RMM Vol. 4, 2013, 131–156
Special Topic: Can the Social Contract Be Signed by an Invisible Hand?
http://www.rmm-journal.de/

Peter Vanderschraaf

A Governing Convention?[*]

Abstract:

In this essay I argue that one can understand the relationship between those who rule and those who are ruled in civil society as an implicit contractual relationship or *contract by convention*. I use variations of the extensive form *Trust Game* to summarize the structures of alternative forms of contracts, and apply these variations to model the relationship between the rulers and those under their rule. One of these variations, the *Irrevocable Sovereignty Game*, summarizes Hobbes' main argument for why it is conceptually impossible for a contract to exist between a sovereign and the subjects under its rule. I argue that Hobbes' argument presupposes a common understanding of a contract as a set of promises enforceable by a third party, such as a legally binding agreement. I use another variation of the Trust Game, the *Repeatable Sovereignty Game*, to show that rulers and ruled can establish and maintain a convention requiring the ruled to obey their rulers' commands in return for these rulers providing the ruled satisfactory benefits. In effect, the ruled and their rulers create an implicit contract that is self-enforcing rather than an explicit contract requiring third-party enforcement. I argue that this idea of a *governing convention* has roots in David Hume's discussions of government, and is even implicit in Hobbes' own treatment of sovereignty.

Keywords: Governing Convention, Simple Trust Game, Trust Problem, Irrevocable Sovereignty Game, Repeatable Sovereignty Game.

1. Introduction

In *A Treatise of Human Nature*, David Hume famously declares "Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them" (*Treatise* 2.3.3:4).[1] In his recent book *Understanding Plato's Republic*, Gerasimos Santas argues that Hume's position effectively stands Plato on Plato's own head, since Plato in *Republic* argues that the rational part of the soul should rule the appetitive and the spirited parts (2010, 93–100). I think Hume stands Plato on his head in a second way. In *Re-*

[*] I thank Hartmut Kliemt and Bernd Lahno for their invaluable comments on an earlier version of this essay.

[1] I will reference Hume's *A Treatise of Human Nature* (2000[1740]) by section and paragraph number, Hobbes' *Elements of Law* (1994[1640]), *De Cive* (1998[1642]) and *Leviathan* (1991[1651]) by chapter and paragraph number, and Locke's *Second Treatise of Government* (1988[1690]) by section number. Here and below I use "*Treatise*" to refer to *A Treatise of Human Nature* and "*Second Treatise*" to refer to *Second Treatise of Government*.

public, Socrates ultimately rejects Glaucon's proposal that justice is the product of a social contract (358e–359b).[2] The Socrates of *Crito* defends the claim that a social contract is at least part of what establishes one's duty to obey the laws of one's state (51d–52d). Assuming Plato maintained essentially consistent views over the period he composed *Crito* and *Republic*, using contemporary vocabulary we might say that at least during this period Plato rejected a contractarian explanation of the origins of morality, but embraced a contractarian defense of political obligation. I think Hume would argue that Plato has it exactly backwards. Moreover, and perhaps more controversially, I believe Hume's discussions of the social contract establish him as an important forerunner of contemporary moral and political contractarianism. I also believe that Hume's analysis points to a way of understanding the relationship between rulers and the ruled in a civil society as a kind of contract, namely a *contract by convention*.[3]

The social contract is an ancient idea, with a long and multi-faceted tradition. According to one common recounting of the history of pre-20th century political philosophy, contractarian discussions of the state first emerge in the Sophist tradition, reappear from time to time in the Roman and Scholastic traditions, reach full flower in the early modern tradition of Hobbes, Pufendorf, Locke and Rousseau, and then suffer abrupt and apparently decisive refutation at the hands of David Hume. Starting in the mid-20th century, contractarianism has enjoyed a tremendous revival, led by the landmark works of John Rawls, Alan Buchanan and David Gauthier. But exactly what is the social contract? In fact, contractarianism encompasses a large family of both political and moral theories.[4] In its most generic sense, a social contract for a given society corresponds to a body of norms that are to regulate the interactions of its members. When Glaucon discusses justice in *Republic* and Gerald Gaus discusses social morality in *The Order of Public Reason* (2011), they discuss the social contract in this generic sense. There is also a more restricted sense of the social contract corresponding specifically to the state that regulates civil society. Locke in *The Second Treatise of Government* and Rousseau in *The Social Contract* concentrate on this segment of contractarianism. I adopt Jean Hampton's helpful terminology and will refer to theories with the more generic scope as *moral contractarian* theories and those focusing on the state as *state contractarian* theories (1991, 32). There are also normative and explanatory sides of the contractarian tradition. The normative side tries to answer the question, 'Why are we obliged to obey certain institutions?', while the explanatory side tries to answer the question, 'How did we come to be regulated by certain institutions?'. The normative and the

[2] Like many, I take Glaucon's proposal in Book II of *Republic* to be the most important contractarian proposal Plato considers in this work. By the time Plato composed the *Laws*, he may have had a somewhat more sympathetic view of a contractarian explanation of some parts of morality. See, for example, *Laws* 793b–d.

[3] Russell Hardin introduced the term 'contract by convention' to refer to the sort of implicit mutual understanding that can support coordination, which is essentially a convention as Hume defines it. See especially Hardin (1982, chapters 10, 12, 13; 1999, chapter 3).

[4] In the inaugural issue of *Journal of Applied Ethics and Philosophy*, Michael Davis (2009) published an essay titled "Fourteen Kinds of Social Contract"!

explanatory sides correspond roughly to the views that a social contract is the product of rational choice or of cultural evolution, respectively. John Rawls and David Gauthier present examples of contemporary normative moral contractarian theories rooted in rational choice in *A Theory of Justice* (1971) and *Morals by Agreement* (1986), while Robert Sugden and Brian Skyrms present examples of contemporary explanatory moral contractarian theories rooted in evolution in *The Economics of Rights Co-operation and Welfare* (2004[1986]) and *Evolution of the Social Contract* (1996).

Until fairly recently, David Hume was viewed primarily as the archetypical anti-contractarian because of his severe criticisms of actual consent theories of government. To recap Hume's celebrated arguments only briefly:[5] Actual consent does no relevant explanatory work, since all the governments of recorded history were created by force. And actual consent does little if any of the real work in establishing political obligation, in no small part because most people do not in fact consent either explicitly or tacitly to be ruled by their governments. Plainly, Hume will have no truck with actual consent state contractarianism. But since the end of the 1970s philosophers can no longer so complacently label Hume an opponent of contractarianism. For in 1979 David Gauthier (1979) and in 1986 Robert Sugden (2004[1986]) independently presented compelling arguments in their works that David Hume is a contractarian of a different stripe. Gauthier and Sugden argued that Hume is in fact a moral contractarian who understands the requirements of justice as a special class of conventions, many of which have emerged as the result of salience or evolutionary forces rather than explicit agreement. I will not review Gauthier's and Sugden's arguments here, although I do think it worth noting that in the two main *Treatise* passages where Hume gives his pathbreaking analysis of convention (*Treatise* 3.2.2:10, 22), he uses the phrase "convention or agreement", indicating that for Hume a convention is a sort of implicit contract. Several authors, including especially Ken Binmore (1994; 1998), Skyrms (1996) and Sugden himself (2004[1986]), have presented fine contemporary explanatory moral contractarian theories incorporating game theory and evolutionary dynamics. These authors all acknowledge their intellectual debt to Hume.

In the sections to follow, I present the outlines of an explanatory contractarian account of the state rooted in convention. I, too, draw inspiration from Hume, although Hume is not usually thought of as a state contractarian of any sort. I also draw inspiration from Hobbes, although this might surprise Hobbes since I propose a conventional relationship between rulers and the ruled while Hobbes' own contractarian theory as I interpret it is a rational choice theory that aims to justify the authority of a sovereign with near absolute authority over its subjects.[6] And I draw inspiration from the late Jean Hampton's more recent work, as Hampton explored the possibility of a conventionalist account of government (1986, chapters 8, 9; 1990; 1997, chapter 3). Indeed, here I appro-

[5] Hume gives these critiques of actual consent contractarianism mainly in *Treatise* 3.2.8–9 and the essay "Of the Original Contract" (1994, 186–201).

[6] I discuss Hobbes' contractarian account of the state in Vanderschraaf (forthcoming, chapter 6).

priate Hampton's term "governing convention" (1997, 78) in her honor, and will use this term to refer to the implicit contract that I will argue can exist between rulers and those they rule.[7] Hampton's analysis of the relationship between the ruled and their rulers in a state is primarily conceptual. My analysis here is somewhat more formal, in that I focus on the structure of the governing convention and present a game theoretic model summarizing this convention.[8] One may read this essay as complementing Hampton's earlier works. In *section 2* I review the *Trust Problem* and *Simple Trust* extensive form games, which summarize the well-known problem of completing a mutually beneficial sequential exchange, and discuss one common view of how contracts "solve" these problems. In *section 3* I introduce an *Irrevocable Sovereignty Game* that summarizes Hobbes' understanding of the relationship between a sovereign and its subjects. I also discuss how this game is related to a *regress argument* Hobbes employs to support his claim that a contractual relationship between sovereign and subjects is conceptually impossible. In *section 4* I introduce an alternate *Repeatable Sovereignty Game* model of the relationship between rulers and the ruled. Here I argue that the Repeatable Sovereignty Game model shows how one can avoid the trap Hobbes tries to set with his regress argument and how a governing convention is indeed possible. I also argue that Hume's account of the origins and the maintenance of government incorporates the idea of a governing convention in an informal manner, and that the possibility of a governing convention is implicit in Hobbes' own works. In the closing *section 5* I discuss possible extensions of the Repeatable Sovereignty Game model that may motivate future work.

2. Trust in Sequential Exchange

An investor is capable of increasing an initial monetary stake by a factor of $\alpha > 1$ but lacks money. An individual with money but lacking the investor's special skills can supply the investor with an initial stake in expectation of a greater return from the larger final sum of money the investor's activities generate. But if this individual does provide the investor an initial stake, what is to prevent the investor from pocketing this entire larger final sum he eventually generates using this stake and returning nothing to his provider? The extensive form game of *figure 1* summarizes this *Trust Problem*.

[7] In her earlier 1990 essay, Hampton uses the term 'leadership convention'. Hampton uses these terms in a somewhat wider sense than I will use the term 'governing convention'. For Hampton, the governing convention defines government offices and officeholders as well as the authority of the governors. In this essay, the governing convention refers specifically to the convention that characterizes this authority.

[8] Hampton does propose a simple game-theoretic model of the relationship between a ruler and the people under its rule, but as I will argue below in *section 3* (note 29), Hampton's model is flawed, and does not in fact characterize a convention.

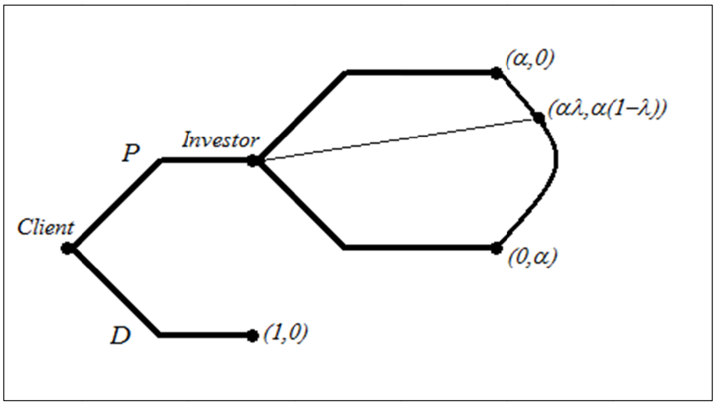


Figure 1: Trust Problem

In this game, the Client moves first, either *performing* (*P*) by providing the Investor a stake worth 1-utile to the Client or *defecting* (*D*) by declining to provide a stake.[9] If the Client performs, the Investor then moves and returns a fraction λ∈[0, 1] of the final sum of money to the Client, keeping 1 – λ for himself. In this game I assume the initial stake is also worth 1-utile to the Investor and that the utility increases for them both linearly according to the money each keeps in the end. Consequently the best possible outcome for the Client occurs when she performs and then receives α while the Investor receives 0, while the best possible outcome for the Investor occurs when the Client performs and then the Investor keeps α while the Client receives 0. The *figure 1* Trust Problem embeds a *Dictator Game* into a larger extensive form game. Should the Client perform and thereby give the Investor the opportunity to move, then in the proper subgame the Investor simply chooses a final division of the sum α and the Client then must simply accept what the Investor chooses to give her, even if the Investor chooses to give her nothing.[10] The larger Trust Problem has a unique subgame perfect equilibrium where the Client defects and the Investor returns

[9] Following the usual conventions for extensive form games, in each of the games discussed here, at each final outcome the payoff for the agent who moves for the first time at the n-th stage is the n-th coordinate of the corresponding payoff vector. For example, in the *figure 1* game at each outcome the Client's (Investor's) payoff is the first (second) coordinate of the corresponding payoff vector.

[10] The Dictator Game has been the subject of many laboratory and field experimental studies. Assuming that the payoff function of the agent in the role of the dictator depends only upon the share 1 – λ of the fixed amount of money at stake she chooses to keep and strictly increases with 1 – λ, λ = 0 characterizes the unique equilibrium. Nevertheless in many experimental studies across various cultures the distributions of the offer λ to the recipient are somewhat dispersed across the interval [0, 1), although in some studies λ = 0 is the modal offer. Explanations for this phenomenon include attributing some altruism or the influence of fairness norms to dictator agents. For summaries of some important laboratory experiments, see Roth (1995) and Camerer (2003, §2.1). Some of the important field experiments using dictator games are summarized in Henrich et al. (2004).

$\lambda = 0$ if the Client performs, and at this equilibrium the Client ends up with 1 and the Investor with 0. Of course, the Client and Investor can both do better if the Client performs and the Investor returns a sufficiently large share of α to the Client. For example, in the *Simple Trust Game* of *figure 2*, $\alpha = 4$ and $\lambda = \frac{1}{2}$, and here the Investor performs by returning 2 to the Client and defects by keeping all 4 for himself.

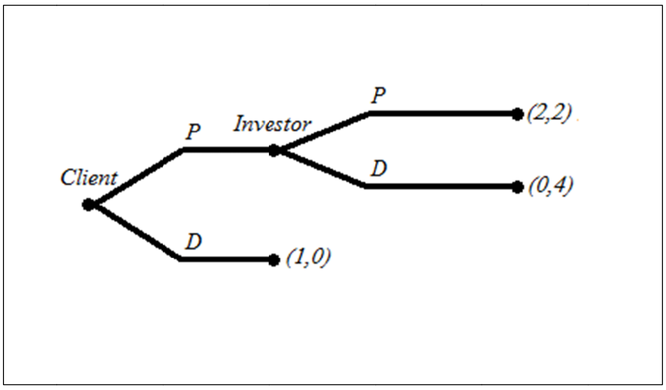


Figure 2: Simple Trust Game

In this game, if the Client and the Investor both perform then each achieves a payoff of 2 which is better than her equilibrium payoff. The Simple Trust Game is sometimes referred to as a *one sided Prisoners' Dilemma* because *D* is the weakly dominant strategy for one agent only, namely, the Investor, but this results in the suboptimal profile (*D, D*) being the only rationalizable strategy profile of the game. The Trust Problem and Simple Trust Game summarize a variety of possible exchanges of goods or labor between two parties where each has little or no direct concern for how the other party benefits from the exchange.[11]

What could remove the temptation to defect in an exchange? The two parties could exchange promises to perform, creating a contract. But why would they keep their promises? To break a promise is to violate justice, but this is no real reply if one does not expect the parties to do what justice requires. If the parties are in a Hobbesian State of Nature, making a contract to complete their exchange would evidently be pointless. As Hobbes himself puts it, "covenants without the sword are but words, and of no strength to secure a man at all (*Leviathan* 17:2)".[12] A penalty for failing to perform as promised in a contract

[11] These exchanges are in a sense unilateral since one side generates all of the cooperative surplus if the other side chooses to invest initially. The Prisoners' Dilemma summarizes a variety of bilateral exchanges where each side can increase the cooperative surplus.

[12] See also *Leviathan* 14:18.

is one “sword” that might make such a promise more than cheap talk. If p_i denotes the cost of such a penalty for Party i, or the expected cost if Party i suffers the penalty with a certain probability, for breaking a promise to perform in an exchange with the Simple Trust Game structure, then the *figure 3* game summarizes the parties’ revised situation.

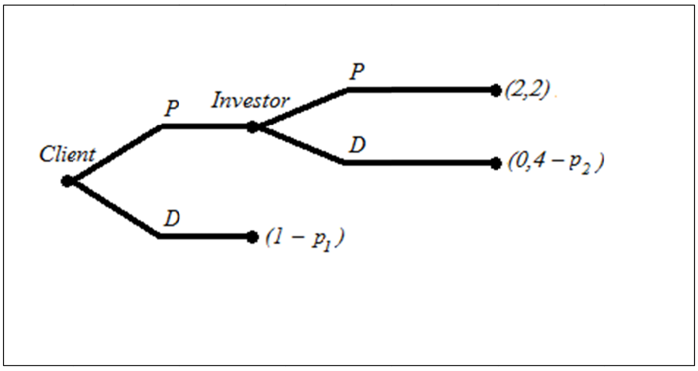


Figure 3: Simple Trust Game Augmented with Penalties

For sufficiently large values of p_i, P becomes each party’s dominant strategy and (P, P) is then the unique equilibrium.[13] The *figure 3* game summarizes in rough manner the strategic structure of the interaction between two parties who have formed a contract enforceable by some third party, typically the officers of a legal system in civil society, capable of meting out punishment in case of failure to perform. If such a third party exists and is willing to administer appropriate punishments against defectors, then the parties have sufficient reason to honor the contract they have formed.

3. Irrevocable Sovereignty and Hobbes’ Regress Argument

I introduced the Trust Problem and Simple Trust Games in *section 2* as foils for illustrating two fundamental points. First, in order to create and maintain a government, those who are to be governed must recognize and support certain offices and the individuals who occupy these offices as their governors, with the caveat that because of their recognition and support, these governors have certain powers over them all, including possibly coercive powers, that ordinarily no

[13] To be more precise, P is the Client’s strictly dominant strategy if $p_1 > 1$ and P is the Investor’s weakly dominant strategy if $p_2 > 2$.

one of them would willingly give others over herself alone. And second, the social contract metaphor for understanding government is based upon interaction structures that are *self-enforcing* rather than contractual according to a certain common understanding of contracts. I will expand upon these points in the following two sections, with an eye towards explaining more clearly how one can understand government in contractarian terms.

I begin with the second point. As several others, including in particular Jean Hampton (1986, §9.2) and Russell Hardin (1999, chapter 3), have argued, the problem of creating and maintaining government is rather different from that of creating and enforcing a contract, at least in the sense of a legally binding contract. I agree with Hampton and Hardin that the metaphor of a social *contract* is somewhat misleading. But I also agree with Hampton that the terminology associated with legally binding contracts is so entrenched in contractarian philosophy that one may as well go along with the use of this terminology. I think it has become a convention among philosophers, in the specialized sense of convention discussed by David Lewis (1969) and his successors, to use 'social contract' and 'contractarianism' to refer to theories that are not really about contracts as we typically think of them!

Hobbes' account of the relationship between sovereign and subjects illustrates this last claim especially well. Hobbes gives a contractarian analysis of the state that is both one of the greatest state social contract theories of the early modern tradition and one that anticipates contemporary contractarian theories in a variety of interesting ways.[14] Hobbes also uses the sort of vocabulary that has become standard among philosophical contractarians. Hobbes defines a *contract* as a reciprocal transfer of rights between two or more parties (*De Cive* 2:9; *Leviathan* 14:9). Each party of a contract performs her part by delivering that which she has transferred her right over to the other party or parties of the contract. In the case where at least one of the parties does not perform immediately upon transferring her rights, so that the others must trust her to perform at some future time, Hobbes calls the contract an *agreement* or *covenant* (*De Cive* 2:9; *Leviathan* 14:11). An express sign of contract indicating future performance is a promise (*Leviathan* 14:13), and performance after the creation of the covenant is keeping one's promise or *keeping faith* (*Leviathan* 14:11). In the problem of the Simple Trust Game of *figure 2*, Client and Investor form a covenant in Hobbes' sense by exchanging promises to perform, since the Client must trust the Investor to perform after the Client performs.[15] Hobbes would certainly maintain that if in a civil society a pair of individuals exchange promises to perform in an investment problem having the Simple Trust Game structure, the party in the role of the Client and the party in the role of the

[14] Kavka (1986) and Hampton (1986; 1997, chapter 2) give particularly illuminating discussion of various ways Hobbes foreshadows contemporary contractarians.

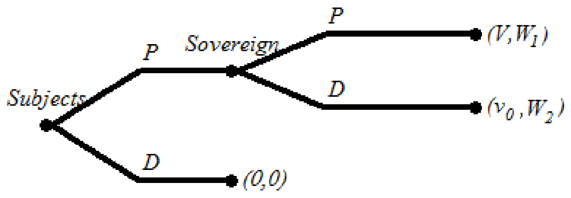
[15] If the Client does perform, then the Investor is in the position of Hobbes' Foole of *Leviathan* 15, who alleges that one has no good reason to honor one's end of a covenant in case the other party has performed already (*Leviathan* 15:5). I discuss the Foole's challenge at greater length in Vanderschraaf (2007; 2010).

Investor are each obliged to perform their ends of this covenant. Such consequences follow at once from Hobbes' third *Leviathan* law of nature, "without which, Covenants are in vain, and but Empty words; [. . .]" (*Leviathan* 15:1). Similarly, if the Simple Trust Game summarizes an exchange between an employee in the Client role and an employer in the Investor role, if they exchange promises to work in return for a later payment then Hobbes would maintain the obligations of this covenant bind each of them, and that in civil society each definitely has good reason to keep faith.

But Hobbes insists that there can be no covenant between a sovereign and the sovereign's subjects (*Leviathan* 18:4). According to Hobbes, the parties who are to be subjects contract with each other, but not with the party that is to be or is already their sovereign. As is well known, Hobbes' famous view contrasts sharply with that of later contractarians, including especially Locke, who makes it clear that he believes that a monarch or legislature that rules civil society is answerable to the people ruled (*Second Treatise* §240, §242–243). Hobbes' position might also seem surprising in light of covenants with a Simple Trust Game structure such as those discussed in the preceding paragraph, for in a certain sense a crucial step in establishing the relationship between subjects and sovereign is structurally much like that of the Simple Trust Game. Hobbes argues that in order to exit the State of Nature and enter into civil society, the members of a population must *authorize* some individual person or assembly of people as their sovereign. Hobbes gives no clear definition of 'authorization', but to authorize some agent implies that one surrenders a number of State of Nature rights exclusively to this agent. In particular, one is obliged to obey the authorized agent's commands to assist in punishing others, and may defend or punish others only with this agent's permission (*Leviathan* 28:2). The now subjects of the authorized sovereign must obey all the sovereign's commands with only very limited exceptions, and they may not try to replace the sovereign (*De Cive* 6; *Leviathan* 17, 18). Creating Hobbesian commonwealth "from scratch" requires those who would exit the State of Nature to solve two problems. First, they must select the individual or assembly of individuals that is to serve as their Sovereign. Once they have chosen some candidate individual or assembly, they must all then authorize the actions of this candidate as if these actions were their own, so that each expects all to generally obey this candidate's commands as sovereign, and consequently has good reasons to obey this candidate herself.[16]

[16] Hobbes uses the term 'authorize' only in *Leviathan* 17:13 and 18:1, when he discusses commonwealth by institution. In the corresponding texts of *Elements of Law* and *De Cive*, Hobbes speaks of all the people surrendering their rights to the sovereign (*Elements of Law* 19:7, 10; *De Cive* 5:7). Hobbes' introduction of the new term 'authorize' in *Leviathan* has sparked a very interesting discussion as to whether or not authorization is a substantive addition to Hobbes' political theory in *Leviathan*. See Gauthier (1969, chapter 4), Hampton (1986, chapter 5) and Kavka (1986, §10.1). Since in *Leviathan* 17:13 Hobbes says "*I authorize and give up my right of governing myself to this man, or to this assembly of men*, [. . .]", I am inclined towards the view that authorization is nothing importantly new in *Leviathan*. But I do find the terminology of authorization quite helpful for describing the establishment of Hobbesian civil society.

The candidate selection problem is quite challenging,[17] but here I will suppose that the parties trying to create their commonwealth have already solved this problem. The parties all know the identity of the sovereign-candidate. This candidate becomes sovereign once they authorize it, and afterwards is able to rule effectively and to provide the subjects with the benefits for which they created their sovereign. I use the *figure 4* game to summarize the final stage of establishing the Hobbesian sovereign.



P = authorize (for Subjects), perform (for Sovereign),

D = abstain (for Subjects), defect (for Sovereign)

$V > 0$, $V > v_0$, $W_1 > 0$, $W_2 > 0$

Figure 4: Irrevocable Sovereignty Game

In this game the Subjects move first and can either *authorize* (*P*) or *abstain* (*D*). If they abstain they will remain the State of Nature. If they authorize, the Sovereign can then either *perform* (*P*) by providing good government or *defect* (*D*) by ignoring the Subjects' needs and possibly even exploiting them. I call this the *Irrevocable Sovereignty Game* because the agents in the roles of Subjects and Sovereign engage in this game but once, reflecting Hobbes' official view that once the Sovereign is established the Subjects have no right to try to replace it. If a Sovereign performs, the Subjects receive a payoff of $V > 0$ as the value of good government and the Sovereign's payoff is $W_1 > 0$. If the Sovereign defects its payoff is then $W_2 > 0$ and the Subjects' payoff is $v_0 < V$. The v_0 payoff the Subjects receive if the Sovereign defects reflects the idea that even if the Sovereign devotes no resources specifically in order to benefit the Subjects, the Subjects might receive a "spillover" benefit resulting from the Sovereign's self-maintenance. For example, the Subjects may enjoy some protection from foreign

[17] For in-depth discussions of the candidate selection problem in Hobbes' theory, see Hampton (1986, §6.4–6), Kavka (1986, §5.1) and Vanderschraaf (forthcoming, chapter 6).

aggression even if the Sovereign defects, since the Sovereign may need to provide for some defense against such aggression in order to maintain itself. Note that I do not assume that $\nu_0 \geq 0$, that is, I allow for the possibility that if the Subjects authorize and the Sovereign then defects, then the Subjects could be in an even worse position than they are in the State of Nature. If $W_2 > W_1$ and $\nu_0 < 0$, then the *figure 4* game has the same preference structure as the Simple Trust Game and the subgame perfect equilibrium is (*D, D*). In this case, the Subjects would do better to stay in the State of Nature.

Both Subjects and Sovereign fare better if both perform, suggesting that the Subjects would want to form a covenant with the Sovereign before they move and promise to authorize in exchange for the Sovereign's promise to perform. But again, Hobbes is clear that no covenant between the subjects and their sovereign is possible. Hobbes concludes this in large part because he concludes that it would be impossible to resolve any claim of breach of such an alleged "covenant".[18] For if some third party could act as judge between sovereign and subjects in such a dispute, then this third party would be the "real" sovereign, and then the question of who could adjudicate between this "real" sovereign and the "real" subjects would arise all over again. Hobbes maintains that the sovereign itself is the final arbiter in all disputes between members of its commonwealth (*Leviathan* 18:11), and that it would be futile to claim that the sovereign ever breaks faith with its subjects.

> "If any one, or more of them [the subjects] pretend a breach of the Covenant made by the Soveraigne at his Institution, and others, or one other of his Subjects, or himselfe alone, pretend there was no such breach, there is in this case no Judge to decide the controversie: it returns to therefore to the Sword again; and every man recovereth the right of Protecting himself by his own strength, contrary to the designe they had in the institution. It is therefore vain to grant Soveraignty by way of precedent Covenant." (*Leviathan* 18:4).[19]

Similarly, Hobbes maintains it makes no more sense to claim the sovereign is bound by laws of the commonwealth than by any supposed covenants between itself and subjects. The following passage illustrates especially well why Jean Hampton calls this part of Hobbes' defense of absolute sovereignty Hobbes' *regress argument*[20]:

[18] This is Hobbes' most important argument against the claim that a sovereign could form a covenant with its subjects, but not his only argument. In *Leviathan*, Hobbes also argues that prior to its authorization, a sovereign could not form a covenant with the whole multitude in the State of Nature as the other party, since they are not yet united, and any covenants it might form with each of the individuals in the multitude would be void upon its authorization "because what act soever can be pretended by any one of them for breach thereof is the act both of himself, and all the rest" (18:4). I doubt the soundness of this argument myself, but this argument of Hobbes is not crucial for the analysis of this section.

[19] See also *De Cive* 6:18.

[20] Hampton gives splendid reconstructions and analyses of Hobbes' regress argument in Hampton (1986, chapter 4) and (1994).

> "For to be subject to Lawes, is to be subject to the Common-wealth, that is, to the Soveraign Representative, that is, to himself; which is not subjection, but freedome from the Lawes. Which errour, because it setteth the Lawes above the Soveraign, setteth also a Judge above him, and a Power to punish him, which is to make a new Soveraign; and again for the same reason a third, to punish the second; and so continually without end, to the Confusion and Dissolution of the Common-wealth." (*Leviathan* 29:9)

Hobbes' argument relies upon a common understanding of a contract as an agreement between parties that is enforceable by some third party. As hinted in *section 2*, parties typically form contracts knowing that the officers of a legal system can serve as this third party. The third party enforcer in the background enables the contracting parties to be confident that their counterparts will perform so that they are willing to exchange their promises in the first place. One can find formal descriptions of such contracts in law texts, but I think Schelling captures this common understanding of contracts especially well in a characteristically sharp-witted passage in *The Strategy of Conflict*:

> "Among the legal privileges of corporations, two that are mentioned in textbooks are the right to sue and the 'right' to be sued. Who wants to be sued! But the right to be sued is the power to make a promise: to borrow money, to enter a contract, to do business with someone who might be damaged. If suit does arise, the 'right' seems a liability in retrospect; beforehand it was a prerequisite to doing business." (1960, 43)

In exchanges such as those between a client and an investor or an employee and her employer, when government enforces contracts using its legal system, a contract could in effect modify the payoff structure of the parties' interaction with third party intervention, as illustrated in the augmented Simple Trust Game of *figure 3*. But in the final stage of establishing a Hobbesian sovereign, there is no third party that can enforce the terms of a covenant between subjects and the sovereign-candidate. Hobbes clearly and notoriously maintains that once established, the power of the sovereign over its subjects is without limits. The regress argument is a key part of Hobbes' defense of this claim that has offended so many of his readers.

Nevertheless, Hobbes also believes that fortunately, the subjects and the sovereign do not need any covenants between them in order to establish and maintain a mutually beneficial relationship. For Hobbes maintains that the interests of the sovereign go hand in hand with the interests of its subjects. The sovereign's power increases just as the subjects' level of well-being increases (*Leviathan* 19:4).[21] And if the subjects do generally suffer disadvantages because their regime "malfunctions", the sovereign suffers disadvantages as well.

[21] Here Hobbes in fact claims that in the case of monarchy, the interests of the sovereign are *identical* with those of its subjects.

> "All the advantages and disadvantages of the regime itself are the same for *ruler* and *subjects* alike and are shared by both of them. The disadvantages which occur to a particular citizen by his own misfortune, stupidity, negligence, ignorance, or extravagance, may be separate from the disadvantages of the Ruler, but they are not disadvantages of the Regime, because they can occur in any commonwealth. If the same things happen as a result of the way the commonwealth was originally set up, they will indeed be called disadvantages of the Regime, but they will be common to ruler and citizen alike, just as their advantages will be common also." (*De Cive* 10:12)

The *figure 4* game reflects Hobbes' reasoning if $W_1 > W_2$, in which case the subgame perfect equilibrium is now (*P, P*). If Hobbes is right about the alignment of the sovereign's and the subjects' interests, then at the last stage the parties in the State of Nature apparently have good reason to authorize the chosen candidate as sovereign without trying in vain to first form a contract with the sovereign-candidate.

This last claim brings me back to the first general point regarding the vesting of power in the governors. In 1789 Benjamin Franklin famously quipped: "Our new Constitution is now established, and has an appearance that promises permanency; but in this world nothing can be said to be certain, except death and taxes." (2002[1855], 321) In fact, those who live under the rule of a state can count on considerably more than taxes levied by this state. They will be subject to laws promulgated by the rulers of this state, to state incursions into their privacy, and occasional conscription into the state's service, possibly even at the risk of life and limb.[22] And at least some of them might suffer severe punishments at the hand of their government, including fines, imprisonment and perhaps even death, while at the same time each of them must abstain from punishing others who may have violated her rights without the government's express permission. Indeed, Max Weber's famous claim that a monopoly over the legitimate use of force over a given territory is an essential defining characteristic of a state is often taken as a starting point for an analysis of the state.[23] Certainly states claim this special monopoly, and to the extent this monopoly is realized it marks perhaps the most important power asymmetry between the governed and their governors. In a recent essay, Christopher Morris (2012) argues that philosophers tend to exaggerate the role coercive force plays in the maintenance of the

[22] Libertarians might claim that the minimal state Robert Nozick describes in *Anarchy, State and Utopia* (1974) would be an exception, since this state exists for the sole purpose of protecting the rights of its inhabitants. However, even if Nozick's minimalist state could be realized, I believe such a state might well require service such as jury and possibly even military service and intrude into privacy via a periodic census.

[23] Weber gives his best-known statement of the monopoly of force claim in his 1919 lecture "Politics as a Vocation" (2004[1919], 33). In *The Theory of Social and Economic Organization*, Weber proposes a much more complete definition of the modern state that includes the monopoly of force claim as but one of several of the state's important characteristics (1947[1925], 15).

state. Morris argues that coercive force need not be incorporated into the definition of the state, and that the state is far more dependent upon consensus and coordination regarding its constitution and laws than is generally recognized.[24] Indeed, Hume (1994, 16) and Hobbes himself (1990[1668], 16) recognize the importance of coordinated opinions in maintaining government. Still, it is a brute fact of political life that states do often use coercive force against their subjects, and even leaving aside coercive force, the state can regulate the conduct of its subjects considerably on the basis of its unique authority, including legal authority, over them.[25] Indeed, the state would appear to have virtually unlimited power over its subject people. Hobbes thinks this should surprise no one.

> "It appeareth plainly, to my understanding, both from Reason, and Scripture, that the Soveraign Power, whether placed in One Man, as in Monarchy, or in one Assembly of Men, as in Popular, and Aristocraticall Common-wealths, is as great, as possibly men can be imagined to make it." (*Leviathan* 20:18)

Hobbes clearly acknowledges that citizens might be oppressed in any commonwealth (*De Cive* 10:2). One of the reasons I use the extensive form *figure 4* game as a simple model of establishing a Hobbesian sovereign is that this game captures the idea that once the State of Nature parties have authorized their chosen sovereign-candidate, they are effectively at their sovereign's mercy. One can reinterpret defection on the Sovereign's part as not merely neglecting the welfare of the Subjects, but actively persecuting them, possibly by taxing them severely, forcing oppressive laws upon them, or imprisoning or even executing some of them arbitrarily. If the Subjects authorize, then the Sovereign has final and complete control of the game and can defect with impunity. The Subjects in the Irrevocable Sovereignty Game can of course forestall all this by abstaining, but then they would remain in the State of Nature. Hobbes acknowledges that people in civil societies frequently express dissatisfaction with their governments, at least in private, but insists that the worst calamities that might befall a people in civil society at the hands of their sovereign pale in comparison when compared with what they might expect in the State of Nature (*Leviathan* 18:20). Hobbes would argue that under a correct interpretation of the Irrevocable Sovereignty Game, even if they fear the Sovereign will defect, the Subjects still should authorize because $\nu_0 > 0$. Moreover, Hobbes would argue that the Sovereign would prefer to perform in this game, on account of the coincidence of the Subjects' and the Sovereign's interests. Hobbes specifically argues that it is to the sovereign's own advantage to provide its subjects with peace and a good defense and to tax them only to the degree necessary in order to provide good government, since failing to protect the subjects adequately or taxing the subjects to the extent that they have trouble providing for themselves will

[24] Hardin (1999) also emphasizes how a state's constitution serves as a coordinating device.

[25] Again I am indebted to Christopher Morris, who has helped me better appreciate this point both by his works and in conversation.

weaken the sovereign's power (*De Cive* 10:2). By authorizing, the Subjects vest the Sovereign with overwhelming power, but this is supposedly nothing for them to fear since they should expect the Sovereign to reward their authorization with the benefits of its performance.

An alternate way of summarizing Hobbes' overall position is to note that in general, no convention is possible in an interaction having an Irrevocable Sovereignty Game structure and this fact should not deter subjects from authorizing, since they should foresee that the sovereign-candidate will after installation follow its part of an optimal equilibrium. David Lewis (1969) presented the first contemporary theory of convention, which has since been generalized by various authors including Robert Sugden (1986) and myself (2001; forthcoming, chapter 2). According to these theories, a convention is characterized by an equilibrium of a game the agents involved follow given that they have common knowledge of: (*i*) the payoff structure of the game, (*ii*) their (Bayesian) rationality, and that (*iii*) they follow *this* equilibrium knowing they would have followed a different equilibrium had their beliefs about each other's intended actions been appropriately different.[26] This equilibrium analysis captures the idea that a convention is social arrangement that is arbitrary in a *discretionary sense*,[27] that is, agents involved follow some practice as their convention when they know they might have followed some alternative practice as their convention. In the *figure 4* game, unless $W_1 = W_2$ only one outcome is compatible with common knowledge of payoff structure and rationality.[28] Given $W_1 > W_2$ as Hobbes would suppose, the backwards induction solution path is the unique subgame perfect equilibrium (*P*, *P*) outcome where Subjects authorize and Sovereign performs. There is no alternative outcome compatible with common knowledge of payoff structure and rationality, so there is no convention available to the Subjects and the Sovereign.[29]

[26] A proposition *A* is common knowledge among a group of agents if each agent knows that all know *A* and knows that all can infer the consequences of this mutual knowledge. Lewis (1969, chapter 2) presented one of the first analyses of common knowledge.

[27] This is my terminology (Vanderschraaf forthcoming, chapter 2). Lewis expressed the idea that a convention is always arbitrary in the discretionary sense (1969, 70). Discretionary sense arbitrariness is not to be confused with *indifference sense arbitrariness*, a special and atypical sense of arbitrariness where some set of conventions are all equally good with respect to each other. Indifference sense arbitrariness, when it obtains for a given set of conventions, is characterized by assigning the equilibria of these conventions identical payoffs.

[28] If $W_1 = W_2$ then if the Subjects follow *P* then the Sovereign can best respond by following a mixed strategy over *P* and *D*. Otherwise backwards induction selects a unique subgame perfect equilibrium outcome.

[29] Hampton (1986, §8.3) and (2007[1990]) presents a one-shot extensive form game model of her own of the relationship between ruler and subjects where the Sovereign moves first, either performing or defecting, and then the Subjects respond either by keeping the Sovereign in power or by deposing the Sovereign. In Hampton's model, the unique subgame perfect equilibrium is the outcome where the Sovereign performs and the Subjects keep the Sovereign in power. I take issue with Hampton's model for two main reasons. First, as Hampton clearly acknowledges, according to her model the Sovereign appears to be in a relatively powerless position with respect to its Subjects, and this strikes me as counterintuitive to say the least. But the second and more serious reason I reject Hampton's model is that even though Hampton explicitly adopts Lewis' account of conven-

4. Revocable Sovereignty and a Governing Convention

Life might not work out so pleasantly for both sovereign and subjects as Hobbes supposes. Hume, whose own views regarding human weaknesses are close to those of Hobbes, claims such weaknesses may easily pervert government and turn the governors into tyrants (*Treatise* 3.2.10:4). Part of one of Locke's acerbic responses to defenders of Hobbesian-style absolutism is worth repeating:

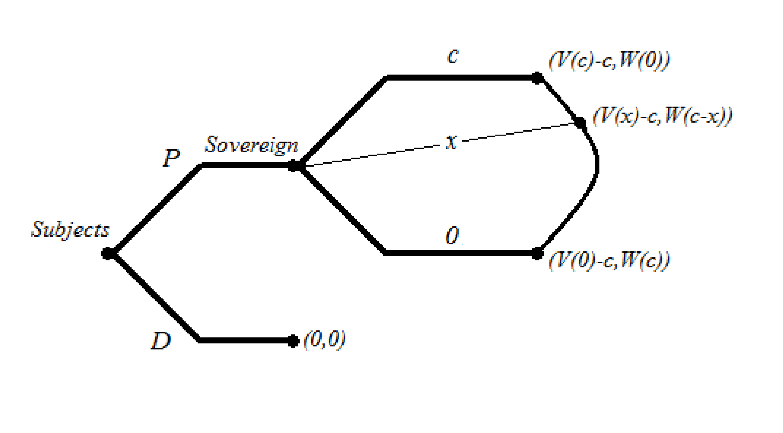
> "As if when men, quitting the State of Nature, entered into Society, they agreed that all of them but one should be under the restraint of Laws; but that he should still retain all the Liberty of the State of Nature, increased with Power, and made licentious by Impunity. This is to think that Men are so foolish that they take care to avoid what Mischiefs may be done them by *Pole-cats* or *Foxes*, but are content, nay, think it Safety, to be devoured by *Lions*." (*Second Treatise* §93)[30]

On the other hand, the subjects might not be so helpless in civil society as the Irrevocable Sovereignty Game model suggests. *Figure 5* depicts a *Repeatable Sovereignty Game*.

As the name suggests, the agents in the roles of Sovereign and Subject engage in this game repeatedly over time periods. One can think of the agent in the role of the Subjects as roughly the same multitude of people in each period that can establish or maintain civil society, while the agent in the role of the Sovereign in a given period might at a later period be replaced by a new agent that assumes this role according to the moves the two agents in the game might follow. At the start of each period, the Subjects can follow P by *obeying* the Sovereign's commands at a cost $c > 0$. If the Subjects obey, then they keep their current Sovereign in power. The Subjects can also neutralize the activity of the incumbent Sovereign by following D and *disobeying*, saving themselves the cost of obedience. Disobedience puts the Subjects and the Sovereign both back in the State of Nature, much like abstaining keeps both Sovereign and Subjects in the State of Nature in the *figure 4* game. If the Subjects obey, then the Sovereign responds by following a strategy $x \in [0, c]$, which for $x > 0$ is *performing to the degree x*. The Sovereign *defects* by following $x = 0$. At the outcome (P, x), the Subjects' payoff is $V(x) - c$ and the Sovereign's payoff is $W(c - x)$, where $V(\bullet)$ and $W(\bullet)$ are both strictly increasing functions. I assume that $W(0) > 0$ and $V(c) > c > V(0)$. In a single play of this game, if Subjects obey then the Sovereign's unique best response is to defect, so the unique subgame perfect equilibrium is

tion, the equilibrium of her game does not characterize a convention. For as in the *figure 4* game, in Hampton's game there are no alternative outcomes compatible with common knowledge of the payoff structure and rationality. So Hampton was mistaken in thinking her game summarizes a governing convention.

[30] See also *Second Treatise* §137. Interestingly, while some read Locke's §93 as a direct response to Hobbes, Peter Laslett argues that it is likely that here Locke's direct target is Filmer. See Laslett's note 32 in the Laslett edited edition of Locke's *Two Treatises of Government*.



$P =$ obey (for Subjects), $D =$ disobey (for Subjects)

$x =$ perform to degree $x \in [0,1]$ (for Sovereign)

$0 =$ defect (for Sovereign)

$V(c) > c, c > V(0), c > 0,\ W(c) > W(0) > 0$

Figure 5: Repeatable Sovereignty Game

(D, 0). In the Repeatable Sovereignty Game, the Sovereign is always tempted to maximize its immediate payoff by defecting, and the Subjects are then left with a lower payoff than their State of Nature payoff. This reflects the idea that a sovereign could gain its own best payoff by exploiting its subjects, thereby leaving them even worse off than they would be in the State of Nature. But if the Sovereign gives in to temptation and defects in a given period, the Subjects might retaliate by disobeying at a future period. Disobedience has the effect of a rebellion, so that if the Subjects disobey in a given period then the agent in the role of Sovereign is replaced by a fresh agent in the following period. A crucial feature of this repeated game is that the ability and willingness of the agents that can assume the role of Sovereign to perform varies, so that from the Subjects' point of view some candidates for the post of Sovereign are better than others. Suppose that at a given period t, the Subjects did not obey at the immediately preceding period $t - 1$. Now, before they choose either to obey or disobey at t, they receive a signal $z_t \in [0, c]$ indicating the degree of performance the Sovereign "offers" if they now obey. From the Subjects' perspective z_t is a random variable drawn from a common distribution. When the Subjects contemplate obeying for the first time a candidate for Sovereign, z_t is their indicator, possibly following an investigation, of how well this candidate is capable of serving them.[31] If the Subjects obey at this period then they allow the Sovereign

[31] I do not assume that the Sovereign's "offer" z_t before first obedience is a promise on the part of the agent in the role of Sovereign to perform at a certain level. Rather, z_t reflects, perhaps imperfectly, a candidate's ability to perform. In fact, I suppose that a candidate's promises give the Subjects

to act, and the Sovereign will then perform to the degree $x_t \in [0, z_t]$. This reflects the idea that in the state of nature, a given sovereign-candidate might indicate that it will if authorized perform to the degree z_t, and then should this candidate receive the power stemming from the subjects' obedience, then it performs in fact at a level x_t that might be as good for the subjects as the "offered" level z_t but might also be lower than z_t. If the Subjects obey, they infer the realized value of the Sovereign's performance x_t from their payoff $V(x_t) - c$. The Subjects can adopt the following contingency strategy recursively defined for $t \geq 0$:

$$f_1(x^*,\ 0) = \begin{cases} P \text{ if } z_0 \geq x^* \\ D \text{ otherwise} \end{cases}$$

and for $t > 0$

$$f_1(x^*,\ t) = \begin{cases} P \text{ if } f_1(x^*,\ t-1) = D \text{ and } z_t \geq x^* \\ P \text{ if } f_1(x^*,\ t-1) = P \text{ and } x_{t-1} \geq x^* \\ D \text{ otherwise} \end{cases}$$

Informally, the Subjects obey the Sovereign either if: (*i*) at the current period they expect the Sovereign, after they did not obey in the previous period, to perform to at least the degree x^*, or (*ii*) they obeyed in the previous period and the Sovereign in that period indeed performed to at least the degree x^*. If (*ii*) is the case, then the Subjects' payoffs at each earlier consecutive period where they have obeyed have been at least as good as an *acceptability threshold* $V(x^*) - c > 0$. But if the Subjects obey at a given period and the Sovereign fails to provide them with at least this acceptability threshold payoff, then at the next period the Subjects defect and continue to defect until they expect that the Sovereign has adopted a new strategy $x \geq x^*$, so that the Subjects expect once more to receive benefits they find acceptable. So it would seem that the Subjects can have some leverage over their Sovereign given that they are in effect able to punish the Sovereign for poor performance. Now suppose the Sovereign's strategy is

$$f_2(t) = \begin{cases} z_t \text{ if } t = 0 \text{ or } f_1(x^*, t-1) = D \\ f_2(t-1) \text{ otherwise} \end{cases}$$

This strategy reflects the idea that if subjects authorize a given sovereign-candidate by obeying this candidate for the first time at some period t_0, then this from then on authorized sovereign adopts a policy of performing at the level it has "offered", that is, $x_t = z_{t_0}$ over each period $t \geq t_0$ the subjects continue to obey. During a "reign" where Subjects follow P over periods $t_0, \ldots,$ t, the Sovereign receives a payoff $W(c - x_t) = W\left(c - z_{t_0}\right)$ while the Subjects receive $V(x_t) - c = V\left(z_{t_0}\right) - c$. If the Subjects do not obey the current candidate for Sovereign at period t or depose the incumbent Sovereign at period t, then at period $t + 1$ a new candidate assumes the role of Sovereign in the *figure 5* game, and this process repeats at each period until the Subjects obey once more and install a new incumbent Sovereign. Now each side intuitively has some lever-

little useful information, since any candidate might well promise to perform at the best possible level from the Subjects' perspective even if this candidate cannot or will not actually perform at this level.

age. If the Sovereign knows the true value of x^*, then if $z_{t_0} \geq x^*$ the Sovereign's best response is to always follow $x_t = x^*$ at each period $t \geq t_0$ the Subjects obey. However, I allow for the possibility that from the Sovereign's perspective, x^* is a random variable, so that a Sovereign might "offer and deliver" to its Subjects more than just the bare minimum needed to maintain its power. If a candidate is unwilling or unable for lack of ability to in effect submit a sufficiently high "offer" z_t to its Subjects, then the Subjects might pass this candidate over in hopes of finding a better candidate in the future. Likewise, if after authorization by the Subjects' obedience an incumbent Sovereign fails to perform up to the level of its "accepted offer", the Subjects may depose the incumbent and obey at a later time a new candidate as Sovereign they expect will serve them better. So the Sovereign has incentive to try to perform over time at a level the Subjects find satisfactory. From their side, the Subjects can depose an incumbent Sovereign if the Sovereign fails to meet their acceptability threshold payoff $V(x^*) - c$, but if they set their standard x^* too high, then they run the risk of having trouble finding a replacement Sovereign that meets their standard if they depose the incumbent.

Interestingly, if the functions $V(\bullet)$ and $W(\bullet)$ are assumed to have derivatives, if the random variables (z_t) have a common probability density $g(\bullet)$, and if the Subjects and the Sovereign both discount their payoffs at future payoffs by an appropriate factor $\delta \in (0, 1)$, then a value x^* exists such that (f_1, f_2) is an equilibrium of the indefinitely repeated *figure 5* game where $V(x^*) - c > 0$ and $W(x^*) > 0$.[32] So a stable and mutually beneficial relationship can obtain between Subjects and Sovereign where the Subjects obey and the Sovereign provides the Subjects benefits they find more satisfactory than the "benefits" of remaining in the State of Nature. The state where at each period, the Subjects disobey and the Sovereign would defect if installed by the Subjects' obedience is also an equilibrium, the suboptimal equilibrium where both Sovereign and Subjects are in the State of Nature. The Subjects and the Sovereign have available to them alternative equilibria they might settle into, according to their conjectures over each other's chosen strategies. So the (f_1, f_2) equilibrium of the indefinitely repeated Sovereignty game characterizes a convention corresponding to a governing convention between a ruled people and their ruler or rulers.

The Repeatable Sovereignty Game model illustrates an important rebuttal to Hobbes: A *self-enforcing* covenant between the subjects and their sovereign is possible in principle. Hobbes maintains that a covenant between subjects and sovereign is conceptually impossible, because he relies upon his belief that any such purported covenant would have to be enforceable by a third party in a manner analogous to a legally binding contract. But not all contracts require third party enforcement, and for that matter, not all contracts require explicit promises. Hobbes may have appreciated this, since he indicates in his response

[32] This result follows from the theory of stochastic dynamic programming. The proof is a variation of the proof of the existence of a *reservation wage* equilibrium in a model of search unemployment that Nancy L. Stokey and Robert E. Lucas give in (1989, §10.7). Stokey and Lucas base their analysis on the model presented in McCall (1970).

to the Foole in *Leviathan* 15 that some covenants made in the State of Nature might be binding without enforcement by a third party (*Leviathan* 15:5), and also allows that under certain circumstances agents may give their tacit consent to provisions of a covenant (*Leviathan* 18:4, 26:7). However, Hobbes evidently believed that a purported covenant between a sovereign and its subjects would have to be of the sort analogous to a legally binding contract. I suspect Hobbes believed this because at bottom he believed that the power asymmetry between a sovereign and its subjects is ordinarily so great that only an even more powerful third party could hope to resolve a dispute between the two. But the subjects might not be so helpless with respect to their sovereign as Hobbes' descriptions of the sovereign's powers suggest. I have used the Repeatable Sovereignty Game model to illustrate the idea that the governed can overthrow their governors at any time by withdrawing their obedience. Hence it can be in a sovereign's long term interests to serve its subjects adequately not because their interests coincide so closely in the ways Hobbes supposes, but because this sovereign needs to preserve the subjects' willingness to obey its commands.

I think Hume recognized that the relationship between governors and the governed is ordinarily a conventional relationship, and that this governing convention is more stable the more nearly the interests of the governed balance with those of their governors. Put another way, for Hume, government is maintained via a contract by convention. This claim may seem surprising given Hume's repeated denials that our obligations to our governments stem from our explicit or tacit consent (*Treatise* 3.2.8–9; 1994, 186–201). For Hume, mutual recognition that their government is a human invention that serves the public interest is the real root of political obligation (*Treatise* 3.2.10:4). He also points out that the governments of recorded history have their origins in usurpation or conquest (*Treatise* 3.2.10:4; 1994, 189–190). But Hume is only an anti-state contractarian according to the more limited understanding of a contract as a system of promises enforceable by some third party. Hume suggests the governed themselves can keep the conduct of their governors at least somewhat in check by the threat of revolt. Indeed, he observes that even the most tyrannical regimes are incapable of depriving the people under their rule the right of resistance (*Treatise* 3.2.10:16). Hume argues that people turn to and respect governors in the first place because they perceive that it is in the direct interests of the governors that certain projects such as the generation of public goods are executed for the benefit of all (*Treatise* 3.2.7:6,8). Moreover, Hume argues that governments can be examined from the perspective of an "impartial examiner" in order to ascertain how well it serves the public interests (1994, 27).[33] So Hume's discussion of governments contains a normative element that complements its explanatory elements. Hume admits that it may be difficult if not impossible to give precise particular conditions either for an optimally designed government or for when armed insurrection is justified (*Treatise* 3.2.10:16; 1994, 27). But he maintains that all would accept the general principle that political obligation ceases

[33] In *Treatise* 3.3.1:14 Hume also speaks of evaluating moral qualities from the perspective of a "judicious spectator".

should the government cease to provide mutual advantage and security (*Treatise* 3.2.10:4, 16). Game theory facilitates a more precise contemporary reformulation of these ideas. The Repeatable Sovereignty Game model summarizes how a people might depose their repressive regime if their plight becomes "bad enough". Hume's insights regarding a "judicious spectator" or "impartial examiner" foreshadow the impartial spectator central to Adam Smith's moral theory and the contemporary moral contractarian theories in the impartial spectator tradition such as John Harsanyi's and John Rawls' veil of ignorance theories.[34] These insights also illuminate how one can understand the relationship between rulers and ruled as a contract by convention. The ruled will continue to obey so long as their ruler provides sufficient benefits, and what counts as sufficient benefits can be ascertained from the perspective of an impartial spectator.

Interestingly, the possibility of a governing convention also reflects another strand of Hobbes' thinking, despite Hobbes' claims that subjects may not try to depose their sovereign. Hobbes explicitly maintains that the subjects' obligation to obey their sovereign ends if this sovereign ceases to be able to protect them (*Leviathan* 21:21). If the Sovereign in the *figure 5* game defects at one or more given periods in time, the Subjects might take this as a sign that the Sovereign has become ineffective and then withdraw their obedience at a later time, ending this Sovereign's rule. Hobbes gives the subjects one "escape clause" in his description of their obligations to the sovereign, namely, the event that the sovereign becomes incapable of protecting them. Yet from the subjects' perspective perhaps it makes no relevant difference if the sovereign is still able to protect them and simply fails to do so. The Repeatable Sovereignty Game model illustrates this idea, as the Subjects may depose their incumbent Sovereign for actual poor performance even if they might believe this Sovereign is still capable of performing adequately. As Hampton puts it in several of her discussions of Hobbes' political theory, Hobbesian subjects are in fact capable of "hiring and firing" their sovereign (1986, 235–236; 1997, 52).

5. Discussion

Hobbes, and Hume after him, rejected the idea that a contract of the sort that obligates parties legally or is subject to third party arbitration could underwrite the relationship between a sovereign and its subjects. But Hume, and perhaps Hobbes before him, recognized in an informal manner that rulers and

[34] See Rawls (1971) and Harsanyi (1977, chapter 4), for summaries of their moral contractarian theories. Rawls' and Harsanyi's theories are foreshadowed by Vickrey's (1948) use of a veil of uncertainty regarding one's own position in a given community. Harsanyi (1977, 48–49) acknowledges that rational choice contract theories such as his own are part of Smith's impartial spectator tradition. Rawls (1971, §30) draws a sharp distinction between his justice as fairness and impartial spectator theories, claiming the latter depend upon sympathy and that his theory avoids this dependence. Obviously, I prefer to regard Rawls' theory as an impartial spectator theory, and I do so because of Rawls' use of a veil of ignorance.

the ruled can establish and maintain a contract by convention or governing convention. Here I have used game theory to show that a governing convention can indeed regulate the interactions between the ruled and their rulers. The Repeatable Sovereignty Game summarizes how a balance of power can exist between Sovereign and Subjects that ensures that the Sovereign serves its Subjects adequately in in return for the Subjects' obedience. In this setting, there can be a covenant between Subjects and their Sovereign that they enforce without appealing to an outside arbitrating power. The equilibrium of the indefinitely repeated *figure 5* game where over time the Subjects obey and the Sovereign performs is a simple example where a governed people and their governor or governors can police an agreement that benefits each side themselves. Moreover, while they could form this self-enforcing agreement explicitly, this is not a necessary feature of such an agreement. The agents in the positions of Subjects and Sovereign could before the start of play first exchange promises to follow their ends of an equilibrium based upon the appropriate threshold x^*. In this manner they might settle into the equilibrium where at each period the Subjects follow P and obey and the Sovereign follows x^*, performing at the necessary threshold level, quite rapidly.[35] This would be analogous to the role Russell Hardin (1999, chapter 3) argues a written constitution plays in focusing the expectations of the governed very rapidly on a set of governing policies, which in a well-designed constitution characterize a complex but stable social equilibrium that includes a governing convention. But as Hume would remind us, the ruled and their rulers seldom if ever draw up such an explicit contract that regulates their relationship. And the ruled and their rulers might not need any such explicit agreement in order to establish and maintain an appropriate balance of power. For the agents involved may be able to learn to follow the equilibrium of a governing convention as Subjects depose an initial unsatisfactory Sovereign and search for various sovereign-candidates until they find a candidate that performs at the equilibrium level. In short, as is the case with conventions in general, a governing convention can emerge via some process other than formal agreement.[36]

Almost needless to say, the Repeatable Sovereignty Game model of *section 4* oversimplifies in several ways the relationships one would expect to find between the rulers and the ruled in actual civil societies. In the remainder of this section I will propose some avenues for future research. The Repeatable Sovereignty Game model presupposes that the Subjects always interact either with an incumbent sovereign or a single sovereign-candidate, and that they can estimate

[35] This might involve a modification of the model where after a pre-play round of negotiation the Subjects can pick an initial sovereign-candidate that is prepared to perform initially at the required level x^* and can renegotiate with further sovereign-candidates that are prepared to meet this standard in the event they depose this initial sovereign. In the model described above, the Subjects receive their sovereign-candidates at random and then wait until they identify a candidate that will meet their standard before initially obeying.

[36] Hume (*Treatise* 3.2.2:10) and Lewis (1969) after him argue that conventions can emerge via processes other than explicit agreement. Sugden (2004[1986]), Young (1998) and Vanderschraaf (2001) give more recent analyses of the evolution of conventions via dynamic learning processes that do not assume the agents explicitly communicate with one another.

with some accuracy how well a sovereign-candidate will perform prior to obeying this candidate for the first time. So if they depose their current Sovereign, the Subjects have some ability to "screen" one at a time each of a sequence of subsequent candidates for their next Sovereign, so that they expect they have a good chance selecting a candidate that will meet their standards before they authorize this candidate by first obedience. Put another way, this model oversimplifies by presupposing that at each period, exactly one incumbent sovereign or sovereign-candidate is salient for the Subjects, who must decide either to obey this incumbent or candidate or to disobey and then either initiate or continue a search for a new sovereign. In fact, the process of identifying a suitable replacement for a deposed sovereign is likely to be more complex, involving multiple competing candidates that might not all be equally salient from the Subjects' perspective. One might model this process using some preliminary set of n-agent impure coordination games whose equilibria correspond to the alternative sovereign-candidates.[37] Another way the Repeatable Sovereignty Game model oversimplifies is by assuming that the reciprocal costs and benefits in equilibrium are constant. A more realistic model might incorporate some variability in these costs and benefits, reflecting the realities that a government might need to increase the costs it imposes on the governed in times of war or other crises and that a governed population might grow discontented if its members perceive a stagnation or decline in their benefits relative to their costs of obedience.

Perhaps the most important oversimplification of this model, one shared by the Irrevocable Sovereignty Game model, is that it treats the Subjects as a single unified coalition that can act at will either to support or to depose the Sovereign. If subjects cannot easily disobey *together*, then the sort of tyranny Hume feared might result. In fact, the dynamics of revolution are exceedingly complicated, and revolutions are quite hard to predict beforehand in large part because those under the rule of any regime might have trouble organizing simultaneous disobedience.[38] Moreover, one might complain that if the Subjects could coordinate their obedience and disobedience so smoothly as they evidently can in the *figure 4 and 5* games, then they would not need a Sovereign in the first place. One response to this complaint is that a widespread withdrawal of obedience might be easier for a populace to achieve than the widespread coordination of activities an effective government can generate, simply because it takes little specific instruction to know how to stop obeying some authority. Whether this last claim is true or not, the facts remain that revolutions occasionally occur even in societies of millions and that few if any of the parties involved can foretell with any precision when such revolutions will occur. In particular, the rulers of an

[37] Some relevant discussions of this candidate selection problem are cited in note 17.

[38] The so-called *paradox of revolution* has sparked an interesting literature. Tullock (1971) and Kavka (1982) are two of the foundational essays in this literature. Kavka also discusses the paradox of revolution in Kavka (1986, 266–279). In Vanderschraaf (2008), I argue that the process of revolution is best modeled by integrating the game-theoretic approach with the threshold analysis approach to analyzing social change first proposed by Thomas Schelling (1978) and Mark Granovetter (1978).

incumbent regime may have some rough sense of the level of simmering discontent among the ruled, but they might well not know just how far the ruled are willing to be further "pushed" before they will rebel. Another way to render the Repeatable Sovereignty Game model more realistic would be to have the Subjects follow a mixed strategy at each period against an incumbent Sovereign where, if the Sovereign defects, the Subjects disobey with a certain probability that increases with the difference $V(x^{*}) - V(x)$ between the Subjects' threshold payoff $V(x^{*})$ and their realized payoff $V(x) < V(x^{*})$. This would reflect the intuition that while a revolt might be a somewhat unpredictable event, the worse an incumbent regime treats those under its rule, the greater the pressure the ruled experience and hence the more likely they "erupt" into widespread mutiny against this regime.

One can view the discussion in *section 4* of the Repeatable Sovereignty Game as presenting an informal existence theorem. Given appropriate constraints on their payoff functions and discount factors, the Subjects and the Sovereign in this indefinitely repeated game have available to them an equilibrium that characterizes a contract by convention. But under what conditions are governing conventions likely to emerge and persist in the real world? As the above discussion in this section suggests, I think this is an open question. I have discussed here in admittedly rough outline some of the conditions that make a governing convention possible and some of the variations of these conditions that may affect the chances that a governing convention rather than a tyranny or anarchy prevails for a given society. Much more work needs to be done if we are to better understand the nature of governing conventions. Still, I hope I have adequately defended this core idea: State social contractarianism makes sense when one interprets the underlying "contract" as a self-enforcing and not necessarily explicit body of rules, that is, as a convention.

References

Bartlett, J. (2002[1855]), *Bartlett's Familiar Quotations*, 17th ed., Justin Kaplan, general ed., Boston: Little, Brown and Company.

Binmore, K. (1994), *Game Theory and the Social Contract Volume I: Playing Fair*, Cambridge/MA: MIT Press.

— (1998), *Game Theory and the Social Contract Volume II: Just Playing*, Cambridge/MA: MIT Press.

— (2005), *Natural Justice*, Oxford–New York: Oxford University Press.

Camerer, C. (2003), *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton: Princeton University Press.

Davis, M. (2009), "Fourteen Kinds of Social Contract", *Journal of Applied Ethics and Philosophy* 1, 8–19.

Gaus, G. (2011), *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, Cambridge: Cambridge University Press.

Gauthier, D. (1969), *The Logic of Leviathan: The Moral and Political Theory of Thomas Hobbes*, Oxford: Clarendon Press.

— (1979), "David Hume: Contractarian", *Philosophical Review* 88, 3–38.

— (1986), *Morals By Agreement*, Oxford: Clarendon Press.

Granovetter, M. (1978), "Threshold Models of Collective Behavior", *The American Journal of Sociology* 83, 1420–1443.

Hampton, J. (1986), *Hobbes and the Social Contract Tradition*, Cambridge: Cambridge University Press.

— (1990), "The Contractarian Explanation of the State", *Midwest Studies in Philosophy* 15, 186–216.

— (1991), "Two Faces of Contractarian Thought", in *Contractarianism and Rational Choice: Essays on David Gauthier's Morals by Agreement*, New York–Cambridge: Cambridge University Press, 31–55.

— (1994), "Democracy and the Rule of Law", in *The Rule of Law*, ed. Ian Shapiro, New York–London: New York University Press, 13–44.

— (1997), *Political Philosophy*, Boulder: Westview Press.

Hardin, R. (1982), *Collective Action*, Baltimore–London: Johns Hopkins University Press.

— (1999), *Liberalism, Constitutionalism and Democracy*, Oxford: Oxford University Press.

Harsanyi, J. (1977), *Rational Behavior and Bargaining Equilibrium in Games and Social Situations*, Cambridge: Cambridge University Press.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr and H. Gintis (2004) (eds.), *Foundations of Human Sociality: Economic Experiments from Fifteen Small-Scale Societies*, Oxford: Oxford University Press.

Hobbes, T. (1994[1640]), *The Elements of Law: Human Nature and De Corpore Politico*, ed. L. C. A. Gaskin, Oxford: Oxford University Press.

— (1998[1642]), *De Cive (On the Citizen)*, trans. and ed. Richard Tuck and Michael Silverthorne, Cambridge: Cambridge University Press.

— (1991[1651]) , *Leviathan*, ed. Richard Tuck, Cambridge: Cambridge University Press.

— (1990[1668]), *Behemoth, or the Long Parliament*, Chicago: University of Chicago Press.

Hume, D. (2000[1740]), *A Treatise of Human Nature*, ed. David Fate Norton and Mary J. Norton, Oxford: Oxford University Press.

— (1994), *Political Essays*, ed. Knud Haakonssen, Cambridge: Cambridge University Press.

Kavka, G. (1982), "Two Solutions to the Paradox of Revolution", *Midwest Studies in Philosophy* 7, 455–472.

— (1986), *Hobbesian Moral and Political Theory*, Princeton: Princeton University Press.

Lewis, D. (1969), *Convention: A Philosophical Study*, Cambridge/MA: Harvard University Press.

Locke, J. (1988[1690]), "The Second Treatise of Government", in: *Two Treatises of Government*, ed. Peter Laslett, Cambridge: Cambridge University Press, 265–428.

McCall, J. (1970), “Economics of Information and Job Search”, *Quarterly Journal of Economics* 84, 113–126.

Morris, C. (2012), “State Coercion and Force”, *Social Philosophy & Policy* 29, 28–49.

Nozick, R. (1974), *Anarchy, State, and Utopia*, New York: Basic Books.

Plato (1997), *Complete Works*, ed. John M. Cooper, Indianapolis: Hackett Publishing Company.

Rawls, J. (1971), *A Theory of Justice*, Cambridge/MA: Harvard University Press.

Roth, A. (1995), “Bargaining Experiments”, in: Kagel, J. H. and A. Roth, *Handbook of Experimental Economics*, Princeton: Princeton University Press, 298–302.

Rousseau, J. (1997[1762]), “The Social Contract”, in: *The Social Contract and Other Later Political Writings*, trans. and ed. Victor Gourevitch, Cambridge: Cambridge University Press, 39–152.

Santas, G. (2010), *Understanding Plato’s Republic*, Chichester: Wiley Blackwell.

Schelling, T. (1960), *The Strategy of Conflict*, Cambridge/MA: Harvard University Press.

— (1978), *Micromotives and Macrobehavior*, New York–London: W. W. Norton and Company.

Skyrms, B. (1996), *Evolution of the Social Contract*, Cambridge: Cambridge University Press.

Stokey, N. and R. Lucas (1989), *Recursive Methods in Economic Dynamics*, Cambridge/MA–London: Harvard University Press.

Sugden, R. (2004[1986]), *The Economics of Rights, Co-operation and Welfare*, 2nd ed. Houndsmills–Basingstoke–Hampshirer–New York: Palgrave MacMillan.

Tullock, G. (1971), “The Paradox of Revolution”, *Public Choice* 11, 89–99.

Vanderschraaf, P. (2001), *Learning and Coordination: Inductive Deliberation, Equilibrium and Convention*, New York–London: Routledge.

— (2007), “Covenants and Reputations”, *Synthese* 157, 155–183.

— (2008), “Game Theory Meets Threshold Analysis: Reappraising the Paradoxes of Anarchy and Revolution”, *British Journal for the Philosophy of Science* 59, 579–617.

— (2010), “The Invisible Foole”, *Philosophical Studies* 147, 37–58.

— (forthcoming), *Strategic Justice*, New York–Oxford: Oxford University Press.

Vickrey, W. (1948) “Measuring Marginal Utility by Reactions to Risk”, *Econometrica*, 319–333.

Weber, M. (2004[1917; 1919]), *The Vocation Lectures*, trans. Rodney Livingstone, ed. David Owen and Tracy B. Strong, Indianapolis: Hackett Publishing Company.

— (1947[1925]), *The Theory of Social and Economic Organization*, trans. A. M. Henderson and T. Parsons, ed. T. Parsons, New York: The Free Press.

Young, H. P. (1998), *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*, Princeton: Princeton University Press.