1 Introduction

1.1 Contractarian Moral Theory

Glaucon's Challenge

Contractarian moral theory enjoys a long tradition in moral philosophy and extends back at least to ancient philosophy.[1] In Plato's *Republic*, Glaucon, one of the main protagonists, challenges Socrates to refute what Glaucon considers to be the common view of the "nature and origin of morality" (Plato 1993: 358c), which differs significantly from Socrates's own view. According to the common view of morality, morality is contractarian. Morality is the result of human agency and established by agreement among primarily self-interested, although not self-sufficient, agents who, if necessary, pursue their own good at the expense of others. Because agents are typically not strong enough to dominate others entirely and want to secure their freedom, they agree to form society and punish so-called immoral behavior. This 'social contract' protects agents from being victimized and, in turn, demands that agents give up the benefits of exploiting others. On the basis of this origin, morality is only ever unwillingly practiced and followed to the extent necessary to ensure peace. Morality has no intrinsic worth but is a means to other ends that agents value. Morality "is a compromise between the ideal of doing wrong without having to pay for it, and the worst situation, which is having wrong done to one while lacking the means of exacting compensation" (Plato 1993: 359a).

Glaucon exemplifies this common view of morality with reference to the myth of Gyges's ring. Gyges's ring allows agents to become invisible. As such, while wearing the ring, agents do not need to fear social sanctions, such as punishment or loss of reputation, for immoral behavior. According to Glaucon, Gyges's ring helps to reveal the true nature of morality.

> Suppose there were two such rings, then – one worn by our moral person, the other by the immoral person. There is no one, on this view, who is iron-willed enough to maintain his morality and find the strength of purpose to keep his hands off what doesn't belong to him, when he is able to take whatever he wants from the market-stalls without fear of being discovered, to enter houses and sleep with whomever he chooses, to kill and to release from prison anyone he wants, and generally to act like a god among men. His behavior would be identical to that of the other person: both of them would be heading in the same direction. (Plato 1993: 360b)

Expressed in modern terms, if agents do not need to fear social sanctions for immoral behavior, then they will free ride and exploit others in order to enjoy

[1] For a brief historical overview of the contractarian tradition, see Sayre-McCord (2000).

the benefits from social cooperation without having to pay the costs for it. According to the common view of morality, the social contract that establishes social moral order serves as a 'straitjacket' that keeps agents in check and from which agents try to escape whenever they can.

In the *Republic*, Socrates offers a response to Glaucon's challenge to show that the origin and nature of morality are more honorable than is expressed by the common view. Socrates's goal is to show that morality is the best type of good that is valued for its own sake and for its natural consequences independent of external benefits such as wealth or reputation. Socrates argues that morality secures harmony in the mind and, in doing so, mental health, which is constitutive of happiness: "Goodness is a state of mental health, bloom, and vitality; badness is a state of mental sickness, deformity, and infirmity" (Plato 1993: 444d). To address the problem of moral motivation (the question of why be moral), especially in his speech on love in the *Symposium* (Plato 1994: 201d–212 c), Socrates, with reference to Diotima, argues that once agents have acquired knowledge of what is morally good they will act out of love of the good. According to Socrates, knowledge of the good is inherently motivating, and thus ultimately agents will comply with moral rules for intrinsic reasons and not out of self-interest. Ultimately, moral behavior is an expression of agents' admiration for moral goodness.

In an important sense, Socrates changes the subject in his response to Glaucon because Socrates simply assumes that the origin and nature of morality are more honorable than is expressed by the common view. Also, Socrates's argument that the demands of morality are universally true, and thus independent of human agreement, relies on substantial metaphysical and epistemological assumptions that are difficult to prove. In this Element, I do not assess Socrates's argument or any other argument that aims to show that the common view of morality is mistaken. Instead, although I acknowledge that competing views of the origin and nature of morality exist, I argue that the common view of morality, which in contemporary moral philosophy is expressed most closely by the position of 'moral contractarianism,' is a plausible view of morality.

In fact, moral contractarianism has significant strengths and, if appropriately conceived, is conceptually coherent, empirically sound, and practically relevant, especially for deeply morally diverse societies. In such societies where, according to Gauthier (1991: 15), "morality faces a foundational crisis," the standards of morality are controversial. More strongly, I argue that, under certain specific conditions, moral contractarianism is the only defensible approach to morality that can ensure mutually beneficial peaceful long-term cooperation. In order to support this claim, this Element clarifies the core features and appropriate place of moral contractarianism in moral theory.

Owing to the general nature of this project, the Element offers a broad view that is necessary to connect the different parts of this argument without focusing on all of its details that have been defended elsewhere.

Moral Contractarianism, Conventionalism, and Contractualism

Contractarian moral theory justifies moral rules through agreement among agents on the moral rules by which the agents are affected. Contractarian moral theory assumes that such implicit or explicit agreement is voluntary in that agents accept the agreed-upon moral rules if they reflect freely on the rules' demands and implications, although the agents may agree with the rules for different reasons. As Sugden (2018: 32) stresses, contractarian theory "takes account of what is good for each party, from its own viewpoint, without needing to consider what is simply good, as viewed from nowhere." The core tenet of contractarian moral theory is that, independent of the precise form of agreement that is assumed, if agents agree with the moral rules that govern their interactions, then the agents have no reason to reject the authority of the rules because the agents themselves are the authors of the rules. Contractarian moral theory is antiauthoritarian and respects the autonomy of agents for the justification of moral rules. There are, however, different approaches within contractarian moral theory. For the discussion in this Element, it is important to distinguish 'moral contractarianism' from its two close cousins, 'moral conventionalism' and 'moral contractualism.'[2] Unfortunately, the distinction among these different approaches and their labeling have not been applied consistently, which has led to much confusion and unwarranted criticisms of contractarian moral theory.[3]

In modern philosophy, the position of 'moral contractarianism' extends back to Hobbes's (1651) moral theory and has been advanced most notably by Gauthier (1969, 1986), Hampton (1986), Kavka (1986), and Moehler (2018a), although Hampton's and Kavka's arguments include a significant discussion of political theory. As a member of the European Enlightenment, Hobbes aimed to expose all considered truths to skeptical doubt and accept only demands that the human intellect can establish. For discovering the truth, Hobbes considered mathematics, especially geometry, as a model form of reasoning (I return to this consideration in Section 2.2). Hobbes's goal was to develop a post-skeptical science of morals that is based on strict conceptual analysis and realistic

[2] For the distinction between contractarianism and contractualism, see Darwall (2003: 1–8), although Darwall does not explicitly distinguish between contractarianism and conventionalism. See also Gauthier (1997: 134–135); Watson (1998: 173–174); D'Agostino, Gaus, and Thrasher (2017); and Moehler (2018a: 11–12).

[3] See Hampton (1991: 32–33).

assumptions about human nature and social cooperation, a morality that agents are actually motivated to follow.[4] To this end, according to Hobbes, morality must appeal to agents' desire for self-preservation and commodious living and, more generally, to the goal of ensuring peaceful long-term cooperation, and not appeal primarily to compassion or a sense of fairness. As Kavka (1986: 310) puts it, "if moral systems are to be *practical* their requirements must link up in appropriate ways with people's motivational capacities."

Specifically, as the discussion of Glaucon's challenge indicates, moral contractarianism assumes that agents are rational and tend to pursue their own interests. In addition, it assumes that agents are roughly equal by nature in that the weakest is able to kill the strongest "either by secret machination, or by confederacy with others, that are in the same danger with himselfe" (Hobbes 1651: Part 1, chapter 13). As a result of these assumptions, moral contractarianism often is associated with bargaining theory and its underlying concept of mutual advantage.[5] However, as I clarify in Sections 3.1 and 4.1, the association of moral contractarianism with bargaining theory must be considered with care because bargaining theory can be applied in different ways to moral theory, and simply because agents agree with a bargaining principle as a moral principle does not mean that they are assumed to bargain with each other over the demands of morality. Also, the application of bargaining theory to moral theory is not unique to moral contractarianism. Instead, it is typically also a core feature of 'moral conventionalism.'

Moral conventionalism originates with Hume's (1739/1740) moral theory.[6] Hume agrees with Hobbes that agents are primarily self-interested. However, according to Hume, agents are also morally sensible. They possess natural virtues, in particular benevolence, which make the agents consider the interests of others. Moreover, for society to be established, Hume argues that agents must acquire artificial virtues in the form of moral conventions that arise from a combination of self-interest and the understanding that reciprocal social behavior is usually mutually beneficial. According to Hume, moral conventions are not the result of a counterfactual social contract but are manifested in agents' actual behavior. Although moral conventionalism does not explicitly invoke the metaphor of the social contract, methodologically the approach fulfills the core

[4] For discussion of Hobbes's method of investigation and his assumptions about human nature, see Gauthier (1969: 1–26). For discussion of Hobbes's moral theory from a metaethical perspective, see Abizadeh (2018).

[5] See Stark (2009: 75), for example.

[6] For contemporary theories of moral conventionalism, see Sugden (1986, 2018), Binmore (1994, 1998, 2005), Skyrms (1996, 2004), Alexander (2007), and Vanderschraaf (2019).

requirements of contractarian moral theory, especially its core notion of agreement.[7]

However, moral conventions are the result of ongoing coordination among agents on which the agents have unequal influence, and thus the specific moral conventions that evolve in society may not always be in the best interest of all current members of society, even if all current members of society consider the existing moral conventions to be better than having no such conventions.[8] That is, although the current system of moral conventions may be strictly Pareto-superior to the state of nature, some members of society may favor other systems of moral conventions that allow them to benefit more than they do under the current system or that match their moral sense more closely. Further, even if the existing moral conventions were maximally beneficial for all members of society and match their moral sense, agents' short-term interests may often conflict with their long-term interests, in which case agents may be tempted to free ride.

According to Hume, agents' continued adherence to the moral conventions of their society can be explained by the fact that agents' private interests are usually closely linked with the common good of having a stable social moral order, and thus agents generally have an interest in adhering to the established moral conventions. Moreover, Hume argues that, if moral conventions are sustained over time, then agents will start to value the existing conventions intrinsically and not merely for instrumental reasons. Over time, agents will internalize the existing moral conventions of their society by developing a moral sense that corresponds to and approves of the established moral conventions. Agents will develop dispositions to follow the established moral conventions and adherence to these conventions becomes the agents' second (moral) nature.[9]

Stated differently, Hume believes that reason alone is not sufficient to motivate agents to follow the established moral conventions of their society permanently. Instead, a transformation of agents' behavioral dispositions must occur that control the agents' self-interest in the short term. Hume assumes that agents develop 'commitment power' that predisposes them to follow the

[7] See Gauthier (1979) and Sugden (2018: 33–37). In this context, see also Thrasher (2015), Hankins (2016), and Vernon Smith and Wilson (2019) for discussions of Adam Smith's moral (and economic) theory that shares similarities with Hume's theory and may also be considered to be part of the contractarian tradition.

[8] In this context, see Gaus (2015), who argues that biological evolution provides strong evidence for the development of egalitarian moral sentiments in the history of human cooperation. By contrast, O'Connor (2019) and Cochran and O'Connor (forthcoming) argue that egalitarian moral sentiments may be fragile. The authors show that, in simple cultural evolutionary models of social groups, inequity is more likely to emerge than equity.

[9] In this context, see Hampton (1998b: 156–165).

existing moral conventions of their society, even if such rule-following behavior is not beneficial to them in each instance.[10] This process of internalization renders moral conventions to be self-enforcing and provides a solution to the free-rider problem that is implicit in Glaucon's challenge (I return to this consideration in Section 3.2, where I discuss Gauthier's moral theory and his notion of constrained maximization).

Compared to moral conventionalism, 'moral contractualism' assumes a more demanding and specific moral basis for the justification of moral rules. Moral contractualism originates with Kant's (1785) moral theory and has been defended systematically by Rawls (1971), Scanlon (1998), Darwall (2006), and Southwood (2010), although Southwood uses the term more broadly.[11] Moral contractualism assumes that agents are rational and reasonable and that agents' reasonableness constrains their behavior in moral interactions. According to Rawls (1993: 51), who defends a moral and political theory, reasonable agents possess a "particular form of moral sensibility that underlies the desire to engage in fair cooperation as such, and to do so on terms that others as equals might reasonably be expected to endorse." Reasonable agents have a desire to justify their actions toward others, not because of their natural equality and potential threat to each other, but because they respect each other as free and equal persons. Reasonable agents consider each other as moral equals.

In this sense, moral contractualism, like moral conventionalism, assumes morally sensible agents. In addition, as typically defended in the literature, moral contractualism assumes that agents possess a particular liberal moral sense that, in some form, relies on the moral ideals of freedom, autonomy, equality, impartiality, and reciprocity. Rawls, for example, by means of his 'original position' (which I discuss in Section 2.3), derives principles of justice that match the moral sense of particularly liberal moral agents. The original position is an analytic device that allows Rawls to rationally derive principles of justice that correspond to the specific moral sense of reasonable liberal moral agents. In Rawls's (2001: 81–82) words, "the reasonable conditions imposed on the parties in the original position constrain them in reaching a rational

[10] For further discussion of the notion commitment, see Schmidtz (1995: 106–111). For support of Hume's view of the evolution of morality, see Bowles and Gintis (2011).

[11] Southwood argues that his 'deliberative model of contractualism' represents an alternative to Hobbesian contractarianism and Kantian contractualism. Southwood's theory (2010: 88–96, 124–128) assumes that agents are deliberatively rational, which demands that agents actively engage in deliberation with others, consider their views, and are accountable to each other. In this sense, agents must respect each other as moral equals in order to be part of society. In addition, the agents' deliberative processes underlie strict norms and require open, good-faith, and receptive back-and-forth communication with the goal to reach consensus on a 'common code' by which to live together.

agreement on principles of justice." Rawls's constructivist procedure ensures that reasonable liberal moral agents will follow the demands of the principles of justice derived in the original position and instituted by the basic structure of society in the real world.

In Socrates's spirit, moral conventionalism and moral contractualism express a more honorable view of morality than is expressed by the common view of morality and captured by moral contractarianism. Moral conventionalism and moral contractualism assume that agents care for each other, consider each other's views, and are intrinsically motivated to do what is morally right. Moral conventionalism and moral contractualism do not consider morality to be purely instrumental. Instead, these two approaches assume a shared moral basis among agents that either evolves over time or is presupposed as a starting point for the justification of moral rules. In this sense, moral conventionalism and moral contractualism are 'traditional moral theories.' Traditional moral theories (as I employ the term) assume, as a basis for the justification of moral rules, that agents value moral ideas at least partially for intrinsic reasons or embrace such ideals for other traditional moral reasons, such as altruistic reasons or similarly motivated other-regarding reasons.[12]

The assumptions of moral conventionalism and moral contractualism and, more generally, the assumptions of traditional morality hold neither conceptually nor empirically for all societies and their members, nor for all morally relevant types of social interaction in such societies. In our world, not all agents are morally sensible or are morally sensible in the same specific way. That is, even if all members of a society were genuine moral agents as traditionally conceived, in morally diverse societies the agents' moral views may conflict with each other and lead to severe conflict. In such cases, the purely instrumental approach to morality, as captured by moral contractarianism, applies if agents share an overarching goal, such as the goal of ensuring peaceful long-term cooperation, despite their conflicting traditional moral views or lack thereof.

1.2 Core Features of Moral Contractarianism

The Tasks of Moral Theory

Moral theory aims to justify moral rules and provide agents with sufficient reasons to comply with these rules. For moral conventionalism and moral

[12] For discussion of potential overdetermination of moral behavior that is motivated by both self- and other-regarding reasons in the context of Kant's moral philosophy, see Herman (1981). Relatedly, see Sugden's (2018: 277–281) discussion of the notion of 'community of advantage' as part of his theory of normative economics.

contractualism, these two tasks are closely related because moral conventionalism and moral contractualism assume that agents share similar moral ideals as traditionally conceived as a starting point for the justification of moral rules, and thus moral theory must determine rules that most closely match the agents' particular moral ideals. If these rules are determined by adequate justificatory procedures that, despite idealization, are justifiable to all current members of society for the domain for which the rules are valid, then all members of society have reasons to follow the rules if others do so too.[13] For moral contractarianism, by contrast, the task is more difficult because this approach does not assume a traditional moral basis as a starting point for the justification of moral rules. Instead, it aims to derive moral rules as a "rational constraint from the non-moral premisses of rational choice" (Gauthier 1986: 4).

The aim of moral contractarianism to derive moral conclusions on the grounds of nonmoral assumptions as traditionally conceived does not necessarily defy the logic of 'is–ought,' because moral contractarianism does not attempt to derive normative conclusions from entirely nonnormatived assumptions.[14] Instead, moral contractarianism aims to derive its conclusions based on a combination of normative assumptions (especially assumptions about the rationality of agents) and empirical assumptions about human nature and the conditions of social cooperation. If successful, however, moral contractarianism does not rely on substantial moral assumptions as traditionally conceived, although the approach does not rule out that some or all members of society may hold traditional moral ideals. Moral contractarianism considers morality to be purely instrumental. It assumes that agents follow moral rules because the rules allow the agents to best fulfill their overarching goals.

Morality, Self-interest, and Instrumental Rationality

To state this feature of moral contractarianism more precisely, moral contractarianism assumes that agents are instrumentally rational. Instrumentally rational agents are goal and outcome oriented and aim to satisfy their interests maximally. Nevertheless, instrumental rationality does not entail the assumption that agents must be self-interested. Different theories of moral contractarianism make different assumptions about agents' motivations.

As discussed, Hobbes *does* assume that agents are rational egoists who primarily pursue their own good. Nevertheless, Hobbes does not defend

[13] For discussion of idealization in the context of normative theory building, in particular public reason theory, see Vallier (forthcoming).

[14] See Kraus (1993: 28–31, 38–39, 319) for discussion of such potential misreading of the project of moral contractarianism.

psychological egoism.[15] Although Hobbes considers self-interest to be the dominant human motivation, he does not assume that all human behavior is selfish. Instead, he allows for other-regarding motivations. Kavka (1983: 293, 1986: 64–80) calls this the assumption of limited altruism or predominant egoism. For my own theory of moral contractarianism, to model the worst type of conflict that may arise among agents, I include negative tuistic interests that may stem from motives such as hate, spite, or envy, and merely exclude positive tuistic interests that express genuine concern for one's conflict partners (Moehler 2018a). For his moral theory, Gauthier (1986: 87) assumes nontuism, and thus assumes that agents, independent of their specific motivations, do not take an interest in the interests of those with whom they cooperate.[16]

Despite the fact that moral contractarianism may constrain in certain ways the motivations and content of agents' interests that form the basis for the justification of moral rules, moral contractarianism in its most general form is, from a traditional moral perspective, morally neutral. Methodologically, moral contractarianism considers moral rules merely as a means that allows agents to reach their goals, independent of the agents' precise reasons for action and the consideration that such means–end reasoning may not always be optimal for all types of moral interaction.[17] This feature of moral contractarianism, that is, to consider moral rules merely as a means that allows agents to reach their individual goals in moral interactions where instrumental reasoning applies, renders the approach well suited for capturing moral diversity.

Moral Diversity

Moral diversity is a common feature of modern societies and a central topic in contemporary moral philosophy.[18] In an interdependent global world, societies must cope with a host of value and value-neglecting tendencies inside and outside of their territories. If disagreement among agents that stems from their diverse moral viewpoints is stark, then such diversity may not always serve as an engine for social progress but as a source for destructive action. In morally diverse societies, especially under the condition of 'deep moral diversity,' which assumes that society is populated by liberal moral agents, nonliberal moral agents, and nonmoral agents as traditionally conceived, the ideal of

[15] See Hampton (1986: 19–24).

[16] For discussion of the notion of 'nontuism' that is relevant especially in the context of economic theory, see Wicksteed (1933: Vol. 1, 180). Gauthier (1987: 212) erroneously assumes that the assumption of nontuism models the worst case scenario from a traditional moral perspective. For clarification of this point, see Morris (1988: 135).

[17] For discussion of this point, see Schmidtz (1995: 19–22).

[18] See Gaus (2011, 2016), Bruner (2015), Thrasher and Vallier (2015), Muldoon (2016), Moehler (2018a), and Müller (2019).

a fully just society as judged from the perspectives of all members of society is unattainable and the topic of moral diversity is not only theoretically but also practically relevant.

Conceptually, moral contractarianism can accommodate the assumption of deep moral diversity, because as long as agents fulfill certain minimal demands of reasoning, moral contractarianism does not exclude anyone's interests for the justification of moral rules. More strongly, moral contractarianism ensures that the views of all members of society are considered equally for the justification of moral rules and, in this sense, ensures the expression of the greatest diversity of moral views as traditionally conceived or lack thereof. Moral contractarianism employs these conditions of autonomy and equality on purely instrumental grounds because if agents were simply to impose their views on others, then, under the assumption of natural equality among agents, the moral rules derived could not ensure mutually beneficial cooperation and would not be stable. The notions of autonomy and equality that underlie the justification of moral rules according to moral contractarianism do not represent moral assumptions as traditionally conceived. Instead, the assumptions are justified instrumentally.

Moreover, moral contractarianism, under the constraint that agents share an overarching goal, is maximally inclusive of different views about moral truth. The approach includes the moral realist who believes that there is moral truth and claims to know it as well as the moral skeptic who does not believe in morality as traditionally conceived. According to moral contractarianism, the moral realist and moral skeptic may try to convince others of their moral views as traditionally conceived or lack thereof. Doing so is a natural part of moral development and not objectionable per se, as long as the agents do not merely impose their views on others but offer reasons that convince others to accept their views from their own perspectives. If, as a result of such processes, convergence arises among agents and they agree on similar moral conclusions as traditionally conceived, then moral theory enters the domain of traditional morality that, in the contractarian tradition, is captured by moral conventionalism and moral contractualism. If deep moral diversity remains and agents do not find a moral ground as traditionally conceived, shared or not, as a starting point for the justification of moral rules, then moral contractarianism represents the most appropriate approach to morality, if the agents share an overarching goal.

Moral and Political Contractarianism

In addition to clarifying the scope of contractarianism as a moral theory, it is important to note that, analytically, the position of contractarianism can be divided into 'moral contractarianism' and 'political contractarianism.' Moral

contractarianism justifies standards of behavior that guide the moral interactions among agents. It justifies rules that apply to agents for the regulation of morally relevant mutually beneficial social interactions. Moral contractarianism deals with the origin and legitimate content of 'social morality' (as opposed to personal morality). Political contractarianism, by contrast, justifies the state as an institution, its power, and its forms of government. Although the distinction between moral contractarianism and political contractarianism is often blurred, this Element focuses on moral contractarianism. Social morality, however, has at least two important implications for the political institutional level.

First, because contractarianism assumes methodological individualism, the design of the political institutional structure of society must be informed by the demands of social morality. According to contractarianism, moral philosophy is fundamental to political philosophy. As Nozick (1974: 6) puts it, "moral philosophy sets the background for, and the boundaries of, political philosophy. What persons may and may not do to one another limits what they may do through the apparatus of a state, or do to establish such an apparatus."[19] According to contractarianism, the task of political philosophy is to translate the demands of moral philosophy from the individual level to the political institutional level. This is not a trivial task because typically there are various ways to institutionalize the demands of social morality, depending on the particular members of society and their specific social, cultural, and historical circumstances. Hobbes's theory follows this logic. Hobbes first develops a moral theory and, as a result of his pessimistic conclusions, justifies monarchy as the only viable political institutional structure. However, Hobbes's argument for monarchy is much contested, and absolute sovereignty is not the only viable solution from a broadly Hobbesian perspective. Muldoon (2016), for example, defends a social contract theory that combines elements of Hobbesian moral theory with Mill's (1859) notion of 'experiments in living.' Muldoon's theory suggests an open-ended process of discovery and experimentation that, in contrast to Hobbes's static approach, relies on a dynamic notion of morality and justice that, as a result of institutional change and reform, promises a stable political order for morally diverse societies.

Second, because moral contractarianism does not demand that agents possess a moral sense or are motivated by traditional moral reasons (as moral conventionalism and moral contractualism demand), the normative demands of moral contractarianism are typically assumed to be enforced by the regulating

[19] Nozick defends a rights-based libertarian theory. However, as Narveson (2001: 154–184) and Thrasher (2017) argue, albeit in different ways, libertarianism can also be viewed as having its roots in contractarianism, which provides libertarian theory with a foundation in the rational choice of individual agents.

institutions of society. According to moral contractarianism, the institutional structures of society serve as a means for rational agents that provide the agents with incentives to do in the short term what is best for them in the long term from their own perspectives. Moreover, such institutional structures are essential to overcome collective action problems that arise if, because of moral diversity, agents are not closely bound together by an entrenched moral fabric (as assumed by moral conventionalism and moral contractualism). For moral contractarianism, the institutional structures of society play a central role in the enforcement of morality to the benefit of all members of society. According to Hobbes, "morality and justice are effectively and lastingly realizable only within the State" (Kavka 1986: 452).

1.3 Core Challenges for Moral Contractarianism

Hobbes: Assurance and Compliance

The main challenges for moral contractarianism stem from its core features and are already present in Hobbes's (1651) moral theory.[20] Hobbes's moral theory is based on a hypothetical state of nature that assumes, apart from rough natural equality among agents, that agents pursue their own good and are free to do what is necessary to preserve their lives. In Hobbes's state of nature, agents have threat capacities against each other that, together with the assumptions of competition, diffidence, glory-seeking, and a desire for power under scarce resources, lead to severe conflict.[21] In this situation, agents recognize that striking first in the struggle for life may provide them with the necessary advantage for survival. Such preemptive action makes the state of nature a war of all against all that destroys most productive efforts and runs counter to agents' desire for self-preservation.[22] According to Hobbes, "the actions which men naturally and reasonably perform in order to secure their ends prove self-defeating" (Gauthier 1969: 17).

In addition to assuming that agents aim to preserve their lives, Hobbes (1651: Part 1, chapters 10 and 11) assumes that agents are prudent and thus have an interest in securing their long-term well-being. According to Hobbes, prudence dictates to rational agents the laws of nature that demand that agents do not

[20] In this Element, I follow the orthodox interpretation of Hobbes's theory that considers Hobbes to defend a purely instrumental approach to morality. For nonorthodox interpretations of Hobbes's theory, see Lloyd (1992, 2009) and Abizadeh (2018), and for discussion of the differences between orthodox and nonorthodox interpretations of Hobbes's theory, see Gaus (2013).

[21] For discussion of glory-seeking in Hobbes's theory and its potential effects on social cooperation, see Gaus (2018a).

[22] Kavka (1983: 297, 1986: 83–125) questions the rationale for preemptive action in the state of nature. For a nuanced discussion of Hobbes's notion of self-preservation, see Kavka (1986: 315–337).

perform actions that are destructive to their lives or take away the means to preserve their lives. The first law of nature demands that agents seek peace because, under the circumstances described, agents expect that peaceful long-term cooperation is most beneficial to them. According to Hobbes, peaceful long-term cooperation is an instrumental good, and the formation of society is a necessary means to this end. The second law of nature states that society can be established only if all agents lay down their rights of nature to do what they consider to be necessary for survival and transfer these rights to a common authority that is not part of the society to be formed. According to Hobbes, only if agents give up their rights of nature and transfer them to an external authority with unlimited and undivided power can society be established and maintained. The third law of nature demands that agents keep their contracts once they are binding, which is a necessary condition for ensuring peaceful long-term cooperation in the postnatural state.

According to Hobbes, two problems of collective action must be addressed to establish society and maintain peace: the problems of assurance and compliance. Concerning the former, if agents maintain their rights of nature, then they will remain in the state of nature and face a war of all against all that is suboptimal for everyone. Nevertheless, in the state of nature agents will cooperate with others only if they can be sure that others will cooperate too, especially by transferring their rights of nature to an external authority that is strong enough to enforce any subsequent agreements. According to Hobbes (1651: Part 2, chapter 17), only if a common power is in place that is strong enough to enforce agreements are such agreements binding and a valid social contract is established. As such, in the state of nature, agents' behavior is conditional on whether the agents can trust that others will cooperate in the establishment of society. Hobbes concludes that, owing to the fierce conditions in the state of nature, agents cannot generally trust that others will cooperate because other agents are potential enemies in the struggle for survival.[23] According to Hobbes, only an external sovereign can solve the problem of assurance in the state of nature by replacing the lack of interpersonal trust among agents with institutional trust. Hobbes's solution to the problem of assurance is captured by his second law of nature.

Once society is formed, the problem of compliance arises that is expressed by Hobbes's discussion of the 'Foole.' The Foole recognizes that, in the long term, being part of society is more beneficial than staying in the state of nature. In the short term, however, the Foole defects from moral standards whenever doing so is beneficial. The Foole, like other members of society, primarily aims to satisfy

[23] For support of this conclusion, see Kavka (1986: 126–178).

her own interests and, without possessing commitment power,[24] decides in each instance about the most beneficial behavior for herself. Such behavior, however, is short sighted. The Foole does not fully consider the negative effects of defecting or assumes that such behavior will remain undetected. The Foole believes that she can have it both ways: secure the benefits of peaceful long-term cooperation and the short-term gains from exploiting others. To solve this problem, Hobbes argues that the sovereign must threaten agents with severe sanctions for defective behavior, including potentially removing agents from society or taking their lives, in order to ensure peace. Hobbes's solution to the problem of compliance is captured by his third law of nature.

One central problem with Hobbes's solution to the problems of assurance and compliance is that Hobbes defends an external sovereign with full legislative, executive, and judicial powers, including the right to punish rule violators and taking their lives. However, if the sovereign has the right to take the lives of rule violators, then the state of nature is not necessarily the worst possible state for such agents any longer. In this case, the sovereign's command conflicts with the agents' right to defend their lives, which remains intact in the postnatural state. As a result, it may be rational for agents to rebel against the sovereign.[25] Also, it is not clear how far agents' right to self-defense extends, especially whether this right protects only the physical integrity of agents or everything else that agents consider to be essential for maintaining their lives. If the latter is true, as Hampton (1986: 197–207) points out, then agents may generally be free to judge whether or not to obey the sovereign's commands.

In more general terms, an essential problem with Hobbes's moral theory is that the sovereign not only has executive and judicial powers but also full legislative power. Hampton (1986: 3) calls this type of social contract an 'alienation contract' because it de facto renders meaningful individual moral agency impossible.[26] Hobbes's moral theory demands that agents give up their autonomy to approve of the specific moral rules by which they are governed. In certain situations of social interaction, this demand may contradict the core tenet of moral contractarianism and license moral rules that, strictly speaking, cannot be justified to all current members of society from their own (prudential) perspectives. Although rational agents may generally outsource the enforcement and adjudication of moral rules to an external authority, they would not

[24] According to Hampton (1986: 93), Hobbes does not attribute commitment power to agents in *Leviathan*.

[25] For discussion of the right to rebellion in Hobbes's theory, see Kavka (1986: 433–436).

[26] Hampton (1986: 208–266) argues that alienation contracts generally fail in social contract theory. Nevertheless, she argues that a Hobbesian argument for absolute sovereignty can be constructed in the form of a strong agency contract, even if Hobbes himself would object to such an argument.

give up their autonomy to approve of such rules.[27] To avoid this problem, later theories of moral contractarianism typically internalize Hobbes's external sovereign by requiring that the members of society themselves decide about the specific moral rules by which they are governed. Such internalization of Hobbes's external sovereign ensures the autonomy of agents and that moral rules are justified that are mutually advantageous from the perspectives of all members of society. Internalizing Hobbes's external sovereign helps to ensure that moral reasoning and prudential reasoning are closely aligned with each other.[28]

Prudence and Morality

Even if moral contractarianism can ensure that moral reasoning and prudential reasoning are closely aligned with each other, the question arises as to whether or not the type of prudential morality defended by moral contractarianism qualifies as a genuine form of morality. More broadly, this charge is expressed by the 'dilemma of contractarian ethics,' according to which contractarian moral theory is either circular, if it relies on moral assumptions as traditionally conceived (as assumed by moral conventionalism and moral contractualism), or is irrelevant as moral theory, if it does not rely on such assumptions for the justification of morality (as assumed by moral contractarianism).[29] To derive moral conclusions as traditionally conceived, moral assumptions, in the traditional understanding of morality, must be made. However, the main goal of moral contractarianism is to avoid such assumptions for the justification of moral rules, and thus moral contractarianism must show that the type of morality that it defends represents a genuine form of morality. This task entails several dimensions.

The first dimension concerns moral motivation and is captured by the 'wrong kind of reasons objection.'[30] Independent of the specific moral rules justified, according to moral contractarianism agents do not need to be motivated to follow moral rules on the basis of what are traditionally conceived to be moral

[27] See Kavka (1984, 1986: 224–236). [28] In this context, see also Schmidtz (1995: xi).

[29] See Park (1992: 11) for one possible statement of this dilemma.

[30] More generally, this objection is expressed by Prichard's (1912) dilemma that suggests that there is no good reason to act morally. Prichard argues that, if one refers to moral reasons, such as altruistic reasons, then one presupposes the persuasive force of morality, which is circular. If one refers to nonmoral reasons, such as self-interest, then one provides the wrong kind of reasons to act morally. In the contractarian tradition, Scanlon's (1998: 147–158) contractualist moral theory directly responds to Prichard's dilemma by suggesting a third type of reason to act morally that, according to Scanlon, is neither a moral nor nonmoral reason, namely, the desire to justify one's actions to others that stems from an action's wrongness. In this context, see also Darwall (2006: 17).

reasons (although some agents may be motivated by such reasons) as long as agents' reasons for action (even if the reasons are selfish) promote an overarching goal among the agents. More fundamentally, the problem concerns the priority of prudence over traditional morality that is assumed by moral contractarianism. As Gaus (2019: 110–111) expresses this concern succinctly, "many people think a moral person is committed to imprudent actions: running the risk of serious conflict and the breakdown of critical features of cooperation is precisely what their moral commitments require." This feature of moral contractarianism does not imply that this approach necessarily offers the wrong kind of reasons for acting morally or that it offers no such reasons at all. Nevertheless, moral contractarianism must show that the types of reasons offered by this approach are legitimate moral reasons.

The second dimension concerns the role and nature of moral emotions. Because moral contractarianism employs instrumental rationality to justify moral rules and explain moral behavior, it arguably does not sufficiently consider the importance of moral emotions, such as anger, indignation, guilt, blame, and resentment, and more generally, the moral psychology of agents. Moreover, even if moral contractarianism were to consider such emotions, it may misconstrue their nature as primarily self-directed.[31] This feature of moral contractarianism, if true, may have implications with regard to the notions of moral accountability and sanctions that are employed by moral contractarianism, in particular because traditional morality typically assumes that agents internalize moral rules at least partially as a result of their moral emotions, and thus moral standards as traditionally conceived typically are assumed to be at least partially self-enforcing.

The third dimension concerns the content of moral rules that are justified by moral contractarianism. To be considered legitimate moral conclusions, the moral rules justified by moral contractarianism must resemble to some extent traditional moral rules, in particular by expressing some of their underlying moral ideals, such as the moral ideals of autonomy, equality, impartiality, and reciprocity, even if the moral rules are justified entirely on instrumental grounds. For his moral theory, Gauthier (1986: 4, 7) stresses the importance of this condition, especially with regard to the moral ideal of impartiality.[32] As Morris (2013: 598) puts it with reference to Gauthier:

> Principles constrain one's actions, and we have reasons to be so constrained. If Gauthier has established this, it would be a major feat. But we would still need to ascertain whether the principles in question and the principled action are moral . . . The strategy seems to be a functional one: the principles and

[31] For such criticism, see Southwood (2008: 185–186, 2010: 34–42).

[32] Barry (1989: 364) considers the moral ideal of impartiality to be "an original principle in human nature and one that develops under the normal conditions of human life."

> dispositions resemble familiar moral ones sufficiently that we can identify them as such. Impartiality he thinks is a defining feature of morality, and it is shared by these principles and dispositions. This and other features allow us to conclude that what has been shown to be rational is in fact a genuine morality.

The fourth dimension concerns the question of moral standing. According to Morris (1991, 1998b, 2011), agents have moral standing if they are owed moral consideration by others. Typically, it is assumed that, according to the mutual advantage approach that underlies moral contractarianism, agents have moral standing only if others can benefit from cooperation with them.[33] Gauthier's (1986: 18, 268) moral theory, for example, includes only agents who are potential constructive net-contributors to society. Such agents have, as Morris puts it, 'primary moral standing.' If agents cannot benefit from each other in cooperation, then they may acquire 'secondary moral standing,' if agents with primary moral standing take an interest in the interests of those who do not have primary moral standing. Agents with primary moral standing may, for example, care for weaker members of society who otherwise would not be protected, such as parents who care for their children. Nevertheless, moral contractarianism may still fail to protect some human agents, such as the elderly, the disabled, and agents with permanent care needs, as well as nonhuman agents that, according to the traditional understanding of morality, should be protected.[34] If so, then moral contractarianism must offer reasons for such omission.

Counterfactual Contractarianism

Finally, if Hobbes's external sovereign is rejected as the ultimate enforcer of morality, then moral contractarianism faces a potential problem of moral authority, because moral contractarianism is a constructivist moral theory that, focusing on the most relevant feature of moral constructivism for my discussion, justifies moral rules based on an idealized and hypothetical process of rational choice and agreement.[35] This feature of moral contractarianism has two implications with regard to moral authority. First, for the moral rules justified by moral contractarianism to have normative force for agents, the justificatory reasons offered by constructivist procedures must be rationally explicable for real-world agents. If this condition is not met, then real-world

[33] See Stark (2009), for example.

[34] For such criticism see Nussbaum (2006: 14–22), although Nussbaum's argument focuses on political theory and the considerations of political theory are not necessarily identical to the considerations of moral theory.

[35] For a general discussion of the features of moral constructivism, see Bagnoli (2017).

agents do not necessarily have sufficient reasons to comply with the moral rules derived by moral contractarianism.

Second, even if the constructivist procedures that are employed by moral contractarianism provide sufficient reasons for real-world agents to comply with the rules justified by these procedures, it is rational for agents to comply with such rules only if other agents follow the rules too. In practice, this condition is guaranteed only if agents actually agree with the rules, because only then do the regulating institutions of society have the authority to enforce the rules. Although this problem may not arise for all constructivist theories that rely on hypothetical agreement, it arises for theories of moral contractarianism that, despite outsourcing the enforcement of moral rules to an external authority, aim to ensure the autonomy of agents, and thereby meaningful individual moral agency. For such theories, counterfactual agreement is not necessarily sufficient. As Dworkin (1989: 18) remarks, "a hypothetical contract is not simply a pale form of an actual contract; it is no contract at all."

In the light of these challenges with regard to moral contractarianism, the remainder of this Element is structured as follows. Section 2 discusses essential methodological considerations, especially concerning the application of rational choice theory to moral philosophy, that are important for understanding the nature of contemporary theories of moral contractarianism. The section also clarifies the multistage structure that typically underlies social contract theory and analyzes, from a game-theoretic perspective, the collective action problems that moral contractarianism must overcome to establish social moral order. Section 3 discusses some of the specific moral principles, in particular bargaining principles, that have been defended as distributive principles by some of the most significant theories of moral contractarianism to maintain social moral order, and the problems of compliance associated with these principles. Finally, and more fundamentally, the section discusses the justification of such principles via suitably tailored moral decision situations. Section 4 integrates the previous discussion and defends a new type of contractarian moral theory that overcomes many of the problems associated with moral contractarianism that are discussed in this Element. This 'multilevel social contract theory' reconciles moral conventionalism, moral contractualism, and moral contractarianism, and defines the appropriate place of moral contractarianism in moral theory. I argue that this triple theory offers the strongest possible defense of moral contractarianism as a moral theory, especially for morally diverse societies.

2 The State of Nature and Social Moral Order

2.1 Rational Choice Theory and Moral Philosophy

Normative Rational Choice Theory

As discussed in the previous section, moral contractarianism justifies moral rules by an idealized and hypothetical process of rational choice and agreement that, if successful, does not rely on moral assumptions as traditionally conceived. Methodologically, this feature brings moral contractarianism close to rational choice theory that has been employed extensively in the social sciences, especially in economic theory, to explain and predict human behavior.[36] Orthodox rational choice theory assumes that agents are instrumentally rational, and thus are goal and outcome oriented (as discussed in Section 1.2). More specifically, in the context of economic theory, agents are assumed to prefer more goods over fewer goods and exhibit preference orderings that can be represented by von Neumann–Morgenstern utility functions, although the precise axiomatic formalization of expected utility theory is typically not decisive from the perspective of moral theory.[37] As a result, agents are assumed to behave as if they were to maximize their expected individual utility. In addition, orthodox rational choice theory assumes that agents are opportunistic case-by-case decision-makers. The theory assumes that agents will always perform the actions that, in a particular instance, will produce the highest expected individual utility for them within their forward-looking perspectives. According to orthodox rational choice theory, agents do not have commitment power. As a consequence, to explain rule-guided behavior, such behavior must be most beneficial for agents in each instance for which the rules prescribe behavior.

As a result of its often descriptively inaccurate predictions about human behavior, orthodox rational choice theory has been criticized as a behavioral model.[38] The various criticisms especially in the context of economic theory have led to the development of behavioral economics that aims to increase the predictive power of rational choice theory by enriching the theory with more realistic empirical and psychological assumptions about human agency. Nevertheless, as Sugden (2018) indicates, most behavioral economists still consider the demands of orthodox rational choice theory to be the ideal and

[36] See Becker (1976) and Gaus (2007: 19–27).

[37] For an axiomatic formulation of von Neumann and Morgenstern's theory, see Luce and Raiffa (1985: 23). For a criticism of orthodox expected utility theory from a normative perspective, see Hampton (1998a: 117) and Buchak (2013). For a criticism of Buchak's argument and defense of expected utility theory, see Thoma and Weisberg (2017), Thoma (2019), and Thoma and Weisberg (forthcoming).

[38] See Sen (1977, 1985, 1997, 2005), Loomes and Sugden (1982, 1983, 1984, 1987), McClennen (1990), and Broome (1991).

assume that psychological processes merely interfere with the demands of ideal theory, instead of moving entirely away from idealized choice and considering actual human psychology as the standard for choice.[39]

In general, when rational choice theory is applied to moral theory, the aim is neither to explain or predict human behavior nor to determine the optimal design of economic institutions. Instead, it is to prescribe the best behavior for rational agents in the moral domain, although maximizing behavior is not necessarily adequate for all domains of morality (as indicated in Section 1.2). Specifically, in the context of moral contractarianism, rational choice theory aims to determine the most beneficial behavior for rational agents in morally relevant types of social interaction in which moral reasoning is reduced to instrumental reasoning, because agents' real-world (moral) behavior, which often does not satisfy the demands of orthodox rational choice theory (as the findings of behavioral economics show), leads to irresolvable conflicts. If rational choice theory is employed normatively within the ideal conceptual space of moral theory and with such restricted scope, then rational choice theory can avoid most of its behavioral criticisms.

Nevertheless, for real-world agents to follow the moral conclusions reached by such 'orthodox rational choice contractarianism,' orthodox rational choice contractarianism must make adequate assumptions about the rationality of agents that allow real-world agents to identify with the ideal model rationality. Although the form of rationality employed by orthodox rational choice contractarianism imposes only weak normative constraints for the justification of moral rules, it defines standards of reasoning that real-world agents must fulfill and, as a result, considers some real-world agents to be potentially irrational. Although orthodox rational choice contractarianism employs only a minimal form of rationality that seems essential for agents to survive in this world, some real-world agents may lack this capacity of reasoning, and thus may not fully comprehend the justification of the moral rules derived by this approach. However, this problem arises for all rationalist moral theories and, by assuming only a minimal form of reasoning, orthodox rational choice contractarianism minimizes this problem.

In addition to the standards of rationality assumed by orthodox rational choice contractarianism, the constructivist procedures employed for the

[39] Sugden (2018) argues that, although the results of behavioral economics pose a problem for neoclassical (welfare) economics, they do not necessarily pose a problem for classical liberal economics that relies less on the notion of rational choice than voluntary, mutually beneficial exchange. Sugden's account of normative economics defends liberal market order that expands agents' opportunity sets, independent of whether or not agents' choices fulfill the demands of orthodox rational choice theory, assuming that agents can generally fulfill their interests better if they have more options rather than fewer.

justification of moral rules must make realistic assumptions about real-world agents and the empirical conditions under which they live. The constructivist justificatory procedures must accurately reflect the real-world agents' abilities (such as their ability of perspective-taking), knowledge, and other relevant circumstances in order for the moral rules justified by such procedures to survive assessment when the agents consider their full sets of reasons in the real world, and thus for the agents to have sufficient reasons to follow the rules in the real world. Also, to be action-guiding in the real world, the moral rules must fulfill the demands of the ought-implies-can principle (see Kant 1797: AK 6:380). The rules cannot demand behavior of agents that they generally cannot perform given their abilities. I will return to some of these considerations more systematically at the end of Section 2.3 in the context of moral theory building.

Rational Choice Theory and the Boundaries of Morality

The employment of orthodox rational choice theory in the context of moral contractarianism adds rigor to moral analysis. Nevertheless, one must be careful not to over-formalize moral theory. Moral theory is an inexact science and the boundaries of rational choice theory in its general standard formulation are not necessarily identical to the boundaries of morality, not even in the case of orthodox rational choice contractarianism, although the former are not necessarily inconsistent with the latter. As such, fully formalized models that are common in the natural sciences and some of the social sciences do not necessarily allow expression of all relevant moral considerations (as my discussion of the 'peace game' in Section 4.1 clarifies) and may even distract from the central moral questions (as my discussion of the Rawls–Harsanyi dispute in Section 3.3 shows).

Moreover, theories of orthodox rational choice contractarianism can employ rational choice theory in different ways. Although orthodox rational choice theory is often used to justify new moral principles, sometimes the theory is used merely to reveal or rationalize moral principles that agents already intuitively hold. Further, orthodox rational choice theory has been applied not only in the context of moral contractarianism but also in the context of moral contractualism, in particular in the context of Rawls's contractualist theory. In this case, rational choice theory is not applied to situations of social interaction in which moral reasoning is fully reduced to instrumental reasoning and, as a result, not all moral considerations as traditionally conceived are necessarily captured by rational choice theory, as my discussion of the Rawls–Harsanyi dispute will show. These considerations must be kept in mind if orthodox

rational choice theory, or some of its concepts, is to be used as a heuristic to guide moral considerations.

Normative Decision Theory and Game Theory

Orthodox rational choice theory includes normative decision theory and game theory. The goal of normative decision theory is to determine the best choices for rational agents, given the agents' beliefs and values, typically under the assumption that agents are fully informed and ideally rational. Normative decision theory focuses on situations in which agents' choices are independent of other agents' choices and, in this sense, the agents play against nature. Game theory, by contrast, addresses interactions in which agents' choices affect the choices of other agents. In such strategic interactions, the goal of rational agents is to choose so that their choices together with the choices of all other rational agents lead to the best outcomes for them. For such interdependent choices, common knowledge of rationality is required that assumes that all players of a game know how all other players think, and all players know that all other players know that they know how they think, and so on.

For the purpose of orthodox rational choice contractarianism, noncooperative game theory, which assumes that agents cannot (verbally or nonverbally) communicate with each other to make binding agreements on outcomes or form coalitions, is especially relevant. Although coalition formation may be a realistic assumption for many types of social interaction, methodologically the assumption is inapt in the context of moral theory that is concerned with the behavior of individual agents and not with the behavior of groups of agents as assumed by political theory. This methodological consideration does not imply that agents may not increase their strength "by secret machination, or by confederacy with others" (Hobbes 1651: Part 1, chapter 13). However, such coalition formation typically does not add anything essential to moral analysis. For the purpose of moral analysis (although not necessarily for the purpose of social scientific analysis more generally), such dynamics generally can be broken down into two-party situations in which the parties play a noncooperative game with each other.[40]

Also, in the context of orthodox moral contractarianism, non-zero-sum games are the most relevant games because they assume that mutually beneficial cooperation is possible. Non-zero-sum games are divided into two subclasses: games of pure coordination and mixed-motive games. The former types

[40] One may also employ a hybrid game in which agents can form coalitions in a cooperative game and then play in a noncooperative fashion with each other. See Kavka (1986: 189) for a slightly different reason to neglect coalition formation in the context of moral theory.

of games capture situations in which agents' interests completely converge, such as driving on the right side or the left side of the road, and agents must tacitly agree on a focal point to coordinate a mutually beneficial outcome (I return to this type of game in Section 4.1 in the context of my discussion of the 'peace game').[41] The latter games capture situations in which there is tension between agents' cooperative and competitive motives. In moral theory, the most widely discussed mixed-motive games are the prisoner's dilemma game, assurance game, assurance dilemma game, battle of the sexes game, and chicken game.[42] Typically, moral philosophers employ game theory to analyze the structure of specific morally relevant social interactions, identify the game that best matches these structures, and then formulate (moral) demands that solve the game or transform the game into another game that allows mutually beneficial cooperation.

3.2 The Rationale for Rule-Following

Gauthier's Notion of Constrained Maximization

As discussed in Section 2.1, according to orthodox rational choice theory, rational agents are opportunistic case-by-case decision-makers. As such, after society is established, agents will break the agreed moral rules whenever the costs for doing so are lower than the expected gains from such defective behavior. If, however, too many agents defect in the short term, then society

[60] For support of this claim, see Barry (1989: 22–24), Binmore (1993, 1994: 80–84, 1998: 77–95, 2005: 26), Skyrms (1996: 107), Moehler (2009b: 452–454, 2013: 36, 2016: 118, 2018a: 56), and Muldoon (2016: 79–80).

[61] For discussion of the notion of fairness (impartiality) as traditionally conceived that is implicit in Gauthier's bargaining theory, see Copp (1991: 209–212).

will break down and, as judged from their own perspectives, the agents will be worse off in the long term. To address this problem, Hobbes argues for the institution of the absolute sovereign that excludes free riders from cooperation or takes their lives. If such severe sanctions are imposed, then even opportunistic case-by-case decision-makers, such as Hobbes's Foole, will follow the established moral rules if free-riding is likely to be detected now or in the future (as discussed in Section 2.3). In this case, rule-guided behavior will be beneficial for rational agents in each instance for which the rules prescribe behavior.

However, external enforcement of moral rules is costly and, in the following, I consider Gauthier's specific attempt to solve the problem of compliance with regard to his bargaining principle without instituting an external sovereign or entirely abstracting from this problem, as bargaining theory typically does. To solve the problem of compliance, Gauthier (1986: 157–189) replaces the orthodox concept of rationality with an alternative notion of rational choice that is captured by his notion of 'constrained maximization.' In similar fashion to Hume (as discussed in Section 1.1), Gauthier argues that, for certain strategic choices, it may be rational for agents to develop 'dispositions' to follow moral rules because such rule-following dispositions may increase the agents' gains from cooperation by allowing the realization of additional gains that otherwise would be unattainable. Once agents have developed such rule-following dispositions that constrain their choices, the agents no longer need to decide what to do in each instance because the agents are bound to follow the rules even if such rule-guided behavior pays less in particular instances than case-by-case decision-making would do.

Gauthier's notion of constrained maximization deviates from the case-by-case maximization assumed by orthodox rational choice theory in favor of an account of genuine rule-following. De facto, Gauthier assumes that agents develop commitment power for moral rules that the agents consider to be beneficial for them in the long term. As Gauthier (1993: 186) clarifies, "a *constrained maximizer* . . . is someone who takes her reasons for acting, not only directly from the utilities of possible outcomes she may bring about, but also from her plans and commitments." That is, to solve the problem of compliance, Gauthier introduces a stronger notion of rationality that, as Gaus (2011: 76–86) clarifies, is incompatible with orthodox rational choice theory.

However, even if agents are constrained maximizers, it is generally rational for such agents to cooperate only with agents who will also cooperate, such as other constrained maximizers, and not with agents who will exploit their cooperative behavior. To solve this problem of assurance that arises for Gauthier's theory in the postnatural state, Gauthier (1986: 174) introduces the assumption of 'translucency.' Gauthier assumes that, in moral interactions,

rational agents typically can guess whether other agents will cooperate or defect, which, if successful, allows constrained maximizers to cooperate only with those who will also cooperate. It is unclear, however, to what extent Gauthier's assumption of translucency is empirically sound. Although the assumption may have some empirical plausibility for small groups in which agents know each other well or they know other agents who know their cooperative partners well, the assumption does not necessarily hold for large anonymous groups and social interactions where agents cannot freely choose their cooperative partners.[62]

Gauthier's Notion of Agreed Pareto-Optimization

The most recent version of Gauthier's (2013) moral theory does not seem to invoke the assumption of translucency. However, from a traditional moral perspective, the theory relies on an even stronger account of practical reasoning than his notion of constrained maximization that Gauthier calls 'agreed Pareto-optimization.' Gauthier's notion of agreed Pareto-optimization rejects the ideal of expected individual utility maximization for non-zero-sum interactions and, in doing so, rejects orthodox rational choice theory that demands that agents settle on equilibrium outcomes that maximize their expected individual utility. To clarify, orthodox rational choice theory embraces two conditions that rational outcomes must fulfill, namely, the Pareto-optimality condition and the equilibrium condition. The former condition, in its strong formulation, demands that outcomes are such that they do not allow improving the situation of one agent without worsening the situation of another agent. The latter condition demands that outcomes are stable in that, ceteris paribus, no agent has incentive to deviate from the selected outcomes.

According to orthodox rational choice theory, to ensure a stable outcome, the equilibrium condition takes priority over the Pareto-optimality condition if the two conditions conflict. This feature of orthodox rational choice theory is the reason that rational agents end up with suboptimal outcomes in situations of social interaction that have the structure of a one-shot PD game, because orthodox rational choice theory demands that agents deviate from distributional outcomes if such (unilateral) defection allows agents to increase their expected gains. In the absence of external enforcement mechanisms, orthodox rational choice theory does not guarantee that the most beneficial Pareto-optimal outcome is always selected.

[62] For a review of some of the experimental literature related to Gauthier's notion of translucency and qualified support for it, see Timmerman (2014: 147–173).

Gauthier's (2013: 605, 609) account of agreed Pareto-optimization avoids this problem because it rejects the ideal of expected individual utility maximization for morally relevant types of non-zero-sum interactions and, for such interactions, prioritizes the Pareto-optimality condition over the equilibrium condition. Also, it assumes that agents do not only maximize their expected individual utility, but also consider their individual good in the same way that they consider the good of others. As a result of this feature of Gauthier's new account of practical reasoning, for morally relevant types of non-zero-sum interactions in which more than one Pareto-optimal outcome exists, agreed Pareto-optimizers will always (even in a one-shot PD game) select the individually and collectively most beneficial Pareto-optimal outcome that is reasonably fair and that may or may not be an equilibrium outcome.

Gauthier's agreed Pareto-optimizers have an interest in realizing reasonably efficient and fair distributional outcomes as traditionally conceived. As such, in theory and practice, if others do so too, agreed Pareto-optimizers will comply with Gauthier's principle of minimax relative concession that Gauthier now calls the 'principle of maximin proportionate gain.' From the perspective of agreed Pareto-optimizers, the principle of maximin proportionate gain ensures reasonably efficient and fair distributional outcomes as traditionally conceived. If all members of society are rational and reasonable in the way assumed by Gauthier, then the problems of assurance and compliance do not arise. Under these conditions, agreed Pareto-optimizers will follow the demands of Gauthier's distributive principle even if they could improve their individual situations by unilateral defection, because agreed Pareto-optimizers, in Gauthier's sense, are cooperators by choice.[63]

In the traditional understanding of morality, Gauthier moralizes his account of practical reasoning by assuming that agents consider others as moral equals and have an interest in reaching fair distributional outcomes as traditionally conceived. In this sense, the latest version of Gauthier's moral theory may be considered to be closer to the position of moral contractualism than moral contractarianism because the theory relies on traditional moral assumptions, especially the assumption of moral equality. In the next section, I discuss Rawls's (1971) theory that is explicitly a contractualist theory. In contrast to Gauthier's moral theory, however, the central analytic device of Rawls's theory of justice, the original position, does not employ a revisionist account of rationality but employs orthodox rational choice theory, which has led to much confusion in the literature.

[63] For a critical discussion of Gauthier's notion of Pareto-optimization, see Neill (2017).