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