must fulfill these standards in order for agents to be considered rational.[5]

Independent of the precise normative and epistemic standards that different conceptions of instrumental rationality impose on agents, instrumental rationality demands only that agents take the means that are necessary to reach the ends that they aim to reach, all things considered. Instrumental rationality does not, as categorical imperatives do, prescribe ends independently of agents' specific goals.

[4] For a more detailed analysis of the normative demands of instrumental rationality, see Schroeder (2004).

[5] Gaus (2011a, p. 62) regards agents to be instrumentally rational if their beliefs, considered from their own epistemic perspectives, and the deliberations that lead to their actions are not 'grossly defective'.

Moreover, instrumental rationality is neutral with regard to agents' reasons for action. In motivational terms, agents' reasons for action may be self- or other-regarding, and in terms of content, they may be based on social, economic, cultural, religious, political, moral, or any other kinds of interest. Instrumental rationality assumes only that agents have interests that they aim to fulfill in their lives, independent of the precise nature of these interests and their motivations.

As a result of this goal-oriented nature of instrumental rationality, instrumentally rational agents are assumed to behave outcome-based. They aim to advance their individual prospects within their forward-looking perspectives and evaluate all actions with regard to their expected consequences. In addition, orthodox rational choice theory assumes that agents behave in an act-specific manner. Rational agents are assumed to approach each decision independently and always perform the action that produces, within their forward-looking perspectives, the highest interest satisfaction for them. Agents who are guided by such dominance and modular reasoning cannot commit themselves today to act in a certain way tomorrow if circumstances change, and such agents 'follow' rules only if rule-guided behavior is most beneficial for them in each instance for which rules prescribe certain behavior.

Instrumental rationality, as a basic form of reasoning, is essential for most human beings in a world of (moderately) scarce resources, especially in situations that concern survival. In such situations, according to Lionel Robbins, rejecting the demands of instrumental rationality amounts essentially to a 'revolt against life itself'.[6] Nevertheless, despite the importance of instrumental rationality for such pragmatic matters, the reasons that instrumental rationality should be the basis for a moral theory are not immediately clear. Three reasons are especially relevant in this regard.

First, instrumental rationality is minimally normatively demanding compared to other forms of reasoning, such as Kant's conception of pure practical reason,[7] and thus, it is intelligible to most human beings. Consequently, a moral theory that relies on instrumental rationality as an account of practical reasoning has maximal scope. As Nozick puts it, "[i]nstrumental rationality is within the intersection of all theories of rationality (and perhaps nothing else is). In this sense it is the default theory, the theory that all can take for granted, whatever else they think."[8] Second, instrumental rationality offers a convincing account of action motivation, because it assumes that agents act on the basis of their interests, whatever their interests may be. According to instrumental rationality, agents themselves, not an external authority, decide what is valuable to them.

Third, as a result of these two features, the instrumental approach to morality can model adequately the situation of deeply pluralistic societies in which not all members of society, due to their divergent moral and non-moral ideals, agree on what is morally right. In contrast to traditional moral theories, the instrumental approach to morality does not assume agreement among agents on substantive moral ideals as a basis for the derivation of social moral rules. In fact, the purely

[6] See Robbins (1935, p. 157).

[7] See Kant (1785, AK 4:389).

[8] Nozick (1993, p. 133). Cf. Gaus (2011a, p. 58).

instrumental approach to morality does not even demand that agents are moral, as traditionally conceived, because the approach is neutral with regard to agents' motivations and the content of their interests. The purely instrumental approach to morality assumes only that agents have at least one end in common that they aim to reach, independent of the nature of this end and the precise reasons for embracing it.

In order for a theory of instrumental morality to model the broadest plurality of interests, the assumptions of freedom and autonomy must be introduced.[9] In short, the theory must allow agents to have any interests that they happen to have and that they regard, on reflection, as part of their conceptions of the good life. Further, the theory must allow agents (at least hypothetically) to approve of the social moral rules that are justified by it, if the agents were asked to reflect freely on these rules and their implications based on their interests. The assumptions of freedom and autonomy formally ensure that the broadest plurality of interests is considered by a moral theory, because no interests of agents are ruled out *per se*, and the assumptions ensure the manifestation of these interests by the justified social moral rules. As such, the inclusion of the assumptions of freedom and autonomy, in the sense described, in a theory of instrumental morality does not represent a substantial moral premise that may be controversial in pluralistic societies. Instead, it is a necessary means to model adequately the plurality of interests in deeply pluralistic societies in which agents may hold any moral and non-moral ideals, and in which no member of society can claim to know the ultimate moral truth, if such a truth exists.

The core features of a theory of instrumental morality, including the assumptions of freedom and autonomy that are necessary to model the broadest plurality of interests that may arise in deeply pluralistic societies, are best expressed in the taxonomy of contemporary moral philosophy by a contractarian framework. Contractarian moral theories justify moral rules based on agreement among the members of a society, and they reflect the pragmatic nature of instrumental morality that considers social moral rules as a means to ensure peaceful long-term cooperation among the members of a society.

3 Models of rational agency

Gaus generally agrees with the described features of the instrumental approach to morality and its contractarian nature.[10] However, he argues that the purely instrumental approach to morality does not offer an adequate account of social morality because it cannot explain the emergence of social moral rules that are essential for large-scale human cooperation, in particular because the approach cannot explain the ability of agents to follow social moral rules that are necessary to maintain social cooperation. The purely instrumental approach to morality cannot explain the existing phenomenon of social morality.[11]

[9] See Moehler (2009, pp. 200–201).

[10] See Gaus (2011a, Chap. II).

[11] Ibid., p. 54.

To support this claim, Gaus discusses, in the context of the social contract, two variations of the instrumental approach to morality that rely on (i) orthodox theories of instrumental rationality or (ii) alternative theories of rational choice.[12] The former theories aim to show that rational agents may settle into rule-following behavior in repeated interactions, because such behavior may be beneficial for agents in an ongoing system of social cooperation, as compared to their situations in the state of nature. The latter theories modify core assumptions of orthodox rational choice theory in order to show that rational agents are capable of genuine rule-following, if they consider social moral rules to be beneficial for them in the long run. Gaus argues, for different reasons, that both variations of the instrumental approach to morality fail to explain the phenomenon of social morality for large-scale human cooperation. Gaus' criticism of alternative theories of rational choice is especially relevant for my argument.

As indicated, the instrumental approach to morality assumes that agents agree with social moral rules only if they consider these rules to be beneficial for them based on their interests, whatever their interests may be. As such, the instrumental approach to morality addresses convincingly the problems of moral motivation and moral authority because, according to the instrumental approach to morality, agents are assumed to act on the basis of their interests and, in this sense, they are their own sovereigns. According to Gaus, however, the instrumental approach to morality cannot convincingly solve the problem of compliance that is exemplified by Hobbes' discussion of the 'Foole' and is debated extensively in philosophical literature.[13]

In short, rational forward-looking agents consider it to be advantageous to leave the state of nature and enter a social framework that ensures mutually beneficial peaceful long-term cooperation. Nevertheless, once society is established and agents can enjoy the gains from peaceful long-term cooperation, it is most advantageous for rational agents, in the short run, to break the rules that establish society, if their behavior remains undetected or without punishment, because then agents can secure the benefits from peaceful long-term cooperation without paying (all of) its costs. More precisely, orthodox rationality *demands* that agents defect in the short run whenever the costs for doing so are lower than the expected gains from such non-cooperative behavior. In other words, orthodox rational choice theory assumes, as a result of its postulates of dominance and modular reasoning, that agents are opportunistic case-by-case decision makers. If, however, a significant number of agents defect in the short run, then the whole cooperative framework breaks down and the agents are, considered from their own perspectives, worse off in the long run.

It seems that orthodox rational choice theory cannot adequately solve this problem of compliance without the institution of external enforcement mechanisms that guarantee a sufficiently high apprehension and conviction rate. External enforcement mechanisms are costly, however, and thus suboptimal. In an attempt to solve this problem, Gauthier in particular, with his notion of constrained

[12] Ibid., pp. 70–100.

[13] For a recent discussion of Hobbes' Foole, see Vanderschraaf (2010, pp. 37–58).

maximization, suggests an alternative theory of rational choice in the context of the social contract.[14] Gaus, however, is skeptical of such alternative theories of rational choice. He argues that, in order for these theories to solve the problem of compliance, they must reject core assumptions of orthodox rational choice theory, in particular the assumption of modular rationality. In doing so, these theories, *de facto*, change the assumed concept of rationality. In essence, I agree with Gaus' criticism of alternative theories of rational choice. However, I argue later in this article that the instrumental approach to morality is not inherently inconsistent. Under certain conditions that define the adequate scope of pure moral instrumentalism, the instrumental approach to morality can solve the problem of compliance in the context of the social contract, without rejecting its core assumptions.

Gaus concludes that the purely instrumental approach to morality, both in its orthodox form that assumes repeated interactions among agents and alternative theories of rational choice, cannot explain the existing phenomenon of social morality and, in particular, agents' ability to follow social moral rules. This conclusion suggests that social morality, as an existing social phenomenon, cannot be purely instrumental. Agents must have, *de facto*, not merely instrumental reasons for establishing social moral rules, but also reasons that are independent of such purely means-end considerations. As a result of this conclusion, Gaus develops a model of agency that can explain the existing phenomenon of social morality and is, according to Gaus, empirically supported. In short, Gaus argues that human beings are 'rule-following punishers'.[15]

According to Gaus, rule-following punishers have the ability to internalize social moral rules that they consider to be personally beneficial in the long run and to follow such rules if sufficiently many other group members follow the rules, too. That is, rule-following punishers are not merely instrumentally rational. Instead, they value the social moral rules of their society at least partly intrinsically, and thus independently of their *immediate* causal consequences on their wellbeing. For Gaus, moral personhood "…consists in the capacity to care for moral rules in such a way that one recognizes a compelling reason to abide by the rule even when such conformity does not promote one's wants, ends, or goals…."[16] Rule-following punishers generally do not break the social moral rules of their society because doing so would conflict with their moral nature. Stated differently, rule-following punishers possess a moral sense that typically, as a result of social and cultural education, is deeply entrenched in their personalities and governs their actions based primarily on their moral emotions.

In addition, and partly as a consequence of their moral sense, rule-following punishers take, according to Gaus, an active interest in the cooperative behavior of others and invest some of their resources in punishing defectors, even if such behavior is costly for the punishers. In this sense, rule-following punishers are altruistic, because they engage in activities that not merely advance their own goals

[14] See Gauthier (1986, Chap. VI). For another alternative theory of rational choice, see McClennen (1988, pp. 95–118).

[15] See Gaus (2011a, p. 103).

[16] Ibid., p. 19.

narrowly construed, but also the common good of their society. As a result of these features, rule-following punishers are able to bridge the compliance gap between the short run and the long run that arises for merely instrumentally rational agents, and thus they are able to maintain a system of social morality that ensures mutually peaceful long-term cooperation, as compared to agents' situations in the state of nature.

Gaus argues that rule-following punishers evolve even if agents start out as purely self-interested agents. With the help of a simple evolutionary model, Gaus shows that social and cultural evolution may pressure rational agents to become rule-following punishers, because "[t]o be a member of a society dominated by Rule-following Punishers is the most effective way to advance one's ends."[17] With reference to Brian Skyrms,[18] Gaus argues that, although purely instrumentally rational agents cannot reason themselves into genuine rule following, as Gauthier suggests, evolutionary pressure can do so. Rationality must respect modular reasoning, evolution does not. If we assume a society of rule-following punishers instead of a society of purely instrumentally rational agents, then the evolution of social morality and its continuing existence can be explained without the institution of external enforcement mechanisms, because rule-following punishers have the ability to follow social moral rules genuinely, and they police themselves.

However, Gaus' model of agency is even more demanding than outlined in this section. For Gaus' moral theory, it is not sufficient that agents develop just any moral sense that corresponds to the existing social moral rules of their society. Instead, according to Gaus' moral theory, the existing social moral rules must also be rationally justifiable to all current members of society. Although social moral rules evolve on the basis of moral emotions and social practices, reason takes priority in justificatory terms, according to Gaus. As such, Gausian rule-following punishers must not only be morally sensible, but they must also share a certain Kantian rational nature, as becomes clear from discussion of Gaus' moral theory.

4 Gaus' moral theory: evolution and reason

According to Gaus, true social morality is rooted in existing moral practices that are a product of social and cultural evolution and, to this extent, social morality is responsive to the traditions and practices of particular societies. However, evolutionary processes do not necessarily lead to moral equilibria that are beneficial for all members of society compared to their situations before the establishment of social morality, or 'state of nature' for short, because some members of society or social groups may be so powerful, or lucky, that they can impose social moral rules on others that unjustifiably favor their own positions and that may worsen the situations of others compared to their situations in the state of nature. In other words, although evolutionary approaches to morality may offer an adequate genealogy of social morality, they do not necessarily lead to moral equilibria that

[17] Ibid., p. 112.

[18] See Skyrms (1996, p. 44).

are justifiable to all current members of society. For this reason, evolutionary morality must be complemented by a rational justificatory procedure, according to Gaus.[19]

For his moral theory, Gaus assumes agents who are, apart from their characteristics as rule-following punishers, Kantian in the sense that they desire to justify the social moral rules of their society rationally to other group members, because the agents conceive of themselves and others as morally free and equal persons.[20] Gaus claims that the concepts of freedom and equality, in his specific understanding, are implicit in existing social moral practices, and thus do not represent controversial moral assumptions. Gaus supports this claim by empirical findings concerning the nature of moral reasoning that, according to him, provide agents with decisive reasons to embrace free and equal moral personhood as a structural feature of social morality.[21]

Based on this understanding of the foundations of social morality, Gaus develops a Kantian account of public reason that aims to determine the set of evolved social moral rules that is justifiable to all current members of a society when they decide on such rules in their role as 'members of the public'.[22] Members of the public are idealized versions of actual citizens, and such idealization is necessary, according to Gaus, to avoid potentially faulty reasoning and partiality of actual citizens in the public decision-making process. That is, Gaus' members of society not only must embrace the moral ideals of freedom and equality, but they also must be impartial in their role as members of the public, as generally is assumed, in some form, by the liberal tradition.

> [A] Member of the Public is an *idealization* of some actual individual; a Member of the Public deliberates well and judges only on the relevant and intelligible values, reasons, and concerns of the real agent she represents and always seeks to legislate impartially for all other Members of the Public.[23]

In addition, although Gaus' account of public reason does not demand that agents, in their role as members of the public, embrace social moral rules on the basis of universally shared reasons but only from their own perspectives, the standards of justification that are employed by Gaus' account of public reason must be reasonable. The standards must pass a test of coherence and plausibility such that agents in their role as members of the public can recognize each others' views as moral perspectives and not merely as the expression of egoism, malice, or other antisocial preferences. Gaus calls this requirement the 'mutual intelligibility' condition, which restricts the evaluative diversity among public deliberators to reasonable viewpoints.[24]

[19] See Gaus (2011a, pp. 176–177).

[20] Ibid., p. 14.

[21] Ibid., Chap. VI.

[22] Ibid., p. 48.

[23] Ibid., p. 26.

[24] Ibid., pp. 277–280.

Moreover, Gaus argues that the outcomes of his public decision-making procedure must fulfill certain reasonable conditions. In order for social moral rules to be justifiable, they must be (i) general, (ii) teachable, (iii) adequate for conflict resolution, (iv) normatively weighty and overriding, (v) reversible,[25] (vi) respectful of the core interests of others (including their bodily integrity and basic liberties), and (vii) not excessive in terms of enforcement costs. Evolved social moral rules that fulfill these standards and make everyone better off than they would be in the state of nature are part of what Gaus calls the 'socially eligible set' of systems of social moral rules. If there are systems of social moral rules in this set that are preferred by all members of society in their role as members of the public, and the non-preferred systems of social moral rules are excluded, then the 'socially optimal eligible' set of systems of social moral rules is determined.[26] Gaus argues that, as a result of moral pluralism, this set of systems of social moral rules is usually neither empty nor singleton. Instead, it usually contains multiple sets of systems of social moral rules that are all publicly justifiable to all current members of society in their role as members of the public.

In order to narrow the socially optimal eligible set of systems of social moral rules further, Gaus introduces the notion of rights. He argues that all members of society in their role as members of the public would agree with certain personal rights that protect individual agency, such as rights against (i) coercion, (ii) deception, and (iii) physical harm, and more positively, rights to (iv) freedom of thought and conscience, and (v) weak assistance concerning individual wellbeing. In addition, Gaus claims that the members of a society in their role as members of the public would agree with certain jurisdictional rights that grant all group members full moral authority over certain parts of the social moral domain. Such rights include (vi) strong private property rights, (vii) the right to privacy, and (viii) the right to freedom of association.[27]

Despite this substantive list of rights that narrows the outcome space of Gaus' procedure of public reason, Gaus argues that, as a result of moral pluralism, public reason cannot determine a uniquely justified system of social moral rules. Public reason is indeterminate under the assumption of moral pluralism. In order to determine a unique system of social moral rules, the helping hand of social and cultural evolution once more is needed. In order to support this claim, Gaus presents a simple game-theoretic model that he calls the 'Kantian coordination game'.[28] This game shows that it is possible for a uniquely justified system of social moral rules to emerge as a result of social interactions among agents who act exclusively on their own reasons. In this game, iterated social interactions ensure that agents gain sufficient reasons to accept a particular system of social moral rules simply because others already have adopted this system, and the more agents adopt this system, the more other agents have reasons to adopt it too, which ultimately leads to a

[25] The reversibility condition demands that agents' endorsements of specific social moral rules do not depend on agents' knowledge that they occupy specific social positions.

[26] See Gaus (2011a, p. 323).

[27] Ibid., Chap. VI.

[28] Ibid., p. 395.

bandwagon effect and convergence on one system of social moral rules. Overall, for Gaus, publicly justified social morality consists of a system of social moral rules that is selected by social and cultural evolutionary processes from a variety of systems of social moral rules that are optimal in relation to rules that lie outside of Gaus' eligible set of systems of social moral rules and that can be justified rationally to all current members of society. Gaus defends an *evolutionary account of public justification*.

The discussion of the core features of Gaus' moral theory shows that Gaus' evolutionary account of public justification assumes that the members of a society in their role as members of the public agree, for the justification of social moral rules, not only on core liberal moral ideals, such as freedom, equality, and impartiality, and core liberal personal and procedural rights, but also on certain reasonable standards that social moral rules must fulfill. As such, although Gaus' moral theory combines evolutionary morality with a form of Kantian moral rationalism, and thus, has a broad scope, Gaus' moral theory is valid only for broadly liberal reasonable moral agents who accept the substantive moral premises that underlie Gaus' evolutionary account of public justification.

In other words, Gaus' moral theory is valid only for reasonably pluralistic societies in a distinctively liberal sense. It is valid only for societies of "…free and equal *moral persons* in a world characterized by deep and pervasive yet *reasonable disagreements* about the standards by which to evaluate the justifiability of claims to moral authority…."[29] The members of such reasonably pluralistic societies may still have significant moral disagreements that may leave empty the socially optimal eligible set of systems of social moral rules that is defined by Gaus' evolutionary account of public justification. However, Gaus is optimistic that, for the societies assumed by his moral theory, the socially optimal eligible set of systems of social moral rules is generally not empty and that a determinate social moral equilibrium always evolves.[30]

Nevertheless, the discussion shows that Gaus' moral theory does not apply to *deeply pluralistic societies* in which not all members of society are liberal moral agents of a certain kind, but in which the members of society may also hold non-liberal moral ideals, or no moral ideals that are sufficiently strong to determine their reasoning and behavior.[31] Such non-liberal moral agents and, according to the traditional understanding of social morality, non-moral agents will not accept the substantive liberal moral ideals and rights upon which Gaus' evolutionary account of public justification is based, and thus, they will not accept the social moral rules that are determined by this account. Gaus' moral theory does not apply to deeply pluralistic societies that may be populated by liberal moral agents, non-liberal moral agents, and, according to the traditional understanding of social morality, non-moral agents.

In other words, Gaus' moral theory operates within the domain of traditional morality that assumes agreement on at least some moral ideals as a basis for the

[29] Ibid., xv. My italics.

[30] Ibid., p. 323.

[31] Gaus (2011a, p. 281) explicitly excludes non-moral agents from his moral theory.

derivation of social moral rules, and, as such, it is not surprising that the purely instrumental approach to morality fails for the purpose of Gaus' project. By contrast, the purely instrumental approach to morality does not assume such agreement, but only agreement on a shared end, and, as such, is methodologically adequate for the derivation of social moral rules for deeply pluralistic societies. Based on this understanding of the distinction between traditional morality and instrumental morality, I develop in the remainder of this article a comprehensive moral theory that combines traditional morality (evolutionary, rational, or both) and instrumental morality, and clarifies the adequate scope of pure moral instrumentalism and its legitimate place in moral theory. I call this theory a *two-level contractarian theory*, following terminology that I have introduced in a different context.[32]

5 A two-level contractarian theory

A theory of social morality must fulfill primarily two tasks. It must justify standards of behavior that regulate social interactions among the members of a society, and it must provide sufficient reasons for all group members to follow these standards. These two tasks are not unrelated. If, for example, the members of a society share a certain moral sense as a result of shared moral ideals, then the task of a moral theory is to determine the social moral rules that correspond to the specific moral sense of the group members. If these rules are determined by a sound justificatory procedure that is uniquely justifiable to all current members of society for the domain for which the rules prescribe behavior, then the group members have decisive reasons to follow the rules derived, if sufficiently many other group members follow the rules, too.[33]

This, precisely, is the rationale for most traditional moral theories that assume a consensus on moral ideals among the members of a society as a basis for the derivation of social moral rules. Typically, social moral rules, as traditionally conceived, evolve as conventions from actual social practice, but they also may be justified more explicitly by rational procedures, or by a combination of evolutionary and rational procedures, as Gaus suggests. The precise content and form of traditional social moral rules depend on the specific moral ideals, and corresponding moral sense, of the members of a society and the society's empirical conditions, such as its historical, cultural, and geographical contexts. As Gaus points out, we do not artificially design whole systems of social moral rules. Instead, we take the

[32] See Moehler (2009, p. 203). I use the term 'two-level theory' differently from Feldman (2012).

[33] Gaus (2011a, p. 392) rejects this commonly accepted condition of moral justification that he calls, in a more precise formulation, the *procedural justification requirement*, in particular because he argues that it cannot be fulfilled under the assumption of significant evaluative diversity. I am not as pessimistic as Gaus and, in the following, argue implicitly that, even under the assumption of extreme pluralism, a uniquely publicly justifiable procedure that takes pluralism seriously can be determined for the derivation of a social moral rule, if the domain of this procedure and the moral rule derived by it are adequately restricted.

systems that we have, select from them, and adapt them over time, and thereby we may, at least in some cases, improve our moral systems in some respects.[34]

The two-level contractarian theory considers the (implicit or explicit) agreements of the members of a society with their traditional social moral rules as the *first contract* into which they enter and, as long as all members of society agree with the existing social moral rules of their society, either on the basis of their evolved moral sense or on rational grounds, or both, after they have reflected on the normative demands and implications of these rules, the group members can establish any social moral rules that reflect their interests and traditions. The two-level contractarian theory regards these first-level social moral rules, as traditionally conceived, to be the primary source for regulating social behavior, because these social moral rules are tailored specifically to a group's moral history and its conditions of social cooperation. The better the first-level social moral rules express the specific situation and moral sense of the members of a society, the more the members of society have reasons to comply with these rules, because doing so allows them to live in accordance with their moral ideals.

As Gaus argues, in order for social morality, as traditionally conceived, to emerge and be maintained over time, it seems that at least some members of society must be rule-following punishers. For the reasons discussed in Sect. 3, purely instrumentally rational agents are not likely to be able to establish and maintain a framework of social morality as traditionally conceived.[35] However, in modern pluralistic societies not all members of society are necessarily rule-following punishers, and even if they were, they would not necessarily share, due to their divergent moral ideals, the same moral sense that could serve as a basis for the derivation of social moral rules for all relevant types of social interaction. In modern pluralistic societies, agents embrace often irreconcilable moral ideals, and some agents may hold no moral ideals that are sufficiently strong to determine their reasoning and behavior. As such, for modern pluralistic societies it cannot be assumed that a consensus on moral ideals among all members of society necessarily exists or evolves that could serve as a basis for the derivation of social moral rules that are valid for all members of society for all types of social dispute. For such deeply pluralistic societies in which, according to Gaus' terms and the traditional understanding of morality, the eligible set of systems of social moral rules is empty, the traditional approach to morality reaches its limitations.

However, that the traditional approach to morality reaches its limitations does not mean that moral theory *per se* reaches its limitations, because if the members of deeply pluralistic societies have at least one end in common that they aim to reach despite their differing starting points and resultant conflicts, then the purely instrumental approach to morality applies. For societies, or for types of social interaction within societies, in which agents do not have a shared moral basis as a starting point for resolving social disputes but only a shared end, the purely

[34] For a potential genealogy of (parts of) our social morality and the notion of moral progress, see Kitcher (2011, Chaps. 1–4 and 6).

[35] See also Gaus (2011b, pp. 83–86), where he argues that, although not all members of society must be rule-following punishers, at least some members of society must be rule-following punishers in order for large-scale social cooperation to be maintained in the presence of free-riders.

instrumental approach to morality has normative authority. More precisely, the purely instrumental approach to morality has full normative authority for social disputes that have the following characteristics. I call these social disputes *cases of conflict*.[36]

First, in cases of conflict, the conflicting parties already have tried to resolve their disputes based on the specific moral and epistemic ideals and procedures that they embrace, and based on implicitly or explicitly mutually agreed conventions that may be historically justified or the product of (biological, social, or cultural) evolutionary processes. That is, the parties to a conflict have considered the specific ideals, reasons, and histories of their opponents as well as the existing evolved social morality of their society in an informed dialogue, but all such attempts to resolve their disputes have failed, although they may have led to clarification of the disputed issues. Further, all attempts to settle the disputes by appeal to a third party, such as an independent arbitrator, also have failed.

Second, in cases of conflict, the parties are so severely negatively affected by the points of contention that they cannot simply go on with their lives without some form of conflict resolution. As a consequence, if the disputes remain unresolved, the conflicting parties may consider that destructive actions are more beneficial for them than remaining in their current situations. In such situations, the agents may be prepared to endanger the lives of their opponents or those who are close to them, or they may endanger the benefits of peaceful long-term cooperation by slowing social development and economic growth and by destroying scarce resources. Or, they may create high costs to deter such negative actions. In short, if cases of conflict in the strict sense defined cannot be settled, not only are the conflicting parties unable to realize the immediate gains that may arise from sharing the goods that are in dispute, but their future gains also may be diminished compared to the outcomes of non-violent conflict resolution.

Given the dependence of human beings on each other and their vulnerability in a world of (moderately) scarce resources, rational forward-looking agents who do not discount the benefits of future social cooperation too much have a *prima facie* interest in peace. In the empirical world in which human beings live, peaceful cooperation is an instrumental good, because it allows agents to secure their lives and significantly raise their standards of living by the exchange of social and private goods, specialized cooperation, and the accumulation of capital.[37] Further, peaceful cooperation is likely to advance social development and higher economic growth in the long run, and thus, if the members of a society cooperate peacefully with one another over time, then they can realize additional gains from stable long-term cooperation and from avoiding the costs that are associated with the destruction of scarce resources through conflict, including the destruction of life itself.

As such, if the members of deeply pluralistic societies regard living in a society to be more beneficial than living in the state of nature, and they expect that the gains from peaceful long-term cooperation are greater than the gains from violent conflict resolution, then they have an interest in resolving the hard cases of conflict that I

[36] For the following, see also Moehler (2012, pp. 87–88).

[37] See, for example, Smith (1776) and Ricardo (1817).

have just described. In such cases in which traditional morality cannot ensure peaceful long-term cooperation, the purely instrumental approach to morality is authoritative, because it is neutral with regard to the motivation and content of agents' interests. According to the purely instrumental approach to morality, agents do not need to be motivated to follow social moral rules on the basis of what are traditionally conceived to be moral reasons, although agents may be motived by such reasons.

This does not mean, however, that the purely instrumental approach to morality offers generally the 'wrong sort of reasons' for acting morally.[38] Instead, the purely instrumental approach to morality may offer the 'wrong sort of reasons' only from the perspective of traditional morality, which is not authoritative for situations of social interaction that are regulated by the purely instrumental approach to morality, because, by definition, no agreement exists in these situations about the precise nature and demands of social morality. That is, the 'wrong sort of reasons' objection arises only if the distinction between the domain of the traditional approach to morality and the domain of the purely instrumental approach to morality is not adequately considered.

The adequate domain of the purely instrumental approach to morality is restricted to situations of social interaction in which moral reasoning is reduced to instrumental reasoning, and in these situations traditional moral reasons are only one sort of reason that may guide agents' moral behavior, which aims to advance a shared end among the members of a society. According to the instrumental approach to morality, there is no 'wrong sort of reasons', as long as agents' reasons for action aim to promote a common end. Further, as long as the purely instrumental approach to morality aims to harmonize the behavior of the members of a society in order to advance their wellbeing and/or to protect their status as persons, the social moral rules derived by the purely instrumental approach to morality will not be mere conventions, such as driving on the left side or the right side of the road. Instead, the rules will have the character and properties of traditional social moral rules, although, as I clarify in the next section, these rules are not necessarily socially and culturally entrenched by a corresponding moral sense.

For Hobbes, the purely instrumental approach to morality, as expressed by the 'science' of the laws of nature that determine the behavioral restrictions that are necessary to secure peaceful long-term cooperation, is the 'true and only moral philosophy'.[39] The two-level contractarian theory, by contrast, considers the purely instrumental approach to morality only as one moral approach that is valid for a specific domain of social morality, namely, for situations of social interaction in which moral reasoning is reduced to instrumental reasoning, as expressed by cases of conflict in the strict sense defined. These situations of social interaction represent the adequate domain of pure moral instrumentalism, and they constitute the *second level* of the two-level contractarian theory that combines traditional morality with instrumental morality.

[38] For such criticism, see Gaus (2011a, pp. 185–187), for example.

[39] See Hobbes (1651, Part I, Chap. XV).

6 Pure instrumental morality

If the members of deeply pluralistic societies aim to secure peaceful long-term cooperation, then they must find agreement on a rule of conflict resolution that allows them to settle cases of conflict in the strict sense defined in the previous section. This rule must be more general and abstract than traditional social moral rules that are specifically tailored to the conditions of particular societies, and the procedure for the derivation of this rule cannot be based on substantial moral ideals as a starting point, because such moral ideals may be controversial in deeply pluralistic societies, and thus, they are not necessarily accepted by all members of society. The only basis for the derivation of the rule of conflict resolution is that agents are rational and that they have interests that they aim to fulfill, irrespective of the motivations and the nature of their interests.

Previously, I derived such a rule of conflict resolution in a game-theoretic framework for the circumstances described.[40] In the following, I do not repeat the details of my argument. Instead, I summarize the core features of (i) the model of agency that underlies the derivation of the rule of conflict resolution, (ii) the game-theoretic procedure that is employed to derive the rule, and (iii) the rule itself. The purpose of the discussion is to make my conclusions accessible to a broader readership that is less interested in the technical details of my argument than in its implications for moral theory.

The derivation of the rule of conflict resolution relies on an account of instrumental rationality that I call the *homo prudens* model. Agents who reason in the mode of *homo prudens* aim to fulfill their interests not only today but also in the future, although they may discount future benefits according to their time preference. The *homo prudens* model considers all types of interest that agents may have, independent of the motivation and content of their interests, apart from a positive concern of agents for the interests of their opponents in conflict situations. That is, although agents may have nontuistic or negative tuistic interests in cases of conflict, such as a desire to dominate their opponents or to thwart their opponents' interest satisfaction, which may be the result of motives such as envy or hate, agents are not assumed to take a genuine interest in promoting the interests of their opponents for the derivation of the rule of conflict resolution, because such an interest may help to mitigate the conflict, and attempts to do so are assumed to have failed already in cases of conflict in the strict sense defined.

The only two substantial requirements that the *homo prudens* model imposes on agents are that the agents are assumed to have an interest in preserving their lives in cases of conflict, and they are assumed to value their lives in the long run, together with the benefits from peaceful long-term cooperation, more than they value non-cooperation *per se* in any particular case of conflict, whatever the conflict issue and the agents' attitudes towards their opponents may be. In addition, the agents are assumed to have reflected on their interests when they reason in the mode of *homo prudens*, and their ordering of interests must fulfill

[40] See Moehler (2012, pp. 88–101). In the following, I borrow from my previous discussion.

certain consistency requirements that allow the agents' interests to be represented by well-defined utility functions.[41]

Like Gaus' moral theory, which assumes that agents reason as members of the public when they decide on social moral rules, my theory assumes that agents reason in the mode of *homo prudens* when they decide on the rule of conflict resolution, because the rationality of *homo prudens* adequately expresses the situation of forward-looking agents who, although they have an interest in peaceful long-term cooperation, aim to fulfill their interests maximally in conflict situations. In cases of conflict in the strict sense defined, agents cannot rely on a richer form of rationality than is assumed by the *homo prudens* model, and the concept of rationality that is employed for the derivation of the rule of conflict resolution cannot rely on specific empirical conditions about agents' rational capacities that do not necessarily apply to all members of society.

In contrast to Gaus' moral theory, the agents who reason in the mode of *homo prudens* are placed into an orthodox game-theoretic framework for the derivation of the rule of conflict resolution, and not into an evolutionary framework or a framework that combines evolutionary with rational considerations in the way described by Gaus. Such evolutionary models are inadequate for the derivation of the rule of conflict resolution, in particular because the results of these models may be (i) path-dependent, and thus the product of empirical contingencies that may not necessarily be acceptable to all current members of deeply pluralistic societies, (ii) based on an insufficiently transparent history that may not allow all current members of society to judge whether the evolved social moral rules are fully justifiable to them, and (iii) biased towards the moral *status quo*.[42] As a consequence of these features, the results of evolutionary models of justification do not necessarily have sufficient normative authority for all current members of society in cases of conflict in the strict sense defined, in which agents cannot rely on a shared moral sense or experiences of past interactions as mechanisms to coordinate their behavior, because attempts to settle their disputes on the basis of such, at least partially, empirically contingent mechanisms of conflict resolution are assumed to have failed already.

The game-theoretic procedure that is employed for the derivation of the rule of conflict resolution is simple. It requires only one essential modification compared to the game-theoretic models that are commonly used in the social sciences. Although all agents in the idealized decision situation, in which they decide on a rule of conflict resolution, are assumed to reason in the mode of *homo prudens*, the actual behavior of agents is not necessarily assumed to be guided by this idealized form of reasoning, but by the agents' natural and evolved dispositions, emotions, ideals, rules of thumb and, for truly moral agents as traditionally conceived, by the social moral rules of their society that represent the first contract of the two-level contractarian theory into which agents enter. In order to allow individuals in the

[41] The precise demands are specified in Moehler (2012, p. 90).

[42] Gaus' (2011a, p. 425) evolutionary account of public justification, for example, holds that "[i]f an existing rule *is* within the optimal eligible set, it is publicly *justified* for that reason—just because it *is* the existing rule." My italics.

idealized decision situation to identify with real-world agents, including their own real-world counterparts, the *empathetic contractor position* is introduced. The empathetic contractor position is an analytic device that, in the current context, allows rational agents who reason in the mode of *homo prudens* to place themselves into the behavioral shoes of their fellows in cases of conflict based on their own experiences, and thus to model adequately an *n*-person decision situation in deeply pluralistic societies.

A consequence of this kind of incomplete information that agents experience in the empathetic contractor position as well as in the real world is that agents cannot precisely predict the behavior of their fellows and, in this sense, face a situation of weak uncertainty with regard to their future. That is, for the derivation of the rule of conflict resolution, agents are assumed to be covered by a *veil of uncertainty* under which they know their empirical reality, but they do not know the specific cases of conflict in which they may become involved in the future and their precise positions in these conflicts.[43] In particular, the agents do not know whether they will be stronger or weaker than their opponents in (all) future cases of conflict. Because the empathetic contractor theory preserves only the form of uncertainty that individuals face in the real world, the procedure is realistic, although it is idealized, and most importantly, it does not introduce controversial moral assumptions into the decision-making process for the derivation of the rule of conflict resolution. What is the outcome of this decision-making procedure?

Rational agents follow the rule of conflict resolution only if the rule prescribes behavior that the agents regard, under consideration of the behavior of other agents, as most beneficial for them in cases of conflict in the strict sense defined. As such, the aim of rational agents must be to identify the rule of conflict resolution from the set of possible social moral rules that imposes the fewest restrictions on their behavior that are necessary to secure peace, because in this case, rule-guided behavior, with regard to the rule of conflict resolution, can be explained by standard opportunistic case-by-case decision making. If the rule of conflict resolution specifies the *minimal restrictions* on the behavior of rational agents that must be fulfilled in each case of conflict in order for peaceful long-term cooperation to be secured, and agents expect that institutionalizing the rule of conflict resolution is more beneficial for them than violent conflict resolution, then rational agents will follow the rule of conflict resolution in each case of conflict, because such rule-guided behavior allows agents best to fulfill their long-term interests in the world in which they live, assuming that other agents follow the rule, too.

On the basis of this consideration, I have argued that rational agents would choose the *weak principle of universalization* as a rule of conflict resolution in the outlined decision situation. The principle demands the following:

> In cases of conflict, only pursue your interests subject to the side constraints that your opponents can (i) enter the process of conflict resolution at least from their minimum standards of living, if the goods that are in dispute permit it,

[43] For the notion of the veil of uncertainty, see Buchanan and Tullock (1962, p. 78).

> and (ii) fulfill their interests above this level according to their relative bargaining power.[44]

The weak principle of universalization defends morality in the form 'each according to her basic needs and above this level according to her relative bargaining power'. The principle is valid for human beings who live in this empirical world and who reason in the mode of *homo prudens* in cases of conflict in the strict sense defined. In these cases, moral reasoning is reduced to instrumental reasoning, and the weak principle of universalization has normative authority for instrumentally rational agents who, all things considered, have an overarching interest in securing peaceful long-term cooperation. For other types of social interaction, traditional social morality is authoritative.

From a methodological point of view, the proposed two-level contractarian theory combines two different moral approaches within one comprehensive moral theory: (i) the traditional approach to morality that presupposes agreement on moral ideals as a basis for the derivation of social moral rules, and (ii) the purely instrumental approach to morality that assumes only a shared end among the members of a society, such as the aim of securing peaceful long-term cooperation. As long as traditional morality can regulate all morally relevant types of social interaction among the members of a society, peaceful cooperation is secured (first-level contract). Only in cases where traditional morality falls short because agents cannot find a way to resolve their disputes based on the moral ideals that they embrace does pure instrumental morality represent the only remaining common denominator to guarantee peace (second-level contract). In these cases of social interaction, agents act both irrationally and immorally if they do not take the means that are necessary to reach an end that they aim to reach, all things considered.

In terms of the two models of agency discussed in this article, for most social interactions the members of society may be assumed to be Gausian rule-following punishers, or an approximation of this model of agency, and their social interactions are assumed to be guided predominantly by traditional social morality.[45] In modern pluralistic societies, however, not all agents necessarily share the same moral sense, if they have a functioning moral sense at all that determines their reasoning and behavior. As such, social moral rules, as traditionally conceived, that are valid for all types of social interaction for all current members of society cannot necessarily be determined. For cases of social interaction in which traditional social morality falls short, the members of society must go beyond their moral sensibility, as traditionally conceived, and reason in the mode of *homo prudens* in order to secure peaceful-long term cooperation. For such cases of conflict, agents are assumed to recognize that, under the condition of extreme pluralism, they must abstract from at least some of their specific interests in order to find agreement on a mechanism of conflict resolution that is acceptable to all members of society, and thus, is able to

[44] Moehler (2012, p. 100).

[45] The two-level contractarian theory is not committed to one particular model of agency for the domain of traditional social morality. However, the model of agency that Gaus suggests, or a close cousin of it, seems to be a plausible candidate for this domain.

secure peaceful long-term cooperation. These cases of conflict, in the strict sense defined, represent the domain of pure moral instrumentalism.

The two-level contractarian theory recognizes that the traditional and purely instrumental approaches to social morality, and their underlying models of agency, do not compete with each other, but are valid for different domains. The two-level contractarian theory combines these two moral approaches into one comprehensive moral theory to allow agents, under the assumption of extreme pluralism, to pursue their interests maximally according to their own conceptions of the good and, as such, to exercise their individual freedom maximally in a world of conflicting interests. As a result of this feature of the two-level contractarian theory, agents have maximal reasons to follow the established social moral order, and thus to maintain peace, if other group members follow the agreed social moral rules, too. If the division of labor between the traditional approach to morality and the instrumental approach to morality is respected, and the two moral approaches are combined adequately, then the problem of compliance is minimized and the two-level contractarian theory, if implemented, is most promising for securing mutually beneficial peaceful long-term cooperation in deeply pluralistic societies, as compared to cooperation in a non-moralized state of nature.

In particular, no additional problem of compliance arises for the instrumental approach to morality. If a decision-making procedure for the derivation of the rule of conflict resolution is employed that is uniquely justifiable to all current members of society for the domain for which the rule prescribes behavior, and the procedure models realistically the real-world situations of all members of society, as the empathetic contractor theory does, and the moral rule derived by this procedure imposes on agents only the minimal restrictions that are necessary to secure peaceful long-term cooperation, then the instrumental approach to morality can solve the problem of compliance, because by definition following the rule derived under these conditions pays off for each member of society in each instance for which the rule prescribes behavior, if all members of society assume that the rule is beneficial for them in the long run. In addition, if the members of society are rational, then they have no incentive to free-ride with regard to the rule of conflict resolution, because doing so would be disadvantageous for them in each case for which the rule prescribes behavior. The purely instrumental approach to morality is not inherently inconsistent, if it is applied in its adequate domain.

The only difference between the rule of conflict resolution and traditional social morality is that the rule of conflict resolution must be institutionalized, because the rule is not necessarily socially and culturally entrenched. To this end, social regulating institutions must be established, if they are not already in place, and all members of society must sign an actual contract with these institutions in order for the rule to gain full moral authority.[46] In this process, two parties must be distinguished. The first party is the starter generation that consists of all agents who currently live in a society and who have reached the age of consent. The second party consists of all agents who come new to the contract, such as children who have not yet reached the age of consent, and new citizens. Signing a contract to follow the

[46] See Moehler (2009, pp. 209–210) for a similar argument in the context of the topic of global justice.

rule of conflict resolution in cases of conflict makes the members of a society aware of their dependency on each other and the respect that they owe each other if they want to live peacefully with one another in a deeply pluralistic society.

Acknowledgments I am very grateful to Jerry Gaus not only for helping me to understand his complex moral theory, but also for pressing me to formulate my own theory more carefully.

References

Buchanan, J., & Tullock, G. (1962). *The calculus of consent*. Ann Arbor, MI: University of Michigan Press.

Feldman, F. (2012). True and useful: On the structure of a two level normative theory. *Utilitas, 24*, 151–171.

Gaus, G. (2011a). *The order of public reason: A theory of freedom and morality in a diverse and bounded world*. Cambridge: Cambridge University Press.

Gaus, G. (2011b). Retributive justice and social cooperation. In M. White (Ed.), *Retributivism: Essays on theory and practice* (pp. 73–90). Oxford: Oxford University Press.

Gaus, G. (2013). Hobbesian contractarianism, orthodox and revisionist. In S. A. Lloyd (Ed.), *The Bloomsbury companion to Hobbes* (pp. 263–278). New York: Bloomsbury.

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

Hobbes, T. (1651). In R. Tuck (Ed.) (1996), *Leviathan*. Cambridge: Cambridge University Press.

Kant, I. (1785). In M. Gregor (Ed.) (1998), *Groundwork of the metaphysics of morals*. Cambridge: Cambridge University Press.

Kitcher, P. (2011). *The ethical project*. Cambridge, MA: Harvard University Press.

Lloyd, S. (1992). *Ideals as interests in Hobbes's Leviathan*. Cambridge: Cambridge University Press.

Lloyd, S. (2009). *Morality in the philosophy of Thomas Hobbes*. Cambridge: Cambridge University Press.

McClennen, E. (1988). Constrained maximization and resolute choice. *Social Philosophy and Policy, 5*, 95–118.

Moehler, M. (2009). Justice and peaceful cooperation. *Journal of Global Ethics, 5*, 195–214.

Moehler, M. (2012). A Hobbesian derivation of the principle of universalization. *Philosophical Studies, 158*, 87–88.

Nozick, R. (1993). *The nature of rationality*. Princeton, NJ: Princeton University Press.

Ricardo, D. (1817). *On the principles of political economy and taxation*. London: John Murray.

Robbins, L. (1935). *An essay on the nature and significance of economic science* (2nd ed.). London: MacMillan.

Schroeder, M. (2004). The scope of instrumental rationality. *Philosophical Perspectives, 18*, 337–364.

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

Smith, A. (1776). In E. Cannan (Ed.), *An inquiry into the nature and causes of the wealth of nations*. London: Methuen.

Vanderschraaf, P. (2010). The invisible foole. *Philosophical Studies, 147*, 37–58.