


ARTICLE

# Rationality, uncertainty, and unanimity: an epistemic critique of contractarianism

Alexander Schaefer 

University of Arizona, Tucson, AZ, USA  
Email: [schaefer1@email.arizona.edu](mailto:schaefer1@email.arizona.edu)

(Received 14 December 2018; revised 14 November 2019; accepted 28 November 2019; first published online 18 February 2020)

## Abstract

This paper considers contractarianism as a method of justification. The analysis accepts the key tenets of contractarianism: expected utility maximization, unanimity as the criteria of acceptance, and social-scientific uncertainty of modelled agents. In addition to these three features, however, the analysis introduces a fourth feature: a criteria of rational belief formation, viz. Bayesian belief updating. Using a formal model, this paper identifies a decisive objection to contractarian justification. Insofar as contractarian projects approximate the Agreement Model, therefore, they fail to justify their conclusions. Insofar as they fail to approximate the Agreement Model, they must explain which modelling assumption they reject.

**Keywords:** Social contract; contractarianism; game theory; unanimity

What theory of morals can ever serve any useful purpose, unless it can show that all the duties it recommends are truly endorsed in each individual's reason?

(David Gauthier)

## 1. The contractarian schema

Is a hypothetical contract worth the hypothetical paper that it isn't written on? Many have wondered (Dworkin 1973; Hume 1978; Simmons 1979; Schmidtz 1990; Cohen 2009; Huemer 2013). Nevertheless, a major tradition in political theory relies heavily on such contracts as a method of justification. Theorists in this tradition have proposed a wide variety of social contracts, highly diverse in their aims and constructions. While all such contracts invoke agreement or consent to play a fundamental role in the defence of some normative conclusion, they differ as to the object of agreement, the nature of the parties that do the

agreeing, and how such agreement bears on the normative situation of real people who have, in fact, made no such agreement.

A helpful, though imperfect, taxonomy begins with a distinction between *contractualism* and *contractarianism*. These two approaches to social contract theory differ primarily in how they characterize the parties to agreement and in how the modelled agreement bears on the normative situation of real people. For the contractualist, parties are moralized either in motivation or in virtue of their decision-making environment. Thus, in Scanlon's model, parties are all committed to and motivated by the task of 'finding principles for the general regulation of behaviour that others, similarly motivated, could not reasonably reject' (Scanlon 1998: 4).<sup>1</sup> Rawls and Harsanyi, on the other hand, propose a moralized choice scenario in which moral distortions are eliminated by rendering parties ignorant as to their particular position in society. The task of such models is to reveal the principles that real individuals, insofar as they are committed to morality, must accept as properly regulating interpersonal behaviour, that is, social-moral rules or political institutions. Contractarians, on the other hand, assume no common moral perspective. In contractarian models of agreement, 'Moral principles of right are... rules that individuals would prescribe, and attempt to gain acceptance for, from their different individual perspectives, bargaining out of self-interest' (Darwall 2008: 5). The central aim of the contractarian is *not* to show us what moral principles we should follow *given that we are morally motivated agents*. Instead, the contractarian seeks to explain why real individuals, replete with narrow and personal interests and biases, should endorse political or moral rules at all, that is, why political and moral constraints actually further the non-moral interests of those who abide by them.<sup>2</sup> Accordingly, contractarian theorists – such as Thomas Hobbes, David Gauthier and James Buchanan – employ models of agreement to show how, from a position in which they are absent, moral and political rules or institutions emerge from the mere pursuit of personal interests or preferences (Cudd and Eftekhari 2000).

Contractarianism, the focus of this paper, begins by identifying a fundamental predicament faced by all human societies: a disconnect between individual goal pursuit and social welfare.<sup>3</sup> Through cooperation and coordination, human beings can improve their lot. Yet, individual rationality provides no guarantee of efficient coordination, and, surprisingly, it often undermines the prospect of cooperation.<sup>4</sup>

<sup>1</sup>See also Southwood (2009).

<sup>2</sup>Although, in the process of doing so, contractarians usually end up offering some positive guidance as to which moral or political rules should structure interpersonal interactions. See d'Agostino *et al.* (1996). In refraining from providing any specification of the results of his social contract, James Buchanan offers a notable exception: 'I do not try to identify either the 'limits of liberty', or the set of principles that might be used to define such limits ...' (Buchanan 1975: 222).

<sup>3</sup>If all political philosophies begin with a predicament, as Michael Oakshott asserts, then the predicament that contractarians identify concerns a disconnect between individual rationality and social rationality (Oakshott 1965: 221–294).

<sup>4</sup>Hobbes, for instance, has been read as emphasizing both sorts of issue, cooperation and coordination. Interpreters emphasizing cooperation include Kavka (1986: 109) and Gauthier (1969). Hampton (1988: 65),

To escape this predicament, however its details are characterized, requires constraints on individual goal pursuit; it requires both rules and some mechanism of enforcement, whether this be formal or informal, external or internal. Such constraints give rise to two related questions that contractarians seek to answer: (1) What rules or constraints should a society adopt? And (2) Why should I, an individual in that society, willingly abide by such constraints? The task faced by the contractarian is to show that each individual subject to a certain set of rules has sufficient reason to endorse them. The main step in doing so is to show that all individuals would agree to such constraints in an environment where these constraints were absent. Hypothetical unanimous agreement, the contractarian asserts, demonstrates that a set of rules or institutions exhibit desirable properties and thus command the rational support of those subject to them. Moreover, on the contractarian account, the reasons we have to support these rules or institutions are purely instrumental; they further the personal ends of each and every individual that abides by them. The social contract, in which each and every individual assents to a set of constraints, thus claims to reveal the solution to our fundamental predicament. It is individually rational, in a merely instrumental sense, to constrain one's behaviour in certain ways or to subject it to external constraint. Because the contractual agreement does not, in fact, occur, it is best interpreted as a model that aids in identifying the set of constraints that rational individuals should endorse. The model does so by demonstrating the acceptability of certain terms by parties who are relevantly similar to real people. Such modelled acceptance is meant to indicate that these terms and their enforcement can be 'justified' to – i.e. are rational for – those subject to them. In other words, because the parties to the modelled agreement have sufficient reason to accept a certain solution to the predicament, and because these agents, being rational and cautiously idealized, have roughly the same reasons as real people, we conjecture that real people also have reasons to endorse the solution agreed upon by the parties. A real society of utility-maximizing individuals, therefore, should be guided by the solution arrived at in the modelled agreement. Pulling these various threads together, we have *The Contractarian Schema*:

For a modelled agent, *A*, in a modelled deliberative environment *E*, considering a rule/principle/institution *R*, the fact that agreeing to *R* maximizes the (expected) utility of *A* suggests that a real agent, *A'*, in the real world, *E'*, maximizes *A'*'s (expected) utility by endorsing and complying with *R*.<sup>5</sup>

---

on the other hand, focuses primarily on coordination. Vanderschraaf (2006) argues that different players will have different preferences, meaning that for some the state of nature poses a prisoner's dilemma, while for others it poses, instead, an assurance game. For another interesting formal analysis along these lines – one that locates the source of war, not primarily in competition, diffidence and glory, but in uncertainty – see Chung (2015). For a concise presentation of various readings of Hobbes, see Gaus (2013).

<sup>5</sup>This is a refinement of the 'General Model of the Social Contract' laid out by d'Agostino *et al.* (1996). The use of expected utility theory is unnecessary and has the drawback of imposing certain assumptions on the preferences of the modelled agents. But its use by contemporary contractarians is so ubiquitous that a more careful formulation is an unnecessary hindrance.

There are several important implications of this schema. First, agents are modelled as *expected utility maximizers* (Gauthier 1986: 65–78; Buchanan and Tullock 2004: 23–26; Binmore 2005: 64; Moehler 2018: 101). This is not a logical entailment of anything fundamental in contractarianism. Indeed, Hobbes can hardly be said to have employed expected utility theory. It merely reflects the fact that expected utility maximization is the most well-developed and widely accepted model of rational choice that social science has to offer (Gauthier 1986: 8). Thus, in order to show that the terms of a contract are rationally acceptable – in a narrow, instrumental sense – contemporary contractarians generally seek to demonstrate that expected utility maximizers would endorse it.<sup>6</sup>

Second, and most obviously, all agents in the model must agree upon the terms of the social contract. *Unanimity* must attain. If a social contract were rationally rejected by one or more modelled agents, then its terms could not be justified as rational for real individuals.<sup>7</sup>

Finally, the schema imposes limitations on the level of idealization in which contractarians may indulge. If a real agent  $A'$  is to identify with modelled agent  $A$ , then  $A$  cannot be unidentifiable to  $A'$ .  $A'$  must see  $A$  as sharing a similar set of reasons. Two important modelling constraints follow from this limitation on idealization. First, the modelled agents must have similar evaluative standards to those of real agents (Moehler 2018: 30).<sup>8</sup> If  $A$  has values that are inimical to those of  $A'$ , then why should  $A'$  think the hypothetical choice of  $A$  has any bearing on his real choice? Secondly, the limitation on idealization imposes informational limitations on the modelled agents. Contractarians, in contrast to certain contractualists, present modelled agents as choosing under conditions of *social-scientific uncertainty*.<sup>9</sup> They choose in an expected utility framework, without knowing the actual ramifications of their choice.<sup>10</sup> Identification between real individual and modelled agents requires this restriction on idealization for at least two reasons:

1. To use the contractarian model for normative guidance, real individuals must be able to follow the reasoning and understand the choices of the modelled agents; in some formulations, they must be able to adopt the perspective of the modelled agents. This is not possible when the modelled agents possess relevant information that is unavailable to real individuals.

<sup>6</sup>A notable exception to this feature of contractarianism will be discussed later on (section 4).

<sup>7</sup>Unanimity is so basic to the idea of a contract that it is explicitly or implicitly adopted by all social contract theorists. Some notable examples include Locke (1980: 52), Hobbes (1991: 121), Kant (1996: 296), Brennan and Buchanan (2000: 31–33) and Gauthier (2013: 618).

<sup>8</sup>See also Gaus (2010: 232–258) for an important discussion of identification.

<sup>9</sup>For example, Gauthier (1986: 343–344), Buchanan and Tullock (2004: 37) and Moehler (2018: 110, 113, 115).

<sup>10</sup>To avoid confusion, it is important to bear in mind that social-scientific uncertainty can be completely subjective and is therefore consistent with a deterministic view of the functioning of society. As Poincaré taught us, a system can be difficult or impossible to predict even if it operates in a non-probabilistic manner (Poincaré 1920).

2. For the contractarian model to provide normative guidance, the choice set of modelled agents cannot radically differ from that of real individuals. Yet, real agents never have enough information to make precise predictions about the outcomes of adopting particular rules or institutions. And so, outcomes are not part of the choice sets of real individuals. At best, real individuals choose lotteries, and the contractarian model must reflect this. Treating a best guess as infallible prophecy is often highly irrational, especially when unexpected outcomes might impose heavy losses. Consequently, *A'* will not be able to identify with *A* if *A* chooses rules on the basis of a futuristic or superhuman ability to predict the results of such choice. *A'* simply cannot accept as rational a choice that ignores the possibility of unexpected and adverse outcomes.

Uncertainty profoundly affects the rules or policies that real individuals find desirable.<sup>11</sup> Social-scientific uncertainty, therefore, is crucial to the contractarian model insofar as it aims to demonstrate the rationality (to real individuals) of complying with its results.

While certain idealizations undermine the force of contractarian arguments, others enhance rather than diminish this force. As Michael Moehler explains, a model of human decision-making in the social contract need not be descriptive of the choice behaviour of real individuals. Rather, 'the ... model is normative. It aims to determine how rational agents should ideally choose and behave in order to best fulfil their interests in the world in which they live, even if the agents sometimes fail to do so in the real world' (Moehler 2018: 96).<sup>12</sup> Thus, it is no objection to contractarian models that real individuals are not perfectly rational, often failing to maximize even expected utility. In fact, contractarian models would lose rather than gain justificatory force by including common errors into the reasoning processes of modelled agents. Demonstrating that laws or morality emerge from rational agreement requires that the agents who agree exhibit rationality – and the purer the better. Following this line of reasoning, one might wonder whether contractarians have gone far enough in idealizing the rationality of their modelled agents.

This paper asks what happens when we do go further. What terms of agreement emerge if modelled agents are even more rational than typically construed? The analysis here accepts the key tenets of contractarianism detailed above: rationality as expected utility maximization, unanimity as the criteria of acceptance, and social-scientific uncertainty of modelled agents. In addition to

---

<sup>11</sup>The exact meaning of 'uncertainty' differs between theories. This paper employs the standard textbook conception of choice under uncertainty, viz. expected utility theory, wherein decision-makers have preferences over an exhaustive, disjoint set of final outcomes to which they assign choice-functional probabilities. They then choose among 'lotteries', or collections of outcomes weighted by the probability that they will occur contingent upon the selection of that lottery. See Kreps (2012: 79–116) or Mas-Colell *et al.* (1995: 167–207) for this type of approach. Others distinguish between 'risk' and 'uncertainty', where risk is the standard situation just outlined, while uncertainty involves ignorance as to the set of possible outcomes or an inability to ascribe probabilities to these outcomes. See Knight (1971).

<sup>12</sup>Compare Bacharach (1976: 2–3).

these three features, however, the analysis introduces a fourth feature: a criteria of rational belief formation.

To examine the implications of this enhanced conception of rationality, I first (section 2) discuss the neglect of rational belief in current contractarian theories and introduce the criteria of rational belief formation I employ, namely, Bayesian belief updating. In order to analyse the complicated implications of introducing rational belief formation into our notion of rational choice in the social contract, section 3 presents both an informal (section 3.1) and a formal (sections 3.2–3.4) explanation of a general contractarian model, which I call the *Agreement Model*. After noting several alarming results that emerge from this choice scenario, section 4 examines how to apply these results by considering two contemporary contractarian theories. Finally, in section 5, I conclude by recapitulating the results of this paper, discussing the implications of these results, and gesturing toward important questions that remain to be answered.

## 2. The epistemics of agreement

In contractarian models, and in decision theory more generally, rationality has typically meant that one's behaviour maximizes the satisfaction of rational preferences, where preferences are rational if they satisfy the consistency conditions of transitivity and completeness (Hausman 1992: 25; Mas-Colell *et al.* 1995: 6). However, an important advance in modelling rational behaviour involves paying special attention, not just to rational preferences, but also to rational beliefs. In many models, beliefs are no longer exogenously given, but become an endogenous feature of the model, depending upon the strategy set and payoffs of other players.<sup>13</sup>

Decision-theoretic models typically use a Bayesian framework to model rational belief formation. In such models, individuals begin with certain prior beliefs, often subjective or arbitrary. As they observe new information, they update their prior beliefs according to Bayes' rule. For example, suppose I know that a coin is either fair or it has heads on both sides. I initially believe that the coin is fairly weighted, but I then observe a flip where heads comes up. Given this new evidence, what should I now believe about the coin? If I started out being 90% sure that the coin was fair, then now I should believe that it is fair with only 82% probability.<sup>14</sup>

Despite the fact that one of the pioneers of Bayesian game theory, John Harsanyi, was himself a social contract theorist, the Bayesian framework is rarely applied to social contract theory.<sup>15</sup> And despite the increasing awareness by game theorists and political philosophers of the importance of rational beliefs, few, if any, major

<sup>13</sup>The application of Bayesian belief updating to game theory was pioneered by Harsanyi (1967), and the concept of sequential rationality, in which beliefs reflect the strategies and payoffs available to other players, was contributed by Kreps and Wilson (1982). For a general discussion highlighting the importance of rational beliefs for the analysis of games, see Hausman (2003).

<sup>14</sup> $P(F|H) = \frac{P(H|F)P(F)}{P(H|F)P(F) + P(H|B)P(B)} = (0.5)(0.9)/(0.55) = 0.818182$ , where F means 'coin is Fair', B means 'coin is biased', H means 'heads is observed', and  $P(x|y)$  is the probability of x given that y is observed.

<sup>15</sup>For some notable contributions to this project, see Vanderschraaf (2006) and Chung (2015). Muldoon *et al.* (2014) approach this project, but focus more on consensus building than on *bona fide* belief formation.

contractarian projects of the last several decades address the issue of rational belief formation.<sup>16</sup> Given the increasing reliance on formal models of rational choice in contractarian theories, this seems a glaring omission. The present article aims to improve this situation by applying a Bayesian model of agreement to the contractarian schema defined above. If the modelled agents are rational Bayesians, this article asks, what terms of agreement will result from the model? Will these terms command rational compliance among real people? There arises a surprising result: in reasonably large groups, the combination of unanimity, uncertainty and rational Bayesian choice yields predictably bad choices much of the time.

Or perhaps this is not so surprising. As I argue for shortly (section 3.1), and as we all know from personal experience, choosing in groups is hazardous terrain. We are constantly influenced by the choice behaviour of others, often in pernicious ways. Such influence may even call into question the often-assumed moral power of consent. This paper, focused as it is on contemporary contractarian theories, does not defend or deny any alleged moral power of consent. The contractarian schema concerns only the rationality of adhering to certain rules or standards and the demonstrability of this rationality via some model of agreement among utility-maximizing players. Contractarian models are, fundamentally, normative-epistemological tools for revealing the (instrumental) rationality of abiding by a set of terms or constraints. Yet, as the next section will show, even this basic rationality is undermined by the perverse effects of group pressures, in particular, by the way these pressures affect the beliefs, and consequent behaviour, of rational Bayesian agents.

### 3. The agreement model

#### 3.1. Explanation

Individuals choosing terms in the social contract have preferences over a set of moral or political rules, each representing a lottery over particular policy decisions or operational outcomes. The term ‘rule’ here is used very loosely, as a stand-in for whatever it is – institutions, moral principles, distributional shares, etc. – that the contractarian theorist posits as the object of choice. A ‘rule’ may even be a whole system of rules.<sup>17</sup> In order to determine which moral-political rules to support at the contractual stage, individuals must make predictions that map these rules to actual outcomes. Social science, especially its models of how individuals behave under varying rules or institutions, bridges this gap between abstract rules and concrete outcomes. Given that social science provides imperfect guidance, these models and their predictions will also be imperfect. Individuals will face *social-scientific uncertainty*. They will not choose outcomes,

<sup>16</sup>The projects I have in mind are those of Gauthier, Binmore, Narveson, Lomasky, Buchanan and Moehler.

<sup>17</sup>The advantage of thinking in terms of whole systems of rules is that such an approach will avoid issues of path dependence in the selection of rules. The drawback is that this approach will make massive cognitive demands upon modelled agents and real individuals attempting to follow the reasoning of modelled agents. For further discussion of this point, see section 3.5 and Gaus (2010: 267–276).

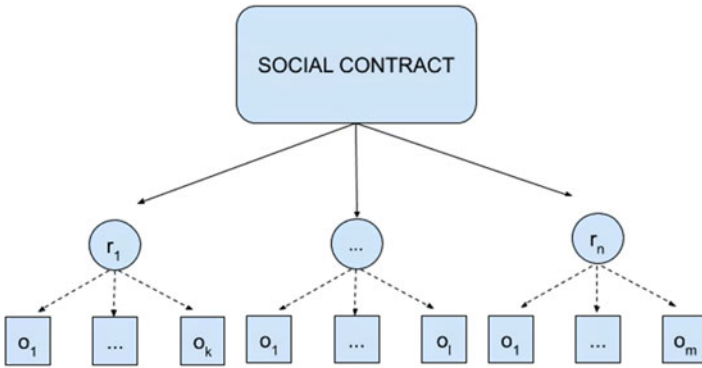


Figure 1. Subjective Uncertainty in the Social Contract.

but lotteries over outcomes.<sup>18</sup> This is not to assert that moral-political rules operate in a probabilistic or indeterminate way. The reason that agents view moral-political rules as lotteries is purely subjective: agents, even utilizing the best social-scientific theory available, are unsure how a given moral-political rule will operate or exactly which outcome(s) it will produce. This is true even if the outcomes arise in a purely deterministic manner once the rule has been chosen and implemented.

Figure 1 depicts the nature of choice in the social contract. Individuals evaluate some rule,  $r_i$ , by considering the value of the outcomes,  $o_1, \dots, o_j$  that it might produce and the probability that it will produce these outcomes.<sup>19</sup> Social-scientific uncertainty appears in the gap between each rule,  $r_i$ , and the various outcomes,  $o_1, \dots, o_j$  that it may give rise to – hence the dotted lines in Figure 1. Under the condition of perfect certainty,  $j = 1$  and the probability of  $o_j$  is equal to 1. Under the condition of social-scientific uncertainty, the value of  $r_1$ , for example, would be its expected utility:  $p_1u(o_1) + p_2u(o_2) + \dots + p_ku(o_k)$ , where  $p_i$  is the probability that outcome  $o_i$  comes about, and  $u(o_i)$  is the evaluation of that outcome.

One barrier to consensus that I wish to set aside is disagreement as to which final outcomes are most desirable. The Agreement Model aims to show that unanimous agreement leads to unsatisfactory results. Granting that parties agree as to which results they want is both charitable and simplifying – charitable in that it allows us to eliminate a serious barrier to achieving satisfactory results; simplifying in that it allows us to isolate a different, epistemic source of poor outcomes. Moreover, certain contractarians have explicitly sought to ‘normalize’ preferences in just this way.<sup>20</sup> So, if the analysis provided here did not grant this assumption, it might fail to apply to these variations of contractarianism.<sup>21</sup> One plausible way,

<sup>18</sup>I thank Jerry Gaus for challenging me to clarify this aspect of the set-up.

<sup>19</sup>Again, remembering that a ‘rule’ is simply any object of choice which generates the lottery over final outcomes.

<sup>20</sup>On normalization, see d’Agostino *et al.* (1996) and Rawls (2008: 226).

<sup>21</sup>See, for example, Buchanan and Tullock (2004), who posit that parties in the constitutional consensus choose as if behind a *veil of uncertainty*, and consequently evaluate final outcomes as if they were a



though certainly not the only way, of understanding this concurrence on the value of final outcomes, is to suppose that such outcomes receive a utility score corresponding to their impact on the well-being of the average individual. This would be the case if, for example, social-scientific uncertainty were so great that individuals could not predict with any accuracy the effect of a policy on any particular person, themselves included. In this case, all individuals would have identical preferences over the final outcomes because they evaluate these outcomes according to how they impact the average individual. Crucially, even in this extreme case, identical preferences over final outcomes do not imply identical preferences over rules or institutions. Although decision-makers may identically rank the final outcomes, or perhaps even lower-level political choices, they may still differ as to the probabilities they assign to each outcome. Disagreement may result from different predictions as to (i) the impact of a lower-level collective choice or (ii) the probability of various lower-level collective decisions following the selection of a moral-political rule at the contractual stage.

This predictive disparity can persist even if all decision-makers have the same exact information – so long as that information is not fully adequate to make accurate predictions. As James Buchanan and Roger Faith have argued, if the given information is insufficient to establish a uniquely reasonable probability distribution over specific outcomes, then subjective differences between individuals will lead to divergent predictions, even if all individuals possess the same (incomplete) information:

If . . . we must acknowledge that the individual’s generalized knowledge about alternative rules must be reflected in predictions rather than in objectively-measurable and observable data, we must also acknowledge that these predictions are inherently subjective . . . Once subjectivity is allowed, differences in predictions about the properties of alternative institutions or rules can be expected to emerge. (Buchanan and Faith 1980: 26)

Importantly, this implies that no amount of deliberation and discussion, insofar as this merely involves the exchange of information, will lead to a convergence of preferences over rules.<sup>22</sup> This result, theoretically grounded in the mathematics of Bayesian belief formation, is also corroborated empirically: our best social scientists exhibit vehement disagreement when it comes to future predictions.<sup>23</sup> If the information possessed, even once aggregated and disseminated, is insufficient to

---

“‘randomly distributed’ participant in the succession of collective choices anticipated’ (Buchanan and Tullock 2004: 91).

<sup>22</sup>Considerations aside from mere information can, however, produce consensus even when information is insufficient to do so. The Wagner–Lehrer model shows that if individuals assign certain ‘respect weights’ to the beliefs of their peers, an iterated process will necessarily lead to agreement (see Lehrer 1976; Wagner and Lehrer 1981). A *contractarian* procedure, in which modelled agents exhibit no moral motivations, a ‘respect weight’ seems out of place. The Wagner–Lehrer model has, however, illuminated aspects of the *contractualist* project (see Muldoon *et al.* 2014).

<sup>23</sup>A survey article published in *The Journal of Economic Forecasting*, for example, examined economic predictions made throughout the 1990s and concluded that ‘the record of failure to predict recessions is virtually unblemished’ (Loungani 2001: 1).

make definite predictions or to determine a uniquely reasonable probability assignment, then subjective differences will generate diverse preferences – not over final outcomes, but over moral-political rules, chosen at the contractual level. Thus, despite having identical preferences over final outcomes, individuals will assign different expected utility to the rules under consideration. In other words, the same moral-political rule will represent a different lottery to each individual. The lottery differs, not in virtue of the desirability of the final outcomes that might occur if it is chosen, but rather in virtue of assigning different probability to the occurrence of those outcomes.

In virtue of individuals' homogeneous preferences over *final outcomes*, the model is free to treat the goodness or badness of each moral-political rule – considered in terms of its outcome and not as an object of choice – as an objective fact. There are two important facts to bear in mind here. First, individuals assign different subjective probability to the final outcomes possible under different rules, but it remains the fact that different rules will lead to one final outcome (or, perhaps, generate an objective probability distribution over final outcomes) independent of the subjective beliefs of agents. Thus, the model treats moral-political rules as lotteries when they are objects of choice, but this does not imply anything about the internal structure or the functioning of such rules in an objective sense. Such rules may produce outcomes that are objectively good or bad, even though subjective beliefs lead to differing expectations and conflicting choice behaviour on the part of agents. Second, for contractarians, the goodness or badness of such rules will depend only on the preferences of decision-makers over the final outcomes, i.e. the actual impact on individuals that will result from the selection of that rule. The 'objective' goodness or badness of a contractually chosen rule depends on the homogeneous preferences of individuals over the final outcomes and not on some objective standard of correctness. This is crucial, since positing some independent standard of good and bad would violate the contractarian's commitment to deriving normative standards solely from instrumental rationality.<sup>24</sup> Also important is the fact that the Agreement Model does not specify exactly how utilities defined over outcomes (not expected utilities) map onto goodness or badness. There are multiple ways to define this relationship – e.g. a good rule is the one that maximizes outcome utility within the budget set – but the Agreement Model leaves the details of this relationship undefined so as to achieve maximal applicability to various contractarian models. These models may posit different relationships between individual preferences and social optimality. The Agreement Model need not choose one specification of this relationship, but should remain consistent with as many such specifications as possible.

With this framework in place, the details of the Agreement Model can now be considered. In determining the terms of their social contract, individuals evaluate one rule at a time, deciding whether to accept or reject it. They each have some prediction, a private signal, as to whether the rule is good or bad – i.e. whether

---

<sup>24</sup>In Buchanan's terminology, it would violate the approach of individualism and the rejection of the collectivist or 'organic' approach to evaluating institutions (Brennan and Buchanan 2000: 25–27; Buchanan and Tullock 2004: 11, 13–14). I thank Thomas Christiano for this point.

its final outcome produces a high or a low average payoff.<sup>25</sup> Since individuals choose under conditions of social-scientific uncertainty, their predictions as to whether a rule is good or bad will differ, even while their preferences over final outcomes remain identical – allowing for the ‘objective’ goodness or badness of the rule. Charitably, the model assumes that predictions are reasonably accurate. That is, every individual’s prediction of the objective goodness or badness of a constitutional rule – its impact on the utility of the average individual – is more likely to be correct than incorrect. After determining whether the rule is more likely to be good or bad, each individual makes a decision: accept or reject. Since unanimity is the condition for group acceptance at the contractual level (section 1), each individual has the ability to veto any rule by choosing to reject it.

Models of naïve choice within groups assume that an individual will observe his or her private signal (i.e. prediction) as to whether the rule is good or bad and then choose in accordance with that signal.<sup>26</sup> Call this *informative choice*. This behavioural assumption fails to capture the additional information that an individual gleans from the choice behaviour of others. It also fails to capture considerations of how an individual’s choice will actually affect that individual’s payoff. We thus require a more sophisticated version of rational choice within groups. Consider, as an alternative, what we may call *Bayesian choice*. Bayesian choice differs from informative choice in two related ways. First, the decision-maker updates her beliefs or predictions based on the assumption that she is pivotal. This aspect of Bayesian choice incorporates the core contribution of this paper, i.e. the inclusion of rational belief formation in the contractarian model. Second, the decision-maker restricts her focus to *pivotal choices*, that is, to situations where her choice actually impacts the group choice. This follows simply from expected utility maximization, since a non-pivotal choice for player  $i$ , by definition, makes no difference to the group choice or to player  $i$ ’s payoff. Under a unanimity rule, player  $i$ ’s choice is *not* pivotal whenever some player  $j \neq i$  has chosen to reject the rule. By contrast, player  $i$ ’s choice *is* pivotal when, and only when, all other players have chosen to accept the rule change. In that scenario an individual finds herself in a position characterized by power and information: power, since her choice determines whether the rule passes or whether it is rejected, and information, since she may now assume that all others have chosen to accept the rule as good. Outside of such a scenario, on the other hand, her choice makes no difference whatsoever to her payoff, since it will not affect the group’s choice.<sup>27</sup>

The key feature of the pivotal choice situation, the feature which affects the choice calculus of the individual by providing additional information for belief formation, is the fact that every single other individual has chosen to accept the rule. An individual who finds himself in the position of a pivotal chooser must

<sup>25</sup>Whether a payoff is high or low may be assessed by comparing it to the status quo or to some alternative rule in the choice set (or perhaps in some other way I have not imagined). The Agreement Model leaves this open in order to remain as general as possible.

<sup>26</sup>Voting models operating under this assumption include McCart (1964), Kalven Jr and Zeisel (1966), Levine (1992) and Adler (1994).

<sup>27</sup>The Agreement Model clearly precludes any intrinsic value that modelled agents may place in choice itself, such as the value of publicly expressing one’s convictions.

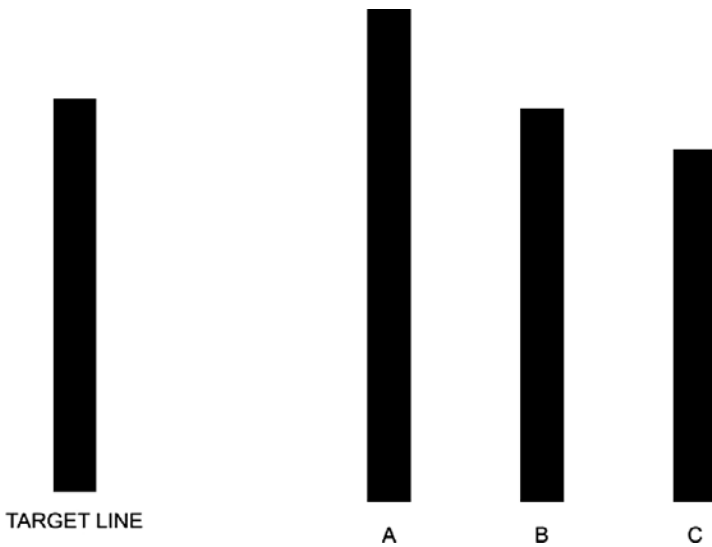


Figure 2. Asch's Lines.

therefore ask himself: 'What information can I glean from the fact that every single other individual has chosen to accept this moral-political rule?'

To better understand this situation, consider the famous Asch Conformity Experiment, in which Solomon Asch studied the tendency to conform one's beliefs to those of the group. In this experiment, eight individuals were lined up in order to conduct a 'vision test' in which each would announce which of three displayed lines – A, B or C – matched the length of a fourth line, the 'target line' (see Figure 2).

Unbeknownst to the eighth individual, the other seven were confederates of the experimenter, who told them to give an incorrect, though uniform answer. A shocking 37% of subjects ended up mirroring the obviously incorrect judgement of the seven confederates, even though fewer than 1% of subjects gave incorrect responses when choosing alone (Asch 1951).

Many interpret this experiment as a cautionary tale against the distorting influence of peer pressure and the tendency towards group conformity. This may very well be the correct interpretation of the results. For the purpose of understanding the model presented in this paper, however, consider another, less pessimistic interpretation: the judgements of other individuals provide additional evidence to the experimental subject. Under the assumption that the other individuals in the room are fellow test subjects, honestly attempting to report the evidence of their senses, their answers provide additional information to the eighth person. Because one's senses are imperfect and sometimes misleading, it seems wise to grant evidential weight to the perceptual judgements of others. In fact, as a general principle, when determining the correct answer to a question, it is usually rational to consider both one's personal belief and the expressed beliefs of others. When these judgements diverge, it becomes necessary

to compare the likelihood that one's belief is erroneous to the likelihood that the expressed beliefs of the others are erroneous. For a significant number of Asch's subjects, the first likelihood seemed greater.<sup>28</sup>

In precisely the same way, individuals in the Agreement Model must decide (when pivotal) whether their signal is more likely to be in error, or whether the choices of every single other member of the group are likely to be in error. Only if the latter is more likely, only if the choices of every single other member of the group are more likely to be erroneous, will the pivotal individual choose according to her private signal. That is, choosing *informatively*, according to one's private signal, will be the optimal strategy only when one's private signal is less likely to be erroneous than the choices of every other member of the group.<sup>29</sup>

A misunderstanding often arises here concerning the information the pivotal player receives from his or her pivotal position. It is not as if some special player gets to choose last, while all the other players must choose beforehand, revealing their choices; nor is it that the other players credibly divulge to some unique pivotal player how they plan to choose. Instead, each player, in determining his or her strategy, considers only the scenario where he or she occupies a pivotal position. No other possible scenario, regardless of how likely it may be, is relevant for deciding how to choose, since no possible scenario except for that in which the player is pivotal affords the player an opportunity to affect the outcome. Choice in the agreement model may therefore be instantaneous, and prior communication as to how each agent plans to vote is unnecessary (though permitted). Each agent simply restricts her attention to the pivotal position and, accordingly, updates her beliefs conditional upon occupying that position.

The complexity of group choice, and the consequent need to rely on a mathematical model rather than intuition, becomes clear when we note that all individuals in the group will reason similarly. In consequence, no individual, even when pivotal, can take for granted that the choices of other individuals reveal their private signals – for other individuals are *also* Bayesian choosers, considering only the situation in which they are pivotal and basing their choice on the information implicit in that situation. Nevertheless, their choice does provide *some* information about their private signal. The extent to which it does so and how this information should affect the updated beliefs and choice

---

<sup>28</sup>The claim that rational beliefs involve Bayesian updating conditional on the choices of one's peers relates this argument to a certain epistemological position with respect to peer disagreement. In particular, it is one way of cashing out the 'equal weight' view of peer disagreement, in which judgements made by one's 'peers', in a technical sense, are given equal epistemic weight as one's own judgements. The equal weight view is not, as Frances and Matheson put it, 'the only game in town' (Matheson and Frances 2018). However, it prevails over alternative theories in terms of the attention and support it receives from epistemologists. Its many defenders include Christensen (2007), Elga (2007), Bogardus (2009) and Matheson (2015).

<sup>29</sup>In such a situation, informative choice is technically a special case of Bayesian choice as defined above, since one still considers only the pivotal case and one does take the judgements of others into account in order to update one's beliefs. What makes this special case so special is merely that, even given one's updated beliefs, the best response is still to choose that action that would be prescribed by the naive strategy that only considers one's private information. In the special case, the result of Bayesian choice and informative choice happen to coincide.

behaviour of a given decision-maker must be investigated in the formal version of this model (sections 3.2–3.4). Intuition has led us to see that individuals may not vote informatively when they seek to maximize expected utility and when they take the choices of others as providing new information. But beyond this insight, intuition leaves us grasping, and further progress requires formal deductions.

In the following subsections, these deductions will demonstrate that informative choice does not (generally) emerge as a Nash equilibrium in the model.<sup>30</sup> In fact, there is only one responsive, symmetric Nash equilibrium. This equilibrium is derived by assuming that individuals apply Bayesian belief updating to their prior beliefs (their initial signals), conditional on being pivotal. This equilibrium, unfortunately, has a very undesirable property: a high prevalence of accepting bad rules. In statistical terms, type-1 error, consisting of ‘false positives’, is rampant. The group frequently chooses rules that no member of the group wants.

If this model accurately represents contractarian choice, then the hypothetical contracts of contractarian theorists would not, in fact, yield justified decisions. If we find, as our model shows, that the unanimity rule under conditions of uncertainty does *not* benefit each member of the social group – and in fact benefits *no* member of the social group – then the justificatory power of the contractarian approach is highly suspect. Aside from the addition of Bayesian belief formation, the model relies only on assumptions, discussed in section 1, that are standard in contractarian models: expected utility maximization, uncertainty and unanimity. Thus, its results appear readily applicable to many, if not all, variants of contractarianism, though this question must be examined further (section 4).

In what follows, the exact specifications of the model are laid out (section 3.2), some of its important features are highlighted (section 3.3), and four propositions are presented with a description of their importance (section 3.4). Proofs of the four propositions, being tedious and unenlightening, are quarantined to the Appendix.

### 3.2. Set-up

Moving beyond the limits of philosophical intuition requires a formal model. Timothy Feddersen and Wolfgang Pesendorfer have developed a model, originally intended to examine jury decisions, that perfectly captures the epistemic interdependence and strategic behaviour of agents under the unanimity rule.<sup>31</sup> By reinterpreting this model as a contractarian procedure among rational Bayesian agents, the results of *Bayesian choice*, i.e. of epistemic interdependence and strategic behaviour, can be examined. Rather than jurors, we have parties to the contract; rather than desiring to convict the guilty and acquit the innocent, agents wish to accept good rules and reject bad ones; rather than private signals about the guilt or innocence of a defendant, agents have private (prior) appraisals of the goodness or badness of the rule under consideration. More precisely . . .

<sup>30</sup>The exception is for small  $n$ , as will be explained below (section 3.4).

<sup>31</sup>The model I refer to is presented in Feddersen and Pesendorfer (1998). For other similar models, see Austen-Smith and Banks (1996), Feddersen and Pesendorfer (1996, 1999) and Wit (1998).

- There are  $n$  ‘choosers’:  $N = \{1, \dots, n\}$
- There are two possible states of the world, one where the rule being considered is Good, ‘G’ and one where the rule being considered is Bad, ‘B’:  $\Omega = \{B, G\}$
- There are two types of decision-maker, defined by the ‘signal’ they receive about the state of the world (i.e. their best guess about whether the rule is good or bad):  $t_i \in \{g, b\}$ 
  - $g$  stands for ‘good’, and  $b$  stands for ‘bad’
- The signals are more likely to be correct than incorrect, that is:<sup>32</sup>  
 $p = Pr(G|g) = Pr(B|b) \in (0.5, 1)$ 
  - $Pr(B|g) = Pr(G|b) = 1 - Pr(G|g) = 1 - Pr(B|b) < 0.5$
- Decision-makers may either choose to Accept, A, or Reject, R, the rule, or they may randomize:
  - Pure strategy set:  $S_i = \{A, R\}$
  - Mixed strategy set:  $Pr(A, t_i) \in \Sigma_i = [0, 1]$
- The preferences of decision-makers can be represented by utility functions with the following payoffs:
  - $U_i(A, G) = U_i(R, B) = 0$
  - $U_i(A, B) = -q; U_i(R, G) = -(1 - q)$
  - Where bold **A**, **R** signify that the group has come to that decision (rather than just the individual  $i$ ) and where  $q \in (0, 1)$
- The unanimity rule stipulates that bold **A** is collectively chosen if and only if  $s_i = A, \forall i \in N$ , and bold **R** is chosen otherwise (i.e.  $\exists j \in N : s_j \neq A$ )

### 3.3. Some important features

A first important feature of the Agreement Model is the individual’s updated belief, call it ‘ $\beta$ ’, that a rule is good upon observing  $k$   $g$ -signals. That is, if we imagine that each decision-maker  $i$ ’s private information were public and that each  $i$  could view the signals of every other  $j \in N$  what would this generic  $i$  believe about the probability of a rule being good or bad?  $\beta$  is a function of  $k$ , the number of good signals, and  $n$ , the number of decision-makers. More specifically, applying Bayes’s rule, we get:

$$\beta(k, n) = \frac{[p^k(1 - p)^{n-k}]}{[p^k(1 - p)^{n-k}] + [p^{(n-k)}(1 - p)^k]} \tag{T}$$

with  $p$  as defined above (Feddersen and Pesendorfer 1998: 24).

Secondly, Federsen and Pesendorfer note an important threshold (without deriving it):

**Lemma.** *If  $q < \beta(k, n)$ , then a pivotal chooser will prefer to Accept, rather than to Reject the rule upon viewing  $k$   $g$ -signals out of  $n$  total signals.*

*Proof.* See Appendix, section 0. □

---

<sup>32</sup>This is a charitable assumption. A standard criticism of the Condorcet jury theorem is that this assumption is often false. By granting this assumption and demonstrating that unanimity still results in poor outcomes, I hope to move beyond the commonplace objection that decision-makers are unreliable.

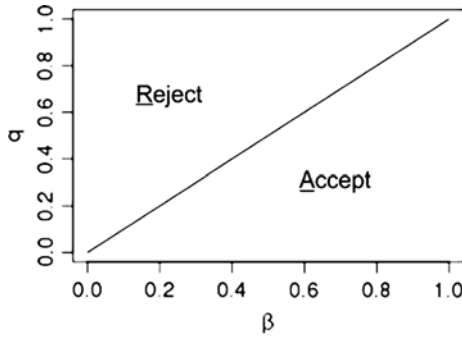


Figure 3. Threshold of Acceptance.

As defined above (section 3.2),  $q$  is a measure of how much decision-makers resent accepting a bad rule. In other words,  $q$  is the cost of type-1 error.  $\beta$ , on the other hand, measures a decision-maker's belief that a rule is good. A higher  $\beta$ , resulting from the 'observation' of more g-signals, means a greater confidence in the goodness of a rule. What our Lemma states is that, for any degree of confidence  $\beta$  that a rule is good, there exists some  $q$  that makes the decision-maker indifferent between accepting and rejecting the rule. Moreover, if  $q$  is higher than this indifference threshold, the decision-maker will reject the rule, while if  $q$  is lower than this threshold, the decision-maker will accept it. In other words, for any level of confidence (less than 100%) in the goodness of a rule, type-1 error can be so odious that the decision-maker will choose reject. Or, viewing things from the other axis,  $q$  determines a threshold such that, if  $\beta$  is greater than  $q$ , choosing to Accept is a best-response, while if  $\beta$  is less than  $q$ , choosing to Reject is a best-response.

In Figure 3, what I call the 'Threshold of Acceptance' is given by the line separating a region of acceptance from a region of rejection. This line gives the value of  $q$  required to make a decision-maker exactly indifferent between Accepting and Rejecting a rule, given that decision-maker's posterior belief,  $\beta$ .

A final important observation is that it is mathematically possible – for certain parameters  $n$ ,  $q$ ,  $p$  – for a decision-maker to observe  $(n - 1)$  g-signals (i.e. every other decision-maker thinks it is a good rule), but to still prefer Reject to Accept. In mathematical terms, it is possible that  $q > \beta(n - 1, n)$ . This will be an important fact for Propositions 1 and 4, laid out in the next subsection.

### 3.4. Four propositions

**Proposition 1.** *Choosing informatively is not a Nash Equilibrium for large  $n$  (holding other parameters fixed).*

**Proof.** See Appendix, section 1. □

The significance of this proposition is that it clears away many of the assumptions that weigh in favour of the unanimity rule. For example, the assumption that a unanimity rule will minimize type-1 errors: the error of



accepting a bad rule (a false positive).<sup>33</sup> The unanimity rule does indeed minimize type-1 error in the case of informative choice, where the probability that an accepted rule would be:  $Pr(G|A) = p^n/[p^n + (1 - p)^n]$ . In the case of informative choice, we can apply a logic similar to that operating in the Condorcet jury theorem: if  $(n - 1)$  others choose to Accept, then that means that  $(n - 1)$  others received signal  $g$ . Since the probability of receiving an accurate signal is greater than the probability of receiving an inaccurate signal, this means that the probability that the rule is genuinely good quickly goes to 1 as  $n$  increases. In this case, the best-response of player  $i$  is to choose Accept.

Yet, if there is not a direct link between the choices of other players and the signals that they have received, then the connection between player  $i$ 's best-response and the choices of other players is not so straightforward. In particular, if those other decision-makers are not choosing informatively, but as Bayesians, then the information that player  $i$  may glean from their choice behaviour depends on how much influence their signals actually have on their actions. It remains to be seen how players will respond to one another in the case of non-informative, i.e. Bayesian, choice. Feddersen and Pesendorfer identify a unique (slightly refined) Nash equilibrium, the formulation of which is stated in their original paper.

**Proposition 2.** *There is a unique, responsive symmetric Nash equilibrium, given by:*

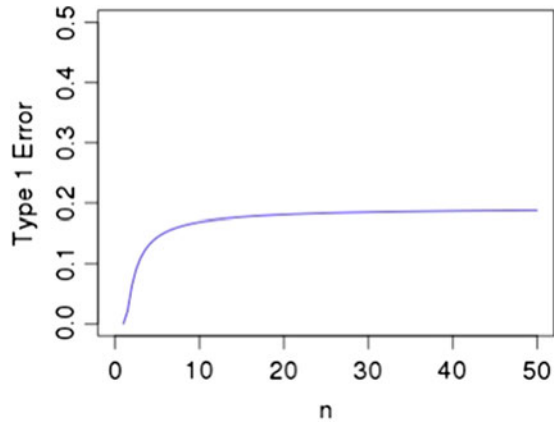
$$\sigma_i^*(p, q, n) = \frac{((1 - q)(1 - p)qp)^{1/(n-1)}p - (1 - p)}{p - ((1 - q)(1 - p)qp)^{1/(n-1)}(1 - p)}. \tag{*}$$

**Proof.** See Appendix, section 2. □

A responsive equilibrium is one where each decision-maker's strategy is responsive in the sense that her action changes as a function of the signal she receives. More technically, the probability of choosing Accept, given that one's signal is  $g$ , is not equal to the probability of choosing Accept, given that the signal is  $b$ . Restricting our attention to responsive equilibria allows us to rule out equilibria that are unrealistic, definitional oddities. That is, equilibria that satisfy the technical definition of Nash equilibrium without presenting a convincing or interesting representation of human behaviour. For example, it is technically a Nash equilibrium for all individuals to Reject no matter what signal they receive. In such a situation, an individual will never be pivotal, so all actions yield the same payoff. By definition, then, choosing Reject is a best-response (so is choosing Accept).

A symmetric equilibrium is one in which all decision-makers play the same strategy. That is, if two individuals have the same information, then they will choose Accept with the same probability as one another. More technically,  $(\sigma_i(b), \sigma_i(g)) = (\sigma_j(b), \sigma_j(g)), \forall i, j \in N$ . The intuition behind symmetry is that identical decision-makers facing identical incentives should choose identical strategies. With identical preferences (assumed throughout), identical priors and

<sup>33</sup>See, for example, Buchanan and Tullock (2004: 6–7) and Moehler (2018: 11).



**Figure 4.** Type-1 Error as a Function of  $n$  (for parameter  $p = 0.75$ ).

identical signals (and hence posterior beliefs), the basis on which we could justify ascribing different actions to such players is unclear. Nevertheless, there are cases where this assumption could break down, and asymmetric equilibria are a real possibility.<sup>34</sup> For the present purpose, however, the existence of a Nash equilibrium that predictably results in high rates of error suffices to establish the main claim: rational belief formation raises issues for contractarian justification. The fact that this Nash equilibrium *uniquely* satisfies certain intuitive conditions – viz. responsiveness and symmetry – serves to compound this worry.

With responsiveness and symmetry in place, the assumption that  $q < \beta(n-1, n)$  implies that we are seeking a mixed strategy profile in which  $\sigma_i(b) \in (0, 1)$  and  $\sigma_i(g) = 1$ . As shown in the formal Appendix, this is the only Nash equilibrium strategy that emerges under these constraints. For my argument, the importance of this Nash equilibrium lies in its various properties, identified by Feddersen and Pesendorfer. These are laid out in Proposition 3.

**Proposition 3.** *The unique, responsive symmetric Nash equilibrium (proposition 2) exhibits the following properties:*

1. Type-I error is bounded below.
2. Type-I error is increasing in  $n$ .
3. Type-I error attains high levels ‘quickly’ as  $n$  grows.

**Proof.** For proof of (1), see Feddersen and Pesendorfer (1998: 32). For proof of (2) and (3), see Appendix, section 3.  $\square$

For our purposes, this is the central proposition of the model. Surprisingly, unanimous agreement leads to the selection of bad rules, rules that no individual actually wants. Figure 4 illustrates this tendency towards high degrees of type-1 error with some specific parameters. If the models employed by contractarians aim to identify justified rules and if these rules are justified in virtue of their

<sup>34</sup>I thank an anonymous reviewer for flagging this issue.

ability to ‘benefit each member of the social group’ (Buchanan and Tullock 2004: 15),<sup>35</sup> then this result poses a problem for contractarian justification. It would seem that unanimous agreement (among uncertain Bayesian decision-makers) *does not* reliably produce beneficial rule changes. This undermines the contractarian’s claim to have provided a tool for evaluation and legitimation.

**Proposition 4.** *Informative choice is attainable as a Nash equilibrium for small  $n$  and  $q \geq 1/2$ .*

**Proof.** See Appendix, section 4. □

Proposition 4 is significant due to the solution it suggests: smaller groups of decision-makers may avoid altogether the appeal of Bayesian choice. Intuitively, this results from the fact that there are not enough other decision-makers to override each individual’s confidence in his or her private information. Even if I suppose that  $(n - 1)$  others Accept, my best response if I think the rule is bad, is to Reject. Due to this fact – i.e. that decision-makers with a  $b$ -signal will Reject even when pivotal – informative choice is prescribed by the Nash equilibrium strategy, and type-1 error, though still present, will not rapidly approach high levels, as occurs when Bayesian choice diverges from informative choice.

### 3.5. Simplifications and limitations of the model

Like all models, the Agreement Model makes a host of simplifying assumptions. It is worthwhile to make these explicit for at least two reasons. First, simplifications can range from clarifying, to innocuous, to misleading. To determine whether a given simplification falls under one of these descriptions requires explicitly identifying it and considering which aspect of the world the scientist aims to model and how the model is meant to illuminate that aspect.<sup>36</sup> To assess the importance of a model’s results, we must determine whether its assumptions, given the model’s purposes, clarify or distort the phenomena it represents. This requires identifying and scrutinizing these assumptions. Second, explicit identification of assumptions facilitates scientific progress by signalling where new models might advance upon old ones. Models rarely, if ever, have the last word on their subjects. New models emerge by modifying, relaxing or abandoning the assumptions of older ones.

Several assumptions of Agreement Model – e.g. expected utility maximization and Bayesian belief updating – have already been discussed (section 1). However, some of these assumptions bear further comment, while others have thus far remained implicit. Before continuing, this section identifies and discusses some of the most conspicuous assumptions of the Agreement Model.

A first assumption worth discussing is that of homogeneous preferences (over final outcomes) among agents in the Agreement Model. This assumption is

<sup>35</sup>See also Buchanan (1975: 51). Buchanan is not the only contractarian who takes unanimous consent to imply universal benefit. Other contractarians frequently endorse this inference, e.g. Gauthier (2013: 618–619).

<sup>36</sup>Such considerations are part of the reason that Daniel Hausman characterizes economics as an ‘inexact and separate’ science (Hausman 1994).

substantive, and given the recent attention to diversity in social contract theory it may appear controversial.<sup>37</sup> Nevertheless, there are several convincing reasons for accepting the assumption of preference homogeneity in this instance.

The primary justification for this assumption is that preference homogeneity has typically been criticized for trivializing the task of the social contract theorist.<sup>38</sup> In other words, homogeneity makes the contractarian's job *too easy*. Since this paper aims to present a problem for contractarianism, it is actually desirable to present a version of contractarianism that is as easy to defend as possible. Homogeneity can thus be viewed as a charitable concession to the contractarian position.

Furthermore, since there are multiple ways in which a contractarian argument might founder, adopting the assumption of preference homogeneity allows us to isolate the particular issue identified in this paper. Disagreement about values in the contractual scenario generates its own issues (Muldoon *et al.* 2014). The present paper, however, wishes to explore a distinct set of problems, arising, not from disagreement about values, but from rational belief formation and strategic choice behaviour.

A third reason for assuming preference homogeneity is the fact that, in many contractarian projects at least, the modelled agents do not know how the chosen rules will affect them as particular individuals. Thus, as Buchanan and Tullock explain, 'The self-interest of the individual participant . . . leads him to take a position as a 'representative' or 'randomly distributed' participant in the succession of collective choices anticipated. Therefore, he may tend to act, from self-interest, as if he were choosing the best set of rules for the social group' (Buchanan and Tullock 2004: 91). Or as Buchanan put it elsewhere, 'A person who remains uncertain as to how a particular rule will impact on her own circumstances will be led, through rational choice precepts, to prefer that rule (or principles or set of rules) that will best further the interest of the anonymous member of the group' (Buchanan 2002: 489). In light of this, it is plausible to model agents as if they all aim to select rules that will maximize the same set of preferences over final outcomes, i.e. rules that further the interests of the average individual. Disagreements will arise, not over the value of final outcomes, but over which rules will best achieve the desired outcome.

A final point on preference homogeneity: in no way do I wish to argue that preference homogeneity is the only, or even the best, assumption to make for all models or projects. For the reasons enumerated above, however, I do claim that it is defensible and worthwhile to examine the contractarian model under this assumption. It might be valuable to explore the implications of a modified Agreement Model in which agents differ in their evaluations of final outcomes. For example, a model that incorporates 'noise' into the signals that agents receive would be an illuminating way of exploring preference diversity. If we consider the main mechanism by which the Agreement Model obtains its result – viz. the informational content of other agents' choices and the strategic responses to this information – then there is good reason to believe that the result will hold under

<sup>37</sup>Recent theorists who explore the implications of diversity for social contract theory include Gaus (2010: 276, 2016: 115–132), Moehler (2010, 2018) and Muldoon (2016).

<sup>38</sup>See, for example, Kymlicka (1991: 193) or Gaus (2010: 276, 2016: 115–132).

some amount of preference heterogeneity. As long as preferences are not *totally* independent, the choice behaviour of other agents will provide information, prompt Bayesian updating, and elicit strategic responses. But this issue should be examined more rigorously. By incorporating rational belief formation into the contractarian model, this paper is novel and exploratory. It thus serves as an invitation for further work, not as a definitive statement on the topic.

A second assumption of the model worth discussing concerns the procedure by which rules are selected. In the Agreement Model, parties consider rules one at a time, making a binary choice between accept and reject. The main reason for adopting this assumption is that most social contract theorists, when explicit on the matter, endorse this choice procedure. For Hobbes there is a single, stark choice between accepting or rejecting an absolute sovereign (Hobbes 1991: 120). Many contractarians, moreover, have employed bargaining theory to derive conclusions about the choice of rules in the social contract. The most notable projects of this type are Gauthier's early theory (henceforth 'Gauthier 1.0') as presented in *Morals by Agreement* and Kenneth Binmore's intricate, naturalistic contractarianism as presented in multiple volumes and papers (Gauthier 1986; Binmore 1994b, 2004, 2005). Although bargaining theorists in both game theory and in philosophy often apply an axiomatic approach, it is not difficult to conceive of bargaining as a game in which players consider proposals and face a binary choice between 'accept' and 'reject.' Indeed, Ariel Rubinstein famously employed this exact set-up to prove that a dynamic bargaining game will yield the same result as Nash's axiomatic approach, which Binmore explicitly endorses.<sup>39</sup> Gauthier 1.0 describes his bargaining process quite similarly, as involving proposals and a binary choice among bargainers between accept and reject (Gauthier 1986: 133). The fact that this binary structure, a choice between accept and reject, is so ubiquitous in contractarianism supports its use in the Agreement Model.

Despite its wide use, this choice procedure exhibits two apparent defects. First, this procedure will often fail to select a unique rule from a list of options. Suppose, for example, that there are three property rules under consideration –  $r_1, r_2, r_3$  – and agents go through each rule, determining by the unanimity criterion which rules are good and which are bad. Will they select at most one rule from the set? Well, no: they might decide to accept both  $r_1$  and  $r_3$  – even if they all strictly prefer  $r_3$  to  $r_1$ . More generally, the decision rule employed here may simply partition the set of rules into two categories, the accepted and the rejected, with no guarantee that a given rule in the accepted category will be optimal. The issue, of course, is that contractarian procedures select a unique rule (or set of rules) to govern each domain; they do not merely partition the set of possible rules into two rough categories. Although this issue would not arise in a Hobbes-like case, where only one rule needs to be chosen, most contractarians posit unanimous agreement as the basis for selecting a large, unique set of moral-political rules. If the Agreement Model differs in this significant way from other contractarian models, then it fails

<sup>39</sup>Rubinstein describes the game as follows: "Two players have to reach an agreement on the partition of a pie of size 1. Each has to make in turn, a proposal as to how it should be divided. After one player has made an offer, the other must decide either to accept it, or to reject it and continue the bargaining" (Rubinstein 1982: 97). See Binmore (1994a: 81, 1994b).

to accurately represent them. If it fails to accurately represent them, then it also fails to offer a cogent critique of these models.

As already pointed out, however, the Agreement Model is quite malleable; its results hold under a wide variety of specifications. This malleability allows it to apply to a diverse set of contractarian models. In this case, we might look at the bargaining process to see if the Agreement Model can be specified so as to avoid the issue of non-uniqueness. Bargaining models typically apply the bargain to one specific case at a time, e.g. how to distribute the gains from a cooperative interaction, and stop the choice procedure once unanimity is attained. So, returning to our simple example, if agents consider rules  $r_1, r_2, r_3$  (each of which presents a mutually exclusive rule for governing a single issue/realm of interaction), and if they reach unanimous agreement on  $r_1$  before they even consider  $r_2$  or  $r_3$ , then  $r_1$  is the chosen rule. It is unique, even though, as noted above, parties would also have agreed unanimously to  $r_3$  if they had considered it. Hence, in the non-Hobbesian case where parties must select from a large set of rules, we can construe the Agreement Model as follows: for each realm or issue, agents will consider rules one at a time, voting to accept or reject them. The process for one realm or issue stops and agents move on to the next realm or issue as soon as unanimity obtains. In this way, agents choose a unique rule for each realm or issue, and the Agreement Model avoids the issue of non-uniqueness.

This procedure, however, brings to light the second defect of the binary choice procedure: *path-dependence*.<sup>40</sup> Since the procedure stops at the first unanimously acceptable rule, there is no reason to believe that the *best* rule has been selected. This is because the order in which the rules are considered will crucially affect which ones are chosen. In the above example, if agents consider rules in the order  $r_1, r_2, r_3$ , then they will choose  $r_1$ , while if they consider them in the order  $r_2, r_3, r_1$ , they will choose  $r_3$ . Not just the quality of the rules, but also the order of consideration thus determines the final choice.

Such path dependence clearly poses a problem for a choice procedure meant to justify its output, but this does not necessarily mean that the Agreement Model fails to represent contractarian choice procedures. First of all, path dependence is a problem with social contract theories, generally. John Thrasher has recently argued that path dependence is endemic in multi-stage social contract choice procedures (Thrasher 2019: 440). Gerald Gaus has considered the path dependence that inevitably arises due to the functional interdependence of different rules, concluding that the costs of avoiding it are prohibitive and that the social contract theorist is better off embracing it, at least in a mitigated form.<sup>41</sup> Path dependence, therefore, seems to be a normal feature of social contract theory. Its appearance in the Agreement Model, which attempts to model such theories, should not surprise us.

<sup>40</sup>See Gaus (2008: 166–170) for a discussion of how path dependence plagues collective choice.

<sup>41</sup>Social contract theorists employ a wide variety of methods to avoid pernicious path-dependence. Hobbes makes the choice singular (Hobbes 1991: 120); Rawls endorses an extreme normalization of agents' preferences (Rawls 2009: 120); Gaus rejects the avoidance of path dependence (in most cases) as unduly demanding (Gaus 2010: 273–274). These approaches have their strengths and weaknesses.

One might object, however, that the path dependence exhibited by the Agreement Model is of a different sort than the inevitable types identified by Thrasher and Gaus. Here, path dependence occurs not because of doxastic diversity, as in Thrasher, nor the interdependence of moral-political rules, as in Gaus, but due to the ability of unanimous agreement to occur when superior options are available. But this points us towards a deeper reply to the issue of path dependence, namely that such path dependence closely relates to the core critique that the Agreement Model presents. Returning to the simple property rule example, suppose  $n$  agents with fairly accurate beliefs (as the Agreement Model assumes) are choosing from  $\{r_1, r_2, r_3\}$ . Suppose, in addition, that  $r_3$  is the best rule from the set. We should expect more than half of the population to prefer  $r_3$  to  $r_1$ , but only *before the epistemic effects of group choice are present*. In other words, most individuals would reject  $r_1$  if considering it in isolation. They consider  $r_3$  to be a preferable alternative, hence  $r_1$  fails to make it into the choice set. Once they start to think strategically, however, and consider the epistemic importance of being a pivotal chooser, they will vote to accept  $r_1$ . Thus, it is only the role of unanimity and the Bayesian choice that players face in a group context that allows path dependence to emerge so forcefully. Path dependence, seen in this way, is not a defect of the Agreement Model, but an aspect of its critique of unanimous choice within contractarian theory. The claim, therefore, is that if existing contractarian models were to incorporate rational belief formation into their choice procedure, they, too, would exhibit this sort of path dependence. The challenge for the contractarian is to contrive some way of avoiding these consequences, perhaps by rejecting Bayesian belief formation, the unanimity rule, social-scientific uncertainty or expected utility maximization.

A final assumption worth discussing is that, following Feddersen and Pesendorfer, the Agreement Model employs (Bayesian) Nash Equilibrium as its solution concept. Although this is a standard approach, it faces challenges from a significant minority of game theorists. Michael Bacharach has argued that, although all solutions must be Nash equilibria, even a unique Nash Equilibrium need not be a *bona fide* solution to the game (Bacharach 1987: 44).<sup>42</sup> Bacharach demonstrates this rather pessimistic conclusion (pessimistic for the prospects of developing an explanatory game theory) using an example involving limited knowledge about the actions of other players. The key idea is that we can define beliefs that render action profiles rationalizable, even if such profiles are not comprised solely of best responses (Bacharach 1987: 45). In other words, a strategy that is not a best response (given the strategies of others) need not be a strictly dominated strategy.

It would go far beyond the present topic to contribute to this important methodological discussion. However, there is reason to believe that the Agreement Model does not exhibit the features that may undermine the solution status of a unique Nash equilibrium. Agents in the Agreement Model do not choose based on their best guess as to what other players will choose, as in Bacharach's example. Rather, agents choose based on the hypothetical scenario in which they are pivotal, since this is the only situation in which their strategy will affect their

<sup>42</sup>I thank an anonymous reviewer for drawing my attention to Bacharach's important paper.

payoffs one way or the other. Hence, unlike Bacharach's example, agents in the Agreement Model can assume that they know the actions of all other players (i.e. that they all choose Agree) with 100% probability. Agents then select their strategy on this basis. This undermines the logic of Bacharach's argument that unique Nash equilibrium and a game's solution can come apart, an argument that relies upon agents not knowing which 'world' they are in, and consequently, not knowing what action the other player(s) will take (i.e. not knowing whether the second player will, in fact, play a best response to the first player's Nash equilibrium strategy). There may be other conditions that render Nash equilibria untenable as solutions, but the Agreement Model escapes, at least, those conditions identified by Bacharach.<sup>43</sup>

In addition, as Bacharach notes elsewhere, game theory can fail as a predictive or explanatory endeavour without failing as a prescriptive one (Bacharach 1976: 2–3). In the context of contractarianism, this point is especially important. Real individuals could hardly be moved by the result of a hypothetical contract if this result were not a Nash equilibrium, since in that case one or more of them would be expected to play a strategy that is not a best response. Employing Nash equilibrium thus seems especially justifiable in the normative, contractarian context – indeed more justifiable than in its use by Feddersen and Pseudofer to explain and predict jury decisions.

### 3.6. Final comment

The conclusions of the Agreement Model are significant. If the contractarian model is essentially a normative-epistemic tool for revealing rational constraint, then it is essential that the agents in the model choose as rationally as possible and that the constraints they choose command rational adherence. Yet, the Agreement Model shows that once we idealize agents to the point of forming rational beliefs, their choices – however individually rational – yield undesirable outcomes. In taking account of the process of rational belief formation, the agreement model has demonstrated that unjustified results can emerge from unanimous agreement among hyper-rational agents. The challenge for contractarianism is thus set: if the unanimous agreement among Bayesian decision-makers under conditions of uncertainty often yields universally undesired results, how can contractarian choice purport to identify justified rules? Or, alternatively, which assumption will the contractarian model reject and why: Unanimity? Uncertainty? Expected utility maximization? Rational, Bayesian belief formation? None of these assumptions are arbitrary. None can be rejected without some explanation.

The purpose of this paper is to explore the implications of accepting a Bayesian framework in the contractarian model. So one of its important conclusions may be that contractarians must reject Bayesian updating in their models of the social contract. Recognizing that one must reject an otherwise plausible assumption in order to make the model work does not, however, justify the rejection of that

<sup>43</sup>An even stronger defence of the solution identified in proposition 2, a defence not undertaken here, would abandon Nash equilibrium, instead employing the weaker (i.e. less presumptive) solution concept of rationalizability. I thank an anonymous reviewer for suggesting the value of such an inquiry.



assumption. Instead, it demands that one either provide a good reason to reject that assumption or construct an alternative model. Before this burden may be attributed to any particular contractarian project, however, one must assess how well the Agreement Model represents the model employed in that particular project. The following section provides two examples of this kind of assessment.

#### 4. Applying the model

If a choice scenario satisfies the assumptions of the model, then the perverse outcome described above must follow. But how well do these assumptions describe the models employed by contractarian theorists? In this section, I consider two examples of contemporary contractarian theories – those of David Gauthier and of James Buchanan and Gordon Tullock – in order to see whether or not their choice models satisfy the assumptions of the Agreement Model in section 3.<sup>44</sup> These two projects were deliberately chosen to provide one example of where the Agreement Models applies and one where it does not. Buchanan and Tullock's model satisfies the assumptions rather straightforwardly, while Gauthier's deviates from these assumptions. In so deviating, however, this latter project becomes suspect *qua* contractarian project.

There are roughly four key assumptions of the Agreement Model. One of them is novel in that it is not incorporated into existing contractarian models. This is the assumption that players are Bayesian updaters. Since the aim of this paper is to introduce the idea of rational belief formation into contractarian modelling, this assumption is simply postulated. The other three assumptions that underly the Agreement Model are generally adopted in contractarian models:

1. Unanimity as a decision rule.
2. Social-scientific uncertainty.
3. Expected utility maximization.

Contractarian models standardly accept assumptions 1–3. To the extent that a theory denies 1, it fails to justify the outcome of the decision process to each and every individual bound by social rules. To the extent that it denies 2, it posits an unrealistic degree of idealization, one that alienates the reasons of agents in the model with respect to the reasons of real individuals (section 1). And to the extent that the theory rejects 3, it rejects the most widely accepted model of individual choice under uncertainty. It must either justify the adoption of some other model of choice, or must identify the output of the model without recourse to any formal framework, leaving the reasoning opaque.<sup>45</sup> To the extent that the model of section 3 accurately describes a particular social contract, this contract loses its justificatory force. To the extent that a contract deviates from

<sup>44</sup>Other notable contractarian projects, each worth consideration in their own right, include those of early Gauthier (1986), Binmore (1994b, 2005) and Moehler (2018).

<sup>45</sup>An exciting analysis that employs an alternative model of choice, prospect theory, can be found in Chung (2018). Chung, however, focuses on Rawls's contractualist project. It remains to explore how contractarian models might operate under this alternative choice framework.

the model of section 3, an explanation is due as to which assumption has been abandoned and why.

Returning to the question asked in the first paragraph of this paper: how do hypothetical contracts bind actual agents? As the contractarian schema suggests, the normative force of contractarianism springs, not from the moral power of consent, but rather from what the modelled agreement reveals about the object of agreement.<sup>46</sup> In order to attain unanimous agreement in the model of the social contract, a set of terms must exhibit certain desirable features or desiderata on the basis of which the set is attractive to modelled agents.<sup>47</sup> The contractarian model is a normative-epistemic device for revealing that a set of terms satisfies said desiderata, which often include stability, efficiency and mutual benefit. If, however, the Agreement Model correctly represents unanimous agreement in the contractarian theory, then unanimous agreement is highly error-prone. It therefore does not ensure the satisfaction of these desiderata, and contractarian justification founders.

In *The Calculus of Consent*, James Buchanan and Gordon Tullock develop a ‘theory of the political constitution’, which involves both normative and positive elements. The main normative project is to provide an evaluative framework for constitutional design and reform. To do so, Buchanan and Tullock rely on what they call the ‘individualistic method’, which avoids positing any ethical standard for judging constitutional provisions apart from their acceptability for each and every individual. And to this end, they endorse unanimity as the test of a rule’s ethical value:

Analysis should enable us to determine under what conditions a particular individual in the group will judge a constitutional change to be an improvement; and, when all individuals are similarly affected, the rule of unanimity provides us with an extremely weak ethical criterion for ‘betterness’ . . . . We do not propose to go beyond welfare judgements deducible from a rigorous application of the unanimity rule. Only if a specific constitutional change can be shown to be in the interest of all parties shall we judge such a change to be an ‘improvement’. On all other possible changes in the constraints on human behaviour, nothing can be said without the introduction of much stronger, and more questionable, ethical precepts. (Buchanan and Tullock 2004: 14)<sup>48</sup>

Before discussing their use of uncertainty in the choice scenario, it is worth noting that Buchanan and Tullock assume ‘that the individual, as he participates in collective decisions, is guided by the desire to maximize his own utility’ (Buchanan and Tullock 2004: 26). Utility maximization is combined with an acceptance of uncertainty, both in real politics and in the model that Buchanan and Tullock construct to model political choice.<sup>49</sup>

---

<sup>46</sup>Basing social contract theories on the foundation of consent opens them up to powerful criticisms, notably those of Dworkin (1973), Simmons (1979) and Huemer (2013).

<sup>47</sup>A helpful guide through these issues is d’Agostino *et al.* (1996). See also Gaus and Thrasher (2016).

<sup>48</sup>*Nota bene* the seamless transition from unanimous agreement to universal interest fulfilment. It is the tight connection between these two notions that the Agreement Model undermines.

<sup>49</sup>In analyzing the behavior of the individual in the political process, there is an important element of uncertainty present that cannot be left out of account. No longer is there the one-to-one correspondence between individual choice and final action’ (Buchanan and Tullock 2004: 37).

In fact, Buchanan and Tullock endorse a very strong version of uncertainty. This is related to their use of the unanimity rule, for, as Buchanan notes elsewhere, ‘the costs of agreement under a unanimity rule may be extremely high or even prohibitive’ (Buchanan 1975: 55). In effect, the unanimity rule grants each party a monopoly right over a crucial resource, one which they may withhold in order to extract purely distributive gains. As Guido Calabresi and Douglas Melamed have demonstrated, such a situation gives rise to free-rider and hold-out problems, and ultimately to the failure to attain Pareto improvements (Calabresi and Melamed 1972: 1106–1110).

In order to avoid such issues while maintaining their individualistic approach, Buchanan and Tullock propose a model in which individuals must choose their constitutional provisions in a state of extreme uncertainty, or from behind what Buchanan would later call a ‘veil of uncertainty’ (Brennan and Buchanan 2000: 35). Under such conditions, Buchanan and Tullock contend, parties are unable to identify which rules favour their particular interests, and therefore, they each evaluate final outcomes as if they were a “randomly distributed” participant in the succession of collective choices anticipated’ (Buchanan and Tullock 2004: 91).<sup>50</sup> Like John Rawls in constructing his original position, Buchanan and Tullock propose uncertainty as a device of ‘normalization’. As Rawls explains, ‘a normalization of interests attributed to the parties’ is ‘common to social contract doctrines’, and through this normalization, the theorist can construct a ‘shared point of view’ that blunts sharp conflicts of interests, making agreement possible (Rawls 2008: 226).

While resolving one problem, uncertainty introduces another. When individuals lack sufficient information to fully predict the effect that rules will have on final outcomes, they rely on their own subjective prior beliefs to form educated guesses.<sup>51</sup> These agents, in seeking to maximize their utility, will naturally rely on their best guesses as to the outcomes of the various rules from which they must choose. Given that they are utility-maximizers, they will choose pivotally. And given our postulate that individuals form rational Bayesian beliefs, they will update their beliefs based on the information that pivotal voting implicitly contains. Therefore, the agents will exhibit both features of Bayesian choice.

Building from these assumptions, the Agreement Model suggests that, although agreement is often possible, the object of agreement may be undesirable. The selection of rules within groups under conditions of uncertainty leads to an unreliable belief-formation process that, in turn, produces high degrees of error: decision-makers choose rules that none of them actually want. Consequently, the idealized choice scenario fails as a normative-epistemic tool for identifying rational behavioural constraints. It thus fails also as a justificatory construction. Buchanan and Tullock’s remedy to the prohibitive costs of agreement is itself a poison – uncertainty begets Bayesian choice along with its loathsome consequences.

In its reliance on normalization and agreement, rather than a formal bargaining model, the approach of Buchanan and Tullock bears a resemblance to David Gauthier’s recent project, a project that revises *Morals by Agreement* and which

<sup>50</sup>See also Brennan and Buchanan (2000: 33).

<sup>51</sup>As noted above (section 3.1), Buchanan, writing with Roger Faith, explicitly endorses the inference from insufficient information to a diversity of prior beliefs (Buchanan and Faith 1980).

I refer to as ‘Gauthier 2.0’. Like Buchanan and Tullock, Gauthier 2.0 also endorses unanimity as the choice rule determining the contents of the social contract. Although Gauthier 2.0 aims to provide a contractarian defence of social morality, rather than of constitutional provisions or reforms, his application of unanimous agreement to the determination of rules clearly resembles the model of Buchanan and Tullock:

Take a proposed normative requirement or expectation – any one. Ask whether it could be included as part of the normative structure of a society to which you could reasonably agree were you, together with your fellows, able by everyone’s agreement to choose that structure. Now extend this to everyone – ask whether the requirement or expectation could be included as part of the structure to which everyone could reasonably agree were they able by universal agreement to choose that structure. If we can answer the questions affirmatively, then the proposed practical consideration passes the contractarian test and is eligible for inclusion in an actual society that constitutes a cooperative venture for mutual fulfillment. A person in such a society who failed to fulfill the requirement or expectation would be rightly open to criticism and perhaps sanctions . . . (Gauthier 2013: 618)

Gauthier 2.0 also accepts uncertainty, not just of social-science, but even of more local actions. This acceptance is implicit in his endorsement of expected utility maximization as ‘the best theory for one-person decision problems’ (Gauthier 2013: 604).

Where Gauthier 2.0 deviates is in his rejection of assumption 3, expected utility maximization, in the context of group decisions. For Gauthier 2.0, the collective action problems that emerge in group choice under the assumptions of orthodox rationality are not unfortunate, though intriguing, results of sound theoretical foundations. Rather, such results provide a *reductio ad absurdum* of the orthodox conception of rationality as utility-maximization: ‘Instead of supposing that an action is rational only if it maximizes the agent’s payoff given the actions of the other agents, I am proposing that a set of actions, one for each agent, is fully rational only if it yields a Pareto-optimal outcome . . . . To the maximizer’s charge that it cannot be rational for a person to take less than he can get, the Pareto-optimizer replies that it cannot be rational for each of a group of persons to take less than, acting together, each can get’ (Gauthier 2013: 606–607).

Gauthier’s revisionist account of rationality provides an interesting challenge to the hegemony of standard game-theoretical models in economics and normative philosophy, but to fully engage this challenge would drag us far afield. It suffices to note that Gauthier 2.0 has drifted from the original contractarian vision of grounding moral requirements in purely instrumental rationality. When an agent’s best reply comes into conflict with the action that is most conducive to the efficient, cooperative outcome – that is, when equilibrium and optimality diverge – the agent is modelled as choosing cooperation over preference maximization, as allowing optimality to take precedence over equilibrium. This behavioural postulate undermines Gauthier’s claim to present a contractarian project that will demonstrate the instrumental rationality of moral compliance. In proposing that

individuals would choose cooperation even at the expense of frustrating the pursuit of their ends, Gauthier 2.0 invokes something aside from instrumental rationality – perhaps moral standards or some notion of reasonableness – at a foundational point in his social contract. Gauthier 2.0 thereby renounces the *bona fide* contractarian project of grounding moral reasoning in mere instrumental reasoning.<sup>52</sup> As such, the theory of Gauthier 2.0 is outside of the purview of this paper. The model of section 3 does not apply to unorthodox conceptions of rationality insofar as such conceptions reject utility-maximization as the behavioural postulate. By showing that hyper-rational agents, seeking to advance their personal interests, do not generally produce rules worth adopting, the Agreement Model targets only contractarian models premised on instrumental rationality. To the extent that a theory rejects utility-maximization as a behavioural postulate, qualifying an agent's pursuit of his or her goals with some reasonable consideration or moral constraint, it ceases to qualify as a *bona fide* contractarian theory and, by the same token, ceases to be a target of the critique offered here.

The purposes of considering these two projects in detail is to illustrate how one might assess whether or not the Agreement Model applies to a given contractarian project. In addition, these examples reveal an important point: to avoid the Agreement Model requires the rejection of one or more of its assumptions. But these assumptions are either standard in (or even constitutive of) the contractarian approach or intuitively plausible (as in the case of Bayesian updating). In this way, the Agreement Model presents a dilemma of sorts: either accept that modelled unanimity does not identify desirable choices or reject certain plausible modelling assumptions. Thus, the present critique has important implications even for those projects that fall outside the purview of the Agreement Model.

## 5. Conclusion

Despite exhibiting wide diversity, contractarian models share several assumptions. These assumptions notably include expected utility maximization, a unanimity rule as the criteria for rule selection, and uncertainty as to the actual outcomes resulting from the selection of rules. When a standard of rational belief formation, i.e. Bayesian updating, is incorporated into the model, these assumptions together generate undesirable group choices. As noted, the contractarian has two ways out: either reject Bayesian updating as a model of rational belief formation or find some way to render it innocuous. Perhaps good reasons exist for rejecting Bayesian updating as a model of rational belief formation. Yet, given the aim of examining the implications of Bayesian belief formation for contractarian models, it is worth considering how we might embrace Bayesian updating without rejecting contractarian justification. Rather than rejecting Bayesian updating, we might defang it by varying another parameter: group size.

In light of propositions 3 and 4 (section 3.4), a natural approach would maintain the core tenets of contractarianism, along with Bayesian belief formation, but allow the group size to change so as to approximate the optimal size for the task at hand.

<sup>52</sup>For a lengthier criticism along these lines, see Moehler (2018: 59–66).

Investigating the best way to develop such a model presents a new angle from which to approach the contractarian project. Briefly, issues that arise in social interaction involve groups of various sizes. The colour of my mailbox concerns my neighbours, while the amount spent on national defence involves a widespread group of citizens. When externalities are limited, as in the mailbox case, a contractarian model should include very small numbers of agents so as to avoid the error costs implied by large groups operating under the unanimity rule. By contrast, with externalities that are more widespread, the network benefits of a single framework of rules may greatly outweigh the cost of errors that arise from a more inclusive group unit. A solid basis for this approach has been laid out by several diverse thinkers. Directly relevant are Albert Hirschman's models of organizational reform via exit (Hirschman 1970), Robert Nozick's exit-based model of utopia (Robert 1974)<sup>53</sup> and the work of Elinor and Vincent Ostrom on polycentric governance (E. Ostrom 2008a; V. Ostrom 2008b).<sup>54</sup>

Of course, social contract models come in a dizzying array of forms, aiming to accomplish diverse tasks in different ways. The critique provided here will extend only to a subset of those models and, perhaps, only to a subset of the subset containing all and only *bona fide* contractarian models. Rather than a sweeping indictment of social contract theory, I present this critique as a first check. If a choice procedure satisfies certain general assumptions – expected utility maximization, uncertainty and unanimity – then standard Bayesian belief formation implies that its output will be unsatisfactory. Contractarian theorists may produce convincing arguments that the surprising conclusions of the Agreement Model do not undermine their projects. In that case, since the conclusion follows deductively from a set of assumptions, the task is to identify which assumption – uncertainty? unanimity? expected utility maximization? Bayesian belief formation? – can be plausibly rejected. On the other hand, the critique offered here may prove to be quite general, perhaps it is even robust in the face of modifications in its assumptions. In that case, the above suggestion takes on an increased importance in pointing toward a better approach to modelling the social contract. Either way, in forcing ourselves to examine our assumptions, determining how they compare to those of section 3, we clarify to others and to ourselves the nature of our models. As a consequence, we consider more deeply how they produce their results, justify those results, and guide our practical deliberations.

## References

- Adler S.J. 1994. *The Jury: Trial and Error in the American Courtroom*. New York, NY: Crown Publishing Group.
- Asch S.E. 1951. Effects of group pressure upon the modification and distortion of judgments. In *Groups, Leadership, and Men: Research in Human Relations*, ed. H. Guetzkow, 177–190. New York, NY: Carnegie Press.
- Austen-Smith D. and J.S. Banks 1996. Information aggregation, rationality, and the Condorcet jury theorem. *American Political Science Review* 90(1), 34–45.
- Bacharach M. 1976. *Economics and the Theory of Games*. London: Macmillan Press.

<sup>53</sup>See also Kukathas (2003).

<sup>54</sup>See also Boettke *et al.* (2011).

- Bacharach M.** 1987. A theory of rational decision in games. *Erkenntnis* 27(1), 17–55.
- Binmore K.** 1994a. *Game Theory and the Social Contract. Volume 1: Playing Fair*. Cambridge, MA: MIT Press.
- Binmore K.** 1994b. *Game Theory and the Social Contract. Volume 2: Just Playing*. Cambridge, MA: MIT Press.
- Binmore K.** 2004. Reciprocity and the social contract. *Politics, Philosophy & Economics* 3(1), 5–35.
- Binmore K.** 2005. *Natural Justice*. Oxford: Oxford University Press.
- Boettke P.J., C.J. Coyne and P.T. Leeson** 2011. Quasimarket failure. *Public Choice* 149(1), 209–224.
- Bogardus T.** 2009. A vindication of the equal-weight view. *Episteme* 6(3), 324–335.
- Brennan G. and J.M. Buchanan** 2000. *The Reason of Rules: Constitutional Political Economy*. Indianapolis, IN: Liberty Fund.
- Buchanan J.M.** 1975. *The Limits of Liberty: Between Anarchy and Leviathan*. Chicago, IL: University of Chicago Press.
- Buchanan J.M.** 2002. John Rawls, Justice as fairness: a restatement. *Public Choice* 113(3), 488–490.
- Buchanan J.M. and R.L. Faith** 1980. Subjective elements in Rawlsian contractual agreement on distributional rules. *Economic Inquiry* 18(1), 23–38.
- Buchanan J.M. and G. Tullock** 2004. *The Calculus of Consent*. Indianapolis, IN: Liberty Fund.
- Calabresi G. and A.D. Melamed** 1972. Property rules, liability rules, and inalienability: one view of the cathedral. *Harvard Law Review* 85, 1089–1128. Faculty Scholarship Series, 1983. [https://digitalcommons.law.yale.edu/fss\\_papers/1983](https://digitalcommons.law.yale.edu/fss_papers/1983).
- Christensen D.** 2007. Epistemology of disagreement: the good news. *Philosophical Review* 116(2), 187–217.
- Chung H.** 2015. Hobbes's state of nature: a modern Bayesian game-theoretic analysis. *Journal of the American Philosophical Association* 1(3), 485–508.
- Chung H.** 2018. Rawls's self-defeat: a formal analysis. *Erkenntnis*. doi: 10.1007/s10670-018-0079-4.
- Cohen G.A.** 2009. *Rescuing Justice and Equality*. Cambridge, MA: Harvard University Press.
- Cudd A. and S. Eftekhari** 2000. Contractarianism. In *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition), ed. E.N. Zalta. <https://plato.stanford.edu/archives/sum2018/entries/contractarianism/>.
- d'Agostino F., G. Gaus and J. Thrasher** 1996. Contemporary approaches to the social contract. In *The Stanford Encyclopedia of Philosophy* (Fall 2019 Edition), ed. E.N. Zalta. <https://plato.stanford.edu/archives/fall2019/entries/contractarianism-contemporary/>.
- Darwall S.** 2008. *Contractarianism/Contractualism*. Hoboken, NJ: Blackwell Publishing.
- Dworkin R.** 1973. The original position. *The University of Chicago Law Review* 40(3), 500–533.
- Elga A.** 2007. Reflection and disagreement. *Noûs* 41(3), 478–502.
- Feddersen T.J. and W. Pesendorfer** 1996. The swing voter's curse. *American Economic Review* 86, 408–424.
- Feddersen T. and W. Pesendorfer** 1998. Convicting the innocent: the inferiority of unanimous jury verdicts under strategic voting. *American Political Science Review* 92(1), 23–35.
- Feddersen T. and W. Pesendorfer** 1999. Elections, information aggregation, and strategic voting. *Proceedings of the National Academy of Sciences USA* 96(19), 10572–10574.
- Gaus G.** 2008. *On Philosophy, Politics, and Economics*. Belmont, CA: Thomson Wadsworth.
- Gaus G.** 2010. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge: Cambridge University Press.
- Gaus G.** 2013. Hobbesian contractarianism, orthodox and revisionist. In *The Bloomsbury Companion to Hobbes*, ed. S. Lloyd, 263–278. New York, NY: Bloomsbury.
- Gaus G.** 2016. *The Tyranny of the Ideal: Justice in a Diverse Society*. Princeton, NJ: Princeton University Press.
- Gaus G. and J. Thrasher** 2016. Rational choice in the original position: the (many) models of Rawls and Harsanyi. In *The Original Position*, ed. T. Hinton, 39–58. Cambridge: Cambridge University Press.
- Gauthier D.** 1969. *The Logic of Leviathan: The Moral and Political Theory of Thomas Hobbes*. Oxford: Oxford University Press.
- Gauthier D.** 1986. *Morals by Agreement*. Oxford: Oxford University Press.
- Gauthier D.** 2013. Twenty-five on. *Ethics* 123(4), 601–624.
- Hampton J.** 1988. *Hobbes and the Social Contract Tradition*. Cambridge: Cambridge University Press.
- Harsanyi J.C.** 1967. Games with incomplete information played by “Bayesian” players, i–iii. *Management Science* 14(3), 159–182.
- Hausman D.M.** 1992. *The Inexact and Separate Science of Economics*. Cambridge: Cambridge University Press.
- Hausman D.M.**, ed. 1994. Why look under the hood. In *The Philosophy of Economics: An Anthology*, 217–221. Cambridge: Cambridge University Press.

- Hausman D.M.** 2003. Rational belief and social interaction. *Behavioral and Brain Sciences* **26**(2), 163–164.
- Hirschman A.O.** 1970. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Cambridge, MA: Harvard University Press.
- Hobbes T.** 1991. *Leviathan*. Cambridge: Cambridge University Press.
- Huemer M.** 2013. *The Problem of Political Authority*. London: Palgrave Macmillan.
- Hume D.** 1978. *A Treatise of Human Nature*, ed. L.A. Selby-Bigge. Oxford: Clarendon.
- Kalven Jr H. and H. Zeisel** 1966. *The American Jury*. Boston, MA: Little, Brown and Company.
- Kant I.** 1996. *Practical Philosophy*. Cambridge: Cambridge University Press.
- Kavka G.S.** 1986. *Hobbesian Moral and Political Theory*. Princeton, NJ: Princeton University Press.
- Knight F.H.** 1971. *Risk, Uncertainty and Profit*. Chicago: University of Chicago Press.
- Kreps D.M.** 2012. *Microeconomic Foundations I: Choice and Competitive Markets*, Volume 1. Princeton, NJ: Princeton University Press.
- Kreps D.M. and R. Wilson** 1982. Sequential equilibria. *Econometrica* **50**, 863–894.
- Kukathas C.** 2003. *The Liberal Archipelago: A Theory of Diversity and Freedom*. Oxford: Oxford University Press.
- Kymlicka W.** 1991. *Liberalism, Community, and Culture*. Oxford: Oxford University Press.
- Lehrer K.** 1976. When rational disagreement is impossible. *Noûs* **10**, 327–332.
- Levine J.P.** 1992. *Juries and Politics*. Boston, MA: Brooks/Cole Publishing.
- Locke J.** 1980. *Second Treatise of Government*. Indianapolis, IN: Hackett.
- Loungani P.** 2001. How accurate are private sector forecasts? cross-country evidence from consensus forecasts of output growth. *International Journal of Forecasting* **17**(3), 419–432.
- Mas-Colell A., M.D. Whinston and J.R. Green** 1995. *Microeconomic Theory*. Oxford: Oxford University Press.
- Matheson J.** 2015. Disagreement and the ethics of belief. In *The Future of Social Epistemology: A Collective Vision*, ed. J. Collier, 139–148. Lanham, MD: Rowman and Littlefield.
- Matheson, J. and B. Frances** 2018. Disagreement. In *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), ed. E.N. Zalta. <https://plato.stanford.edu/archives/win2019/entries/disagreement/>.
- McCart S.W.** 1964. *Trial by Jury: A Complete Guide to the Jury System*. Boston, MA: Chilton Books.
- Moehler M.** 2010. The (stabilized) Nash bargaining solution as a principle of distributive justice. *Utilitas* **22**(4), 447–473.
- Moehler M.** 2018. *Minimal Morality: A Multilevel Social Contract Theory*. Oxford: Oxford University Press.
- Muldoon R.** 2016. *Social Contract Theory for a Diverse World: Beyond Tolerance*. London: Routledge.
- Muldoon R., C. Lisciandra, M. Colyvan, C. Martini, G. Sillari and J. Sprenger** 2014. Disagreement behind the veil of ignorance. *Philosophical Studies* **170**(3), 377–394.
- Oakeshott M.** 1965. *Rationalism in Politics and other Essays*. Indianapolis, IN: Liberty Fund.
- Ostrom E.** 2008a. Tragedy of the commons. In *The New Palgrave Dictionary of Economics* (2nd edn), ed. S.N. Durlauf and L.E. Blume. Basingstoke: Palgrave Macmillan.
- Ostrom V.** 2008b. *The Intellectual Crisis in American Public Administration*. Tuscaloosa, AL: University of Alabama Press.
- Poincaré H.** 1920. *Science et Méthode*. Paris: Ernest Flammarion, Éditeur.
- Rawls J.** 2008. *Lectures on the History of Political Philosophy*. Cambridge, MA: Harvard University Press.
- Rawls J.** 2009. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
- Robert N.** 1974. *Anarchy, State, and Utopia*. New York, NY: Basic Books.
- Rubinstein A.** 1982. Perfect equilibrium in a bargaining model. *Econometrica* **50**, 97–109.
- Scanlon T.** 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.
- Schmidtz D.** 1990. Justifying the state. *Ethics* **101**(1), 89–102.
- Simmons A.J.** 1979. *Moral Principles and Political Obligations*. Princeton, NJ: Princeton University Press.
- Southwood N.** 2009. Moral contractualism. *Philosophy Compass* **4**(6), 926–937.
- Thrasher J.** 2019. Constructivism, representation, and stability: path-dependence in public reason theories of justice. *Synthese* **196**(1), 429–450.
- Vanderschraaf P.** 2006. War or peace? A dynamical analysis of anarchy. *Economics and Philosophy* **22**(2), 243–279.
- Wagner C. and K. Lehrer** 1981. *Rational Consensus in Science and Society*. Dordrecht: Reidel.
- Wit J.** 1998. Rational choice and the Condorcet jury theorem. *Games and Economic Behavior* **22**(2), 364–376.



## Appendix

### 0. Lemma

**Proof.** The following arithmetic establishes this claim:

$$\begin{aligned}
 U_i(\mathbf{A}|k) &= (k, n)U_i(\mathbf{A}, G) + (1 - \beta(k, n))U_i(\mathbf{A}, B) \\
 &= \beta(k, n)(0) + (1 - \beta(k, n))(-q) \\
 &= -q + \beta(k, n)(q)
 \end{aligned} \tag{1}$$

$$\begin{aligned}
 U_i(\mathbf{R}|k) &= \beta(k, n)U_i(\mathbf{R}, G) + (1 - \beta(k, n))U_i(\mathbf{R}, B) \\
 &= \beta(k, n)(-(1 - q)) + (1 - \beta(k, n))(0) \\
 &= -\beta(k, n) + \beta(k, n)(q)
 \end{aligned} \tag{2}$$

$$\begin{aligned}
 q < \beta(k, n) &\Rightarrow \\
 U_i(\mathbf{A}, k) &= -q + \beta(k, n)(q) \\
 &> -\beta(k, n) + \beta(k, n)(q) = U_i(\mathbf{R}, k) \\
 U_i(\mathbf{A}|k) &> U_i(\mathbf{R}|k)
 \end{aligned} \tag{3}$$

□

### 1. Proposition 1<sup>a</sup>

**Proof.** Recall equation (T). From Bayes’s rule, we determined that the posterior probability that the rule is Good, G, contingent upon observing n signals, k of which are good, g, is:

$$\beta = \frac{[p^k(1 - p)^{n-k}]}{[p^k(1 - p)^{n-k} + p^{n-k}(1 - p)^k]} \tag{4}$$

Each voter considers only the situation in which he or she is pivotal, since otherwise the voter’s choice cannot affect the payoff (i.e. all actions in the non-pivotal scenario are best-responses). Under the unanimity rule, a vote is pivotal if and only if all other voters have voted Accept. That is, if and only if:  $\forall j \in N : j \neq i : s_j = A$ .

In this case, confining ourselves to player type  $t_i = b$ , the posterior probability that the rule is good is:

$$\beta(n - 1, n) = \frac{p^{n-1}(1 - p)}{p^{n-1}(1 - p) + p(1 - p)^{n-1}} \tag{5}$$

Voter  $i$  will therefore vote to Accept, A, the rule when:

$$\begin{aligned}
 \beta U_i(\mathbf{A}, G) + (1 - \beta)U_i(\mathbf{A}, B) &= (1 - \beta)(-q) \\
 &> \beta U_i(\mathbf{R}, G) + (1 - \beta)U_i(\mathbf{R}, B) \\
 &= -\beta(1 - q)
 \end{aligned} \tag{6}$$

i.e. when  $q < \beta$ .

This last condition holds for decision-makers of both signals,  $t_i \in \{g, b\}$ .

In particular, for sufficiently large n,  $i$  will prefer to vote A, even if  $t_i = b$ . This is due to two facts:

(a)  $q \in (0, 1)$ , and (b)

$$\lim_{n \rightarrow \infty} \beta(n - 1, n) = 1$$

- Jointly, (a) and (b) ensure that there exists an  $n^*$  such that  $[n > n^*] \Rightarrow q < \beta(n - 1, n)$ . □

---

<sup>a</sup>Proposition 2, as well as propositions 3(i) and 3(iii) closely follow proofs given in Feddersen and Pendorfer (1998).

2. Proposition 2

**Proof.** Let  $n$  be sufficiently large that informative voting is not a Nash Equilibrium. That is, let  $n$  be sufficiently large to imply that  $q < \beta(n - 1, n)$ .

In addition, we rule out non-responsive Nash equilibria, for example, the equilibrium where all vote Reject regardless of what signals they receive. These equilibria are definitional oddities, trivial and uninteresting. (Formally, a responsive equilibrium is an equilibrium in which  $\varrho_G \neq \varrho_B$ , with these two probabilities defined below.)

Given that  $q < \beta(n - 1, n)$ , informative voting is not a Nash equilibrium, i.e. the strategy of choosing A when  $t_i = g$  and R when  $t_i = b$ , is not a best response. In addition, we can trivially rule out the strategy where voters always vote Accept when they get a b-signal, and Reject when they get a g-signal. Therefore, the only remaining possibility for an equilibrium is one where voters employ mixed strategies.

The only situation where a voter may improve his or her payoff by voting contrary to his or her signal is when that voter is pivotal, i.e.  $(n - 1)$  others have voted to Accept the rule. Thus, where  $\sigma(b)$  is the voter's probability of voting Accept after receiving a bad signal and  $\sigma(g)$  is the voter's probability of voting Accept after receiving a good signal, the mixed strategy profile we seek will have, for each  $i \in N$ ,  $(\sigma(b), \sigma(g)) : \sigma(b) \in (0, 1)$  and  $\sigma(g) = 1$ . (Because I am computing a symmetric Nash equilibrium, I will drop the subscript  $i$ , since all voters who receive the same signal will vote the same.)

Define two probabilities:

1.  $\varrho_G := \Pr(\text{voter Accepts} | G) = \sigma(b)(1 - p) + \sigma(g)p$
2.  $\varrho_B := \Pr(\text{voter Accepts} | B) = \sigma(b)p + \sigma(g)(1 - p)$

Because the equilibrium strategy profile is mixed, we must have  $\sigma(b) > 0$ , and we also know that a pivotal voter who has received a bad signal must be indifferent between Accept and Reject (this is just the equilibrium condition for a mixed strategy). This indifference relation yields:

$$\Pr(G | \text{pivotal}, t_i = b) = q, \forall i \in N \tag{1}$$

Applying Bayes's rule gives us:

$$\Pr(G | \text{pivotal}, t_i = b) = \frac{\varrho_G^{n-1}(1 - p)}{\varrho_G^{n-1}(1 - p) + \varrho_B^{n-1}p} \tag{2}$$

Combining equations 1 and 2 we get:

$$q = \frac{\varrho_G^{n-1}(1 - p)}{\varrho_G^{n-1}(1 - p) + \varrho_B^{n-1}p} \tag{3}$$

Substituting in our expressions for  $\varrho_G$  and  $\varrho_B$  gives:

$$q = \frac{(1 - p)(\sigma(b)(1 - p) + \sigma(g)p)^{n-1}}{(1 - p)(\sigma(b)(1 - p) + \sigma(g)p)^{n-1} + p(\sigma(b)p + \sigma(g)(1 - p))^{n-1}} \tag{4}$$

Noting that, as argued above,  $\sigma(g) = 1$  and simplifying yields:

$$\frac{(1 - p) + \sigma(b)p}{\sigma(b)(1 - p) + p} = \frac{(1 - q)(1 - p)^{1/(n-1)}}{pq} \tag{5}$$

Solving for  $\sigma(b)$ :

$$\sigma_i^*(p, q, n) = \frac{\frac{(1 - q)(1 - p)^{1/(n-1)}}{qp} p - (1 - p)}{p - \frac{(1 - q)(1 - p)^{1/(n-1)}}{qp} (1 - p)} \tag{*}$$

This equation is the Nash equilibrium identified in proposition (2) in the main text. □

**3. Proposition 3(ii)(iii)**

**Proof.** Given our Nash equilibrium strategy profile, we can calculate type 1 error:

$$\begin{aligned} Pr(\mathbf{A}|B) &= (\varrho_B)^n = [\sigma^*(b)p + \sigma^*(g)(1-p)]^n \\ &= \left[ (1-p) + p \left( \frac{\left( \frac{(1-q)(1-p)}{qp} \right)^{1/(n-1)} p - (1-p)}{p - \left( \frac{(1-q)(1-p)}{qp} \right)^{1/(n-1)} (1-p)} \right) \right]^n \\ &= \left[ \frac{\left( \frac{(1-q)(1-p)}{qp} \right)^{1/(n-1)} (2p-1)}{\left( p - \frac{(1-q)(1-p)}{qp} \right)^{1/(n-1)} (1-p)} \right]^n \end{aligned}$$

Since  $p > 0.5$ , the numerator of this expression is greater than the denominator, and therefore it must be increasing in  $n$ .

To establish that type 1 error is increasing ‘quickly’, not formally defined, it will suffice to find the limit and to provide an example (in the main body) that demonstrates that with relatively low  $n$ , our expression closely approximates the limit.

Feddersen and Pesendorfer (1998) show that the limit of type 1 error as  $n$  approaches infinity is:<sup>b</sup>

$$\left( \frac{(1-q)(1-p)}{qp} \right)^{-p/(2p-1)}$$

Numerical examples demonstrate that even at relatively low  $n$ , type 1 error approaches this limit. (See, for example, Figure 2 in the main text.) □

**4. Proposition 4**

**Proof.** When  $t_i = g$ , Accept is always a best response (this is because either the voter is pivotal, in which case all others have voted Accept, which only serves to increase voter  $i$ ’s confidence that the rule is good, or else the voter is not pivotal, in which case all actions yield an identical payoff). We therefore only need to show that for  $q \geq 0.5$ , there is an  $\bar{n}$  such that voter  $i$  prefers to reject even when pivotal. According to our equation T and our Lemma, we only need to show that for  $q \geq 0.5$  there will always be some  $n = \bar{n}$  such that  $q \geq \beta$ .

When  $n = 1$  the condition for voting informatively,  $q \geq \beta$ , is satisfied trivially: when voter  $i$  is pivotal and the signal is  $t_i = b$ , then

$$\beta = \frac{p^0(1-p)^1}{p^0(1-p)^1 + p^1(1-p)^0} < 1/2 \leq q.$$

This suffices to establish proposition 4 (with  $\bar{n} = 1$ ). So the proof is complete.

For further elaboration, note that the proposition also holds necessarily for  $\bar{n} = 2$ . At  $n = 2$ , when  $t_i = b$  and voter  $i$  is pivotal, we have:

$$\beta = \frac{p^1(1-p)^1}{p^1(1-p)^1 + p^1(1-p)^1} = 1/2 \leq q$$

Finally, note that  $\beta$  is increasing in  $n$ . Hence, as  $q$  increases, informative voting can be maintained as a Nash equilibrium for higher and higher  $n$ . For example, if  $p = 0.51$  (very low accuracy of signals) and  $n = 5$ , then

---

<sup>b</sup>For the derivation, see Feddersen and Pesendorfer (1998: 32).

$$\begin{aligned}\beta &= \frac{p^4(1-p)^1}{p^4(1-p)^1 + p^1(1-p)^4} \\ &= \frac{(0.51)^4(0.49)}{(0.51)^4(0.49) + (0.49)^4(0.51)} \\ &\simeq 0.52997.\end{aligned}$$

Therefore, at  $n = 5$ , informative voting will be a sustainable equilibrium for  $q \geq 0.52997$ , since such a  $q$  ensures that rejecting when  $t_i = b$  is a best-response, even when  $i$  is pivotal.  $\square$

**Alexander Schaefer** is a PhD student in philosophy at The University of Arizona and a Politics, Philosophy, Economics and Law Fellow at the Freedom Center. His research interests include social contract theory, political economy and social complexity. His current scholarship focuses on applying the insights of complexity theory to assess the proper scope of state action. URL: <https://freedomcenter.arizona.edu/alex-schaefer>.