Why Hobbes' State of Nature is Best Modeled by an Assurance Game

MICHAEL MOEHLER

In this article, I argue that if one closely follows Hobbes' line of reasoning in *Leviathan*, in particular his distinction between the second and the third law of nature, and the logic of his contractarian theory, then Hobbes’ state of nature is best translated into the language of game theory by an assurance game, and not by a one-shot or iterated prisoner’s dilemma game, nor by an assurance dilemma game. Further, I support Hobbes’ conclusion that the sovereign must always punish the Foole, and even exclude her from the cooperative framework or take her life, if she defects once society is established, which is best expressed in the language of game theory by a grim strategy. That is, compared to existing game-theoretic interpretations of Hobbes, I argue that the sovereign plays a grim strategy with the citizens once society is established, and not the individuals with one another in the state of nature.

I. INTRODUCTION

Modern social philosophy has been strongly influenced by the prisoner’s dilemma (PD) game. The PD game is a mixed-motive game. Mixed-motive games are non-zero sum games and, as such, allow for mutually beneficial cooperation among rational individuals, although there is tension between the individuals’ cooperative and non-cooperative motives. The dilemma from the perspective of social philosophy is that it is never rational to cooperate in a (one-shot) PD game. Instead, defection is the strictly dominant strategy.

As such, if a situation of social interaction among rational individuals, who behave as if they were to maximize their expected utility, has the form of a one-shot PD game, cooperation does not take place, assuming common knowledge of rationality and complete information. Rational individuals alone are not able to realize the possible gains of cooperation in a one-shot PD game. Instead, they end up with a suboptimal outcome, both individually and collectively. To realize the (collectively) optimal outcome, an external authority, such as the state, is needed to transform the one-shot PD game into another game that makes mutual cooperation rational.

It is often argued that Hobbes’ state of nature is best modeled by a one-shot or iterated PD game, or more recently, by an assurance dilemma game. In this article, I reject these three game-theoretic interpretations of Hobbes’ state of nature.

---


Published in *Utilitas* 21 (2009): 297-326. This PDF file of the article is intended for personal academic use. Please cite only the slightly modified published version.
nature. Instead, I argue that if one closely follows Hobbes’ line of reasoning in *Leviathan*, in particular his distinction between the second and the third law of nature, and the logic of his contractarian theory, then Hobbes’ state of nature is best translated into the language of game theory by an assurance game. This finding is not entirely new, and support for it can be found elsewhere, at least in part. However, the reasons offered here to justify this conclusion and the explication of its game-theoretic implication for the post-natural state, are novel.

In a nutshell, I argue first that neither the PD game, in its one-shot or repeated form, nor the assurance dilemma game can model the problem of collective action that must be solved in Hobbes’ state of nature in order for society to be established, which is the problem of assurance. Second, these three games cannot model the situation of prudent individuals, who are the main type of actor in Hobbes’ state of nature and who identify the real game in the state of nature. The assurance game, by contrast, can model both aspects.

However, the one-shot PD game and the assurance dilemma game do not have to be entirely dismissed. If we assume an extended state of nature that includes, apart from prudent individuals, imprudent agents, or *Fooles* as Hobbes calls them, then these two games can model particular types of interaction in Hobbes’ state of nature. The one-shot PD game can model the interaction of Fooles, who misconstrue the game in the state of nature due to their short-sightedness. The assurance dilemma game can model the interaction of Fooles with prudent individuals.

Following these considerations, the sovereign has two tasks in Hobbes’ extended state of nature. First, the sovereign must make the Foole aware of the long-term consequences of her behavior, so that the Foole realizes that she is playing an assurance game and not a one-shot PD game or an assurance dilemma game. As a consequence, defection is no longer the strictly dominant strategy for the Foole in the state of nature. Second, the sovereign must solve the problem of assurance for both the Foole, who is now aware of the true nature of the game in the state of nature, and all prudent individuals. This allows rational individuals, foolish and prudent, to leave the state of nature and to form society.

Once society is established and the individuals repeatedly interact with one another in a social framework, the sovereign’s main task is to ensure peaceful cooperation by preventing the citizens from free riding, even if they are shortsighted from time to time and, in this sense, foolish. To this end, the sovereign must introduce sanctions for non-cooperative behavior that are sufficiently harsh to outweigh the potential gains from unilateral defection.

---

According to Hobbes, the sovereign must exclude free riders from the cooperative framework, or even take their lives, if they are detected.

In game-theoretic terms, the sovereign is assumed to play a grim strategy with the citizens once society is established. I will clarify and support Hobbes’ argument for the grim strategy by arguing that it best allows the sovereign to ensure peaceful cooperation in a world where not all free riders are necessarily detected, and thus best allows the sovereign to fulfill her task in the post-natural state. That is, compared to existing game-theoretic interpretations of Hobbes, I argue that the sovereign plays a grim strategy with the citizens once society is established, and not the individuals with one another in the state of nature.

The argument proceeds as follows. In section II, I lay out the characteristics of Hobbes’ state of nature that are relevant for my analysis. In section III, I translate Hobbes’ description of the state of nature into the language of game theory and explain the standard interpretations of Hobbes’ state of nature as a one-shot and iterated PD game. In section IV, I argue that the real game that rational, prudent individuals play in Hobbes’ state of nature is an assurance game. Thereby, I reject also the interpretation of Hobbes’ state of nature as an assurance dilemma game. In section V, I discuss Hobbes’ extended state of nature that regards the Foole as a member of the state of nature. In section VI, I clarify the rationale for the sovereign to play a grim strategy with the citizens once society is established.

II. HOBBES’ STATE OF NATURE

The starting point for Hobbes’ moral and political theory is, as it is for all contractarian theories, the state of nature. Hobbes thinks that human beings are by nature rational egoists. They pursue their own good, which is determined by their desires. Good is an object of an individual’s appetite, and bad is an object of her aversion. Nevertheless, although the good is agent relative in the state of nature, Hobbes believes that human beings have at least one desire in common. They want to preserve their lives. This desire is generally the strongest human desire and, as such, has priority over all other desires. The fulfillment of this desire, however, is threatened in the state of nature, because the state of nature is a war of almost all against all, according to Hobbes, for the following reasons.

Hobbes believes that human beings think that their mental and physical capacities are roughly equal by nature, so that no person can completely dominate another in the state of nature. The weakest is able to kill the strongest:

[T]he difference between man, and man, is not so considerable [...]. For as to the strength of body, the weakest has strength enough to kill the strongest, either by secret machination, or by confederacy with others, that are in the same danger with himselfe.\(^3\)

In addition, Hobbes assumes that individuals are free to use their mental and physical capacities to do whatever they regard as necessary to preserve their lives in the state of nature. This right of nature allows individuals even to take the lives of others, if such action is necessary to preserve their own lives. As a result, individuals have threat capacities against each other in the natural state.4

In short, Hobbes’ state of nature is a moral vacuum, according to the contractarian view, because individuals have no contractual obligations to each other in the state of nature.5 This characteristic of Hobbes’ state of nature, together with the assumptions of competition, diffidence, glory, and man’s ‘perpetuall and restlesse desire of Power after power’6 in an environment of scarce resources, leads to a situation of severe conflict.

Due to the high potential for conflict in the state of nature, rational individuals recognize, according to Hobbes, that striking first in their struggle for life provides them with the necessary advantage to survive in the state of nature. Rational individuals engage in preemptive action in the state of nature to increase their chances of survival. They will attack their enemies first.7 The potential of preemptive action ultimately makes the state of nature a war of almost all against all that destroys all productive efforts, because each individual is in constant fear of losing her life, which is in conflict with an individual’s desire for self-preservation. What will rational individuals do in this situation?

Hobbes argues that rational individuals have not only a desire to preserve their lives, but also an interest in securing their long-term well-being. All men possess the natural capacity for prudence:

[T]he object of mans desire, is not to enjoy once onely, and for one instant of time; but to assure for ever, the way of his future desire. And therefore the voluntary actions, and inclinations of all men, tend, not only to the procuring, but also to the assuring of a contented life [...].8

Rational, prudent individuals are forward-looking. They aim to secure the satisfaction of their desires not only today, but also in the future.9 According to

---

4 See Hobbes, **Leviathan**, Part I, Chapter XIV.
5 Hobbes argues, however, that rational individuals have certain prudential moral obligations in the state of nature that derive from the laws of nature, which are valid, *in foro interno*, already in the state of nature, as I will clarify. For a more detailed discussion of this point, see Hobbes, **Leviathan**, Part I, Chapter XV and Tom Sorell, ‘Hobbes’s Moral Philosophy’ in *The Cambridge Companion to Hobbes’s Leviathan*, ed. Patricia Springborg (Cambridge, 2007), pp. 128-153, pp. 133-135.
6 Hobbes, **Leviathan**, Part I, Chapter XI.
7 See Hobbes, **Leviathan**, Part I, Chapter XIII. Gregory Kavka argues in ‘Hobbes’s War of All Against All’, p. 297, that it is not necessarily rational always to attack first in the state of nature, because if an individual would behave aggressively as such, she would expose herself to the defensive violence of those attacked. She would identify herself as an especially dangerous person who should be eliminated first. For a further discussion of this point, see also Kavka, *Hobbesian Moral and Political Theory*, pp. 83-125.
8 Hobbes, **Leviathan**, Part I, Chapter XI. See also Part I, Chapter X.
9 To be clear, prudence itself can be acquired only by past experiences, that is, by looking backwards. A prudent individual analyzes the past with the aim of identifying the relevant causal relationships
Hobbes, prudence dictates to rational individuals in the state of nature the laws of nature, which demand that an individual is forbidden to do that which is destructive to her life or takes away the means to preserve it. The laws of nature are eternal theorems of prudence that can be summed up in the maxim: ‘Do not that to another, which thou wouldest not have done to thy selfe’.10

The first law of nature is to seek peace, if there is hope of obtaining it, and to do so by the means that are most adequate to reach this end. That is, under the circumstances described, prudent individuals expect that peaceful cooperation is most beneficial for them, and instrumental rationality then forces them to specify the means that are necessary to reach this end. Hobbes argues that a necessary means for peaceful cooperation is the formation of society.

The second law of nature states that society can be established only if all individuals in the state of nature lay down their rights of nature and transfer them to a common authority, the sovereign, who is not part of the community to be formed. According to Hobbes, only if all individuals entirely give up their rights to govern themselves and transfer these rights to an external authority with unlimited power over them, can society be established.

The third law of nature demands that individuals must keep their contracts (once their contracts are binding).11 That is, once society is established, its citizens must do what they have agreed to do in order for peaceful cooperation to be possible. Hobbes defends the remaining laws of nature by showing that failure to follow them would lead to an outbreak of hostilities, and thus would endanger peaceful cooperation. The laws of nature are the roadmap to enduring peace, according to Hobbes.

In addition, Hobbes argues that to ensure peaceful cooperation once society is established, the sovereign must have not only unlimited power, but also undivided power. Monarchy is the best form of government because it unites, inter alia, the public and private interests more closely than any other form of government, it does not allow for dissent among the rulers, and it guarantees continuity.12 Any more democratic form of government is likely to lead to the dissolution of society.
Democracy cannot absolutely secure internal peace, but can do so only under favorable conditions, according to Hobbes.

I will not discuss further the key concepts of Hobbes’ moral and political philosophy, such as the right of nature, laws of nature, authorization, and obligation, because such an analysis is not essential to my game-theoretic reconstruction of Hobbes’ argument. Further, I will not question the validity or soundness of Hobbes’ arguments that (i) the state of nature is a war of almost all against all,13 (ii) an absolute sovereign with full legislative, executive, and judicative powers is needed to secure peace, and (iii) the sovereign has a right to punish her citizens, which includes the right to take their lives.14 My aim is only to translate Hobbes’ state of nature argument, as closely as possible, into the language of standard game theory.15 To this end, I will make simplifying assumptions, and certain aspects of Hobbes’ argument cannot be modeled adequately by game theory at all. Nevertheless, I do not aim to criticize Hobbes’ argument, as is often the objective of recent game-theoretic literature on Hobbes, but only to reconstruct his argument by means of standard game theory.

III. HOBBES AND THE PRISONER’S DILEMMA GAME

Let us assume that the individuals in the state of nature behave as if they were to maximize their expected utility, and that the assumptions of common knowledge of rationality and complete information hold. Further, let us restrict the analysis, for simplicity, to a two-person state of nature with an external third party that functions as a possible ruler. Given these simplifying assumptions, is Hobbes’ state of nature best modeled by a one-shot PD game?

Rational individuals face the following problem of collective action in Hobbes’ state of nature. If each individual keeps her right to all things, the individuals remain in the state of nature and face a war of almost all against all. This situation

13 For support of Hobbes’ conclusion that anarchy leads to a war of (almost) all against all and a more precise description when this conclusion holds, see, for example, Gregory Kavka, *Hobbesian Moral and Political Theory*, pp. 126-156; and Peter Vanderschraaf, *War or Peace? A Dynamical Analysis of Anarchy*, pp. 243-279 and *Game Theory Meets Threshold Analysis: Reappraising the Paradoxes of Anarchy and Revolution*, *The British Journal for the Philosophy of Science* (2008), pp. 579-617, pp. 600-606.
15 By *standard game theory*, I mean non-evolutionary game theory. Standard game theory is methodologically more adequate to capture Hobbes’ state of nature argument than evolutionary game theory, because Hobbes’ state of nature serves as a device to justify the absolute sovereign. Hobbes’ project is justificatory and static in the sense that he does not want to provide an evolutionary explanation for the actual existence of the absolute sovereign. David Hume’s state of nature argument, for example, is by contrast best modeled by evolutionary game theory, because Hume’s genealogy of morals aims to explain the emergence of the existing social and political structure.
is suboptimal for everyone because rational individuals expect that peaceful cooperation is more beneficial for them than remaining in the state of nature. As such, each individual has, *in foro interno*, an incentive to strive towards peace. *In foro externo*, however, each individual knows that if she lays down her right to everything and her fellows do not, she risks losing her life.

This line of reasoning seems to imply that the worst outcome for an individual in Hobbes’ state of nature is to cooperate when her opponent defects (cd), and the best outcome is to defect when her opponent cooperates (dc). The second best outcome is mutual cooperation (cc), and the third best outcome is mutual defection (dd). Given these options, a rational individual will not cooperate in the state of nature if her primary aim is to preserve her life, because cooperation puts her life at risk.

Following these considerations, the individuals’ situation in the state of nature can be represented, in a stylized form, by a one-shot PD game, as shown in Table 3.1.

<table>
<thead>
<tr>
<th></th>
<th>Cooperate</th>
<th>Defect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperate</td>
<td>4, 4</td>
<td>-2, 6</td>
</tr>
<tr>
<td>Defect</td>
<td>6,-2</td>
<td>0, 0</td>
</tr>
</tbody>
</table>

Table 3.1 – One-shot PD game

The payoffs shown in the matrix reflect the all-things-considered utilities that the individuals expect to gain if they either cooperate or defect in the state of nature in a one-off decision situation, depending on their opponent’s behavior. An individual can choose between two pure strategies, cooperation and defection. In the case where an individual cooperates and her opponent defects, the cooperator’s payoff is negative because she is most likely to lose her life, given the fierce conditions in Hobbes’ state of nature.

In a one-shot PD game, defecting is the strictly dominant strategy. As such, if the situation of rational individuals in Hobbes’ state of nature is adequately modeled by a one-shot PD game, the individuals will remain in the state of nature and end up with a suboptimal outcome, both individually and collectively. In order for rational individuals to leave the state of nature, an external authority is needed that transforms the one-shot PD game into another game in which defection is not the strictly dominant strategy, such as the assurance game, by introducing sanctions for defective behavior. Before I explain the nature of the assurance game in the next section, let me briefly discuss another common game-theoretic interpretation of Hobbes’ state of nature.
So far, the analysis has been restricted to a one-off decision situation where two rational individuals have a choice either to defect or to cooperate in the state of nature. If we now consider a situation of repeated interaction in which the parties can decide what to do each time they meet, then it is tempting to model Hobbes’ state of nature by an iterated PD game. In the following, I argue that, depending on the particular interpretation of the game, the iterated PD game is either unable to add new considerations to the discussion, or it is inadequate to describe Hobbes’ state of nature.

If Hobbes’ state of nature is modeled by an iterated PD game and a finite time horizon (that is, a finite number of iterations) is assumed, which seems adequate from the perspective of mortal human beings, then the problem of backward induction arises under the idealized circumstances described. That is, because it is rational for an individual to defect in the last round of the game because her opponent will not be able to reciprocate, it is rational for the individual’s opponent to defect in the penultimate round of the game, and so on up to the first round of the game. As such, although the game is repeated, rational individuals will defect in each round of the game.

Note that in order for the problem of backward induction to arise, the players do not need to know the exact number of iterations of the game, which is unlikely in the state of nature because the state of nature lacks a social order that can sufficiently stabilize the individuals’ expectations with regard to cooperation. Instead, the players need to know only that the number of plays is limited, and that the chance of entering a new round of the game is very low, which is the case in Hobbes’ state of nature where life is ‘solitary, poore, nasty, brutish, and short.’

Thus, from the perspective of modeling the problem of collective action that arises in Hobbes’ state of nature, the finitely repeated PD game does not add any essential considerations compared to the one-shot PD game. De facto, the individuals are assumed to play a finite sequence of one-shot PD games with one another.

If the state of nature is modeled by an infinitely repeated PD game, which is difficult to justify from the perspective of mortal human beings who live under the

---


19 If the assumption of complete information is dropped, then it may be rational for individuals to cooperate in a finitely repeated PD game. In this case, however, the problem of endgame imperfection arises, as Reinhard Selten and Rolf Stoecker have shown in ‘End Behavior in Sequences of Finite Prisoner’s Dilemma Supergames: A Learning Theory Approach’, Journal of Economic Behavior and Organization 7 (1986), pp. 47-70.
fierce conditions in Hobbes’ state of nature where their lives are constantly at risk, defective behavior is not necessarily the strictly dominant strategy. The reason for this is that cooperative moves may elicit cooperative moves from others, assuming that individuals who generally cooperate survive, which is another unrealistic assumption in Hobbes’ state of nature. In repeated interactions, the behavior of keeping contractual promises, according to Luce and Raiffa, ‘is in a sort of quasi-equilibrium: it is not to the advantage of either player to initiate the chaos that results from not conforming, even though the non-conforming strategy is profitable in the short run (one trial).’

Further, in repeated interactions, rational individuals may also render their choices dependent upon their opponents’ behaviors in previous encounters. This approach may allow individuals to gain higher payoffs in the long run compared to strictly cooperative or defective behavior. The tit-for-tat strategy, for example, where an individual initially cooperates and then responds in kind to her opponent’s behavior, is commonly regarded as an optimal strategy in repeated games.

As such, if rational individuals play an infinitely repeated PD game in the state of nature, cooperative behavior may emerge and may be sustained, which means that rational individuals may be able to leave the state of nature without the help of an external authority. Hobbes argues, however, that under the fierce conditions described by him, rational individuals cannot expect to repeatedly interact with one another in the state of nature and it is generally irrational to cooperate in the absence of the sovereign. Therefore, if we take Hobbes’ state of nature argument seriously, which I do in this article, and we do not try to prove him wrong, then the problem of collective action that arises in Hobbes’ state of nature cannot be adequately modeled by an infinitely repeated PD game.

---


22 See, however, Jean Hampton who defends such an interpretation in *Hobbes and the Social Contract Tradition*, pp. 80-89. I think that Hampton’s argument is mistaken for the following reason. Hampton argues that the individuals in Hobbes’ state of nature play an infinitely repeated PD game in which it may be rational to cooperate. However, due to their short-sightedness, the individuals do not adequately consider the long-term benefits of cooperation and, as a consequence, defect in the state of nature. I agree with Hampton’s argument that some individuals may misconceive their situation in the state of nature, and consequently defect. However, all other forward-looking individuals who identify the real game in the state of nature do not play a one-shot or iterated PD game. Instead, they play an assurance game, because even if rational individuals are forward-looking and they believe that others are too, they will not generally cooperate with their fellows in Hobbes’ state of nature because of the problem of assurance. The problem of assurance is the main problem of collective action that prevents cooperation in Hobbes’ state of nature, and thus
In sum, if Hobbes’ state of nature is to be modeled by a PD game, then the one-shot PD game seems to be the most plausible interpretation of Hobbes’ argument. In the following, however, I argue that only an assurance game can adequately model the problem of collective action that arises in Hobbes’ state of nature.

IV. THE PROBLEM OF ASSURANCE
A (one-shot) assurance game has the following structure (Table 4.1).

<table>
<thead>
<tr>
<th>A</th>
<th>Cooperate</th>
<th>Defect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooperate</td>
<td>4, 4</td>
<td>-2, 3</td>
</tr>
<tr>
<td>Defect</td>
<td>3, -2</td>
<td>0, 0</td>
</tr>
</tbody>
</table>

Table 4.1 – Assurance game

An assurance game has two Nash equilibria in pure strategies, (cc) and (dd), but no strictly dominant strategy. The payoff-dominant outcome of the game is (cc). Whether this outcome is realized depends on the individuals’ beliefs about their opponents’ behaviors. If an individual believes that her opponent will defect, then she will also defect. If an individual believes that her opponent will cooperate, then she will also cooperate. More precisely, given the numbers depicted by Table 4.1, cooperation is the utility maximizing choice for A (B), only if A (B) expects B (A) to cooperate with a probability greater than 2/3.\(^23\) If A (B) believes that the probability for B (A) to cooperate is lower than 2/3, she will defect, and the risk-dominant equilibrium (dd) will be realized.

In the following, I argue that Hobbes’ state of nature is best modeled by an assurance game, and not by a one-shot PD game or an assurance dilemma game, for two reasons. First, neither the one-shot PD game nor the assurance dilemma game can model the problem of collective action that must be solved in Hobbes’ state of nature in order for society to be established. Second, these two games cannot model the situation of prudent individuals, who are the main type of actor in Hobbes’ state of nature and who identify the real game in the state of nature. The assurance game, by contrast, can model both aspects. I will discuss the first part of my argument in this section and the second part in the next section.

To recall, Hobbes’ second law of nature states that society can be established only if all individuals in the state of nature lay down their rights to everything and

\(^{23}\) Cooperate\(=4(p) + (-2)(1-p)\)  Defect\(=3p+0(1-p)\)  Equilibrium: Cooperate=Defect  \(\rightarrow p=2/3\)

the formation of society. The problem of assurance, however, cannot be adequately modeled by the one-shot or iterated PD game, as I will clarify in section IV.
transfer these rights to an external authority, the sovereign. The third law of nature demands that, in order for peaceful cooperation to be possible and maintained, individuals must keep their contracts once their contracts are binding. By separating the second law of nature from the third law of nature, Hobbes clearly recognizes that two distinct problems must be solved in order for (i) rational individuals to leave the state of nature and (ii) peaceful cooperation to be maintained once society is established: the problem of assurance and the problem of compliance, respectively.

The problem of assurance is expressed by the second law of nature. The problem of assurance is the key problem of collective action that must be solved in Hobbes’ state of nature, because only if all individuals lay down their rights to govern themselves and transfer these rights to an external authority who is strong enough to enforce any subsequent agreements will the agreements be binding and a valid social contract that establishes society can be made. This is because, according to Hobbes, ‘[c]ovenants, without the [sovereign’s] Sword, are but Words, and of no strength to secure a man at all.’24 Words alone cannot secure one’s life, and thus cannot be trusted in the state of nature.25

Therefore, once the individuals in the state of nature have transferred their rights of nature to the sovereign, they must agree to keep their contracts, as demanded by the third law of nature, because contracts are binding for individuals only if they voluntarily agree to keep their contracts, and such agreements are binding only once a common power is in place that is strong enough to enforce the agreements, according to Hobbes. The individuals’ agreements to keep their contracts after the sovereign is instituted represent the final step of the formation of society.26 The agreements represent the social contract and the origin of justice, which demands, according to Hobbes, simply to keep one’s agreements.27

24 Hobbes, Leviathan, Part II, Chapter XVII.
26 To clarify, the sovereign herself does not come into power by a contract, but by the individuals’ surrender of their rights of nature, which rational individuals regard as beneficial in the state of nature. That is, the sovereign is not empowered by a contract between the citizens and the leader, but by a self-interested alienation of the individuals’ rights of nature. The individuals give up their rights to govern themselves and transfer them to the sovereign, and in doing so, the sovereign acquires, as a free gift, the power to govern the citizens and to punish them for their failings. The sovereign herself, however, is not bound by any agreement, because she neither transfers any rights to the citizens, nor does she make a contract with them. Instead, the social contract is made among the citizens. The citizens metaphorically promise each other, in bilateral agreements, to obey the sovereign and to keep their private contracts after the sovereign is empowered. For a further discussion of this point, see in particular Gregory Kavka, Hobbesian Moral and Political Theory, pp. 304-305. As Kavka notes (ibidem, p. 386), ‘the parties are bound to their sovereign by a double tie of obligation—by contract, each owes every other citizen obedience to the sovereign, and by the obligation to carry out the terms of a free gift [...], each directly owes the sovereign obedience.’
The content of justice is determined by two different types of agreement. First, the sovereign is authorized to make the civil laws and to regulate all other issues that are relevant to ensure the citizens’ safety. The sovereign is the sole legislator and she has full executive and judicative powers. Second, the citizens are allowed to make private contracts with each other as long as their agreements do not conflict with the civil laws imposed on them by the sovereign.

In sum, for rational individuals to leave the state of nature and to establish society, the individuals must transfer their rights of nature to the sovereign and agree to keep their social and private contracts once the sovereign is in place. A rational individual will cooperate with her fellows only if she can be sure that her fellows have taken both of these steps. Or, vice versa, in order for a rational individual to cooperate with her fellows, she must be sure that only individuals who have laid down their rights of nature and who have agreed to keep their contracts are members of society. Otherwise, an individual risks losing her life if she cooperates with others.

This line of reasoning shows that the primary problem of collective action and, strictly speaking, the only problem of collective action that must be solved in Hobbes’ state of nature in order for society to be established, is the problem of assurance. The problem of compliance does not arise in Hobbes’ state of nature. It arises only after society is established. To clarify, an essential part of the social contract is the individuals’ promises to keep their contracts and their trust in the sovereign’s power to enforce any agreements that have been made. However, only once the sovereign is in place and the promises are made, and thus the social contract is ratified, can the problem of compliance arise, because only then are individuals bound to do what they agreed to do. In the state of nature, rational individuals are not obligated to keep their contracts, precisely because of the problem of assurance.

The only exception to this rule is when an individual’s contract partner (naively) fulfills her part of the contract in the state of nature. Then, an individual is bound also to fulfill her part of the agreement, according to Hobbes. If the individual were to unilaterally defect in this situation, she would risk that others will not cooperate again with her, which lowers her chances of survival in the state of nature, as I will explain shortly.

---

29 Stated differently, in foro interno an individual is obliged to fulfill her part of the contract in the state of nature. In foro externo, however, she is bound to do so only if she can be sure that her opponent fulfills her part, which is the case if her opponent has already performed. Translated into the language of game theory, if the first mover cooperates in an assurance game, it is rational for the second mover also to cooperate.
The problem of assurance, which characterizes Hobbes’ state of nature, cannot be modeled by a one-shot PD game, because it is never rational to cooperate in a one-shot PD game. In Hobbes’ state of nature, however, non-cooperation is not the strictly dominant strategy per se. Instead, a rational individual makes her behavior dependent on whether she can trust her fellows to cooperate in the game of cooperation that establishes society. Of course, Hobbes is pessimistic with regard to this assumption. He argues that a rational individual generally cannot trust that her fellows will cooperate with her in the state of nature due to the fierce conditions in the natural state.

However, this does not mean that it is never rational to cooperate in Hobbes’ state of nature, as the one-shot PD game suggests. Hobbes does not argue that trust relationships do not exist at all in the state of nature. Instead, rational individuals may trust other individuals in the state of nature, for example, individuals who are close to them, such as their family members and good friends. Further, Hobbes argues that it is rational, and even necessary, to form defensive coalitions, or confederacies as he calls them, to increase one’s power and thus one’s chances of survival in the state of nature.

In short, Hobbes argues that rational individuals may trust other individuals in the state of nature and, as a result, cooperate with them. But this is true only for specific individuals with whom one holds a close relationship. In general, it is irrational to trust one’s fellows in the natural state, because they are potential enemies who aim to take one’s life in order to increase their own chances of survival. Under these conditions, a rational individual generally will not cooperate with her fellows in the state of nature if her primary aim is to preserve her life.

Hobbes thinks that the small islands of trust that arise among particular individuals in the state of nature are insufficient to guarantee one’s security, and thus to establish a proper society that encompasses all inhabitants of the state of nature. To this end, the problem of assurance must be solved, which can be done only by an authority that is not part of the social contract, such as the sovereign. The sovereign is the only independent authority that can guarantee that only individuals who lay down their rights to everything and who agree to keep their contracts are members of society. The sovereign will exclude all other individuals from the cooperative framework, and she must exclude them, as I will argue in section VI. Only the sovereign, as an external authority, can solve the problem of assurance in the state of nature by replacing interpersonal trust, or the lack thereof, by institutional trust.

Thus, in order for rational individuals to leave the state of nature, and in this sense to cooperate with each other, the sovereign does not have to change the structure of the game that rational individuals play in Hobbes’ state of nature to

---

31 For support of this conclusion see Gregory Kavka, *Hobbesian Moral and Political Theory*, pp. 126-178.
transform a one-shot PD game into another game that renders mutual cooperation rational. Instead, the sovereign only must assure the individuals that if they lay down their rights to everything and agree to keep their contracts, all other individuals will do so, too. The sovereign must only assure the individuals’ trust in one another to establish society.

To be precise, the sovereign could not change the game that rational individuals play in Hobbes’ state of nature, because the sovereign does not have legislative and executive powers in the state of nature. The sovereign can change the game only after she is instituted and the social contract is signed, but not in the state of nature. As such, if rational individuals were to play a one-shot PD game in Hobbes’ state of nature, the sovereign could not solve the problem of collective action that prevents the formation of society in Hobbes’ state of nature.

Once the problem of assurance is solved by the institution of the sovereign, and the individuals have signed the social contract and interact repeatedly with one another in a social framework, the problem of compliance arises, because then all agreements that have been made are binding and must be kept in order to maintain peaceful cooperation, as expressed by Hobbes’ third law of nature. Once society is established, the individuals’ promises to keep their contracts are insufficient to guarantee the citizens’ security. Instead, the individuals must also do what they have promised, even if they think later, in particular instances, that non-cooperative behavior is more beneficial for them. The problem of compliance is best modeled by a one-shot PD game, because the highest payoff in this game is achieved by unilateral defection, and defection is always the rational strategy in a one-shot PD game.

However, I do not want to suggest that Hobbes’ state of nature is best modeled by a combination of the assurance game and the one-shot PD game, as argued by Gregory Kavka.32 The combined game, which is called the assurance dilemma game, has the following structure (Table 4.2).

---

In our case, individual A’s payoff structure is identical to the structure of the assurance game, and individual B’s payoff structure is identical to the structure of the one-shot PD game. As a consequence, if we assume common knowledge of rationality and complete information, individual B’s strictly dominant strategy is to defect. That is, individual B defects irrespective of individual A’s behavior. If individual B defects, however, then individual A’s best response is also to defect. In short, mutual defection is the only Nash equilibrium in an assurance dilemma game.

Therefore, if rational individuals play an assurance dilemma game in Hobbes’ state of nature, they will strictly defect and not be able to leave the state of nature, even if the sovereign is in place. In this case, the institution of the sovereign is insufficient to establish society. To form society, the sovereign would have to change individual B’s payoff structure from a one-shot PD game (at least) to an assurance game in order to make mutual cooperation rational, which transforms the assurance dilemma game into an ordinary assurance game. However, the sovereign does not have such legislative and executive powers in the state of nature, but only once society is established.

This line of reasoning shows that Hobbes’ state of nature cannot be modeled by an assurance dilemma game, because this game cannot model the key problem of collective action that arises in Hobbes’ state of nature, which is the problem of assurance, and its related trust considerations. To clarify, although the assurance dilemma game carries the word assurance in its name, trust considerations do not play a role in this game, because even if individual B trusts that individual A will cooperate, individual B will defect, and individual A cannot trust that individual B will cooperate because defection is individual B’s strictly dominant strategy. Further, if individual B cooperates, due to a mistake, and individual A anticipates this move, individual B does not have to trust that individual A will cooperate, because in this case cooperation is individual A’s utility maximizing choice. In short, trust cannot render mutual cooperation rational in an assurance dilemma game, but only individual A’s expectation that individual B will make a mistake.

In more general terms, any game that has the structure of a one-shot PD game (and if so only partly, as is the case for the assurance dilemma game) is inadequate
to model Hobbes’ state of nature, because the one-shot PD game cannot model the
problem of assurance. It can model only the problem of compliance. The problem
of compliance, however, does not arise in Hobbes’ state of nature, but only after
society is established. Compliance is the key problem that must be solved in order
for peaceful cooperation to be maintained once society is established.

V. HOBBES’ EXTENDED STATE OF NATURE
The conceptual distinction drawn in the previous section is often overlooked in the
game-theoretic literature on Hobbes, in particular in the literature that is inspired
by evolutionary game theory, in which a broader reading of Hobbes’ state of nature
is adopted that captures the problems of both assurance and compliance, because
the shortsighted Foole is interpreted as a member of Hobbes’ state of nature.

To clarify, if one adopts a strict reading of Hobbes’ state of nature, as in the
previous section, then only prudent individuals exist in Hobbes’ state of nature
from the perspective of our game-theoretic analysis. The Foole is not part of
Hobbes’ state of nature properly conceived:

The Foole hath sayd in his heart, there is no such thing as Justice; […] He does not […] deny, that
there be Covenants […] and that […] breach of them may be called Injustice, and the observance of
them Justice: but he questioneth, whether Injustice […] may not sometimes stand with that Reason,
which dictateth to every man his own good.33

The Foole makes contracts, but then breaks her word when she is called to fulfill
her part, if breaking her word is more beneficial for her in the particular instance.
As such, if we follow Hobbes’ definition of justice, an individual cannot be a Foole
in Hobbes’ state of nature because she cannot act unjustly in the state of nature. An
individual can act unjustly only once society is established.

In this section, however, I follow the broader reading of Hobbes’ state of nature
that makes the Foole a member of the state of nature. That is, I use the term Foole
more broadly to refer to all rational but shortsighted individuals, independent of
whether they are in the state of nature or in society. I argue that even if one adopts
this extended view of Hobbes’ state of nature, there is a second reason that Hobbes’
state of nature is best modeled by an assurance game, and not by a one-shot PD
game or an assurance dilemma game.

The argument, in a nutshell, is this. The main type of actor in Hobbes’ extended
state of nature is prudent individuals, and for prudent individuals only the
problem of assurance arises in the state of nature, which can be adequately
modeled only by the assurance game. However, because Fools, who misconceive
their situation, are also part of Hobbes’ extended state of nature, and this is
common knowledge, the problem of compliance arises in addition to the problem
of assurance. In this case, and only in this case, the one-shot PD game and the
assurance dilemma game become useful tools to model particular types of

33 Hobbes, Leviathan, Part I, Chapter XV.
interaction in Hobbes’ extended state of nature, namely, the interaction of Fooles with each other and the interaction of prudent individuals with Fooles, respectively. Let me first address the situation of prudent individuals.

As argued in section III, Hobbes assumes that human beings are generally prudent. A prudent individual is forward-looking in the sense that she properly considers the future consequences of her behavior within her forward-looking perspective, based on past experiences about her empirical and social world. Because a prudent individual is forward-looking, she understands that if she does not lay down her right of nature in the state of nature and she does not agree to keep her contracts, society cannot be established, according to Hobbes. In other words, a prudent individual is fully aware of the meaning of the laws of nature, because they are theorems of prudence. As a consequence, a prudent individual cannot expect that strictly non-cooperative behavior is beneficial for her in the state of nature, in particular for the following reason.

Hobbes argues that an individual who does not cooperate at all in the state of nature and who deceives even the individuals who trust her in the natural state will not be able to survive, because

[i]n a condition of Warre, wherein every man to every man [...] is an Enemy, there is no man can hope by his own strength, or wit, to defend himself from destruction, without the help of Confederates; where every one expects the same defence by the Confederation, that any one else does: and therefore he which declares he thinks it reason to deceive those that help him, can in reason expect no other means of safety, than what can be had from his single power.34

An individual who deceives others who trust her in the state of nature does not act unjustly, because there is no injustice in Hobbes’ state of nature. However, such an individual acts imprudently because she acquires a bad reputation, and consequently, she will not be able to form defensive coalitions. That is, if trust relationships are established among particular individuals in the state of nature, unilateral defection is likely to decrease an individual’s chances of survival, assuming that communication takes place among the inhabitants of the state of nature.35 A prudent individual considers this fact, and given common knowledge of rationality, she expects that other individuals do so, too. As a consequence, defective behavior is not the strictly dominant strategy for prudent individuals in Hobbes’ state of nature.

This line of reasoning shows that the situation of prudent individuals in Hobbes’ state of nature is best modeled by an assurance game. In an assurance game, defection is not the strictly dominant strategy, as is the case in the one-shot PD game. In the assurance game, the payoff from unilateral defection is lower than the

34 Hobbes, Leviathan, Part I, Chapter XV.
35 For support of Hobbes’ claim that unilateral defection, or offensive violation of agreements, is usually irrational under the circumstances described and the assumption of common knowledge of past behavior, see Peter Vanderschraaf, ‘Covenants and Reputation’, Synthese 157 (2007), pp. 167-195.
payoff from mutual cooperation (3<4, see Table 4.1), because a prudent individual fully takes into account the negative consequences of her non-cooperative behavior. A prudent individual realizes that if she categorically defects in the state of nature, she will not be able to survive and become a member of society. As a consequence, the highest payoff for her is not achieved by unilateral defection, as suggested by the one-shot PD game, but by mutual cooperation. This implies that if all individuals always were to act prudently and if this were common knowledge, only the problem of assurance would arise in Hobbes’ extended state of nature, and not the problem of compliance. In practice, however, not everyone acts prudently all the time, which brings us to the second type of actor in Hobbes’ extended state of nature, the Foole.

The Foole is, like prudent individuals, rational. She behaves as if she were to maximize her expected utility. However, the Foole is shortsighted because (i) she does not fully consider the negative consequences of strictly non-cooperative behavior, in particular not the negative consequences of unilateral defection, or (ii) she overestimates the short-term benefits of defecting due to, for example, disturbing factors such as spontaneous passions, or (iii) she thinks that her defective behavior will not be detected. As a consequence, the Foole expects the short-term benefits of defecting to be greater than the long-term benefits of cooperation, which makes defecting in the short run seem rational to her.\footnote{For a similar interpretation of Hobbes’ argument see Jean Hampton, *Hobbes and the Social Contract Tradition*, pp. 80-89. Hampton does, however, not explicitly distinguish between prudent individuals and Fooles, and I think that her short-sightedness account of conflict is mistaken, because it leads her to the wrong game-theoretic interpretation of Hobbes’ state of nature, as indicated in section III.}

To clarify, the Foole is forward-looking in the sense that she recognizes that establishing society and being a member of the cooperative framework (cc) is more beneficial for her than staying in the state of nature (dd). However, the Foole is shortsighted by not fully considering the negative consequences of strictly non-cooperative behavior in the state of nature. Due to her short-sightedness, the Foole is not fully aware of the precise nature of her situation and the relevance of the laws of nature for reaching her long-term goals. The Foole believes that she can have both the benefits of peaceful cooperation and the gains from exploiting others.

As a consequence, defection seems always to be the best strategy for the Foole in Hobbes’ extended state of nature, and her situation is best modeled by a one-shot PD game. If the Foole were to step back, however, and properly consider the long-term effects of her behavior, she would realize that she plays an assurance game. If the Foole were to subtract the negative consequences of her non-cooperative behavior from the payoff that she expects to achieve from unilateral defection, which are 3 utils in our example (6-3=3, compare Tables 3.1 and 4.1), she would
realize that she plays an assurance game. In this case, the one-shot PD game would be transformed into an assurance game, according to Hobbes’ argument.

This line of reasoning is supported by Brian Skyrms, who argues that if the shortsighted Foole were to consider the ‘shadow of the future’ in a repeated game, assuming for a moment that cooperation is sufficiently stable in the state of nature, the alleged PD game would be transformed into an assurance game for the Foole, under the circumstances described by Hobbes. If the Foole were to ‘open her eyes’, she would realize that the game that she really plays in Hobbes’ extended state of nature, and that prudent individuals correctly identify, is an assurance game. It is also interesting that David Gauthier, in The Logic of Leviathan, recognizes that the adequate long-term representation of the game that rational individuals play in Hobbes’ state of nature is an assurance game, although he does not call the game by its name, and he does not discuss the relevant trust considerations involved.

In sum, two different types of actor exist in Hobbes’ extended state of nature from the perspective of our game-theoretic analysis: prudent individuals and Fooles. The situation of prudent individuals, who are the main type of actor in Hobbes’ extended state of nature and who identify the real game in the state of nature, is best modeled by the assurance game. The situation of Fooles, by contrast, and their interactions with one another, is best modeled by the one-shot PD game, because defection is the strictly dominant strategy for Fooles in the state of nature due to their short-sightedness. Or, to be more realistic, one and the same person may sometimes be prudent and sometimes foolish. An individual may sometimes be forward-looking by fully considering the long-term effects of her behavior, and sometimes she may be shortsighted. Further, another reason that rational individuals may sometimes defect in practice, although they regard cooperation to be most beneficial for them both in the short run and in the long run, is that they may be weak-willed and, as a consequence, fail to put their decisions into action.

There are then two reasons that a prudent individual generally does not trust and thus does not generally cooperate with her fellows in Hobbes’ extended state of nature. First, given the fierce conditions described by Hobbes, a prudent individual cannot trust that her fellows in the state of nature will lay down their rights of nature and agree to keep their contracts, because doing so makes her vulnerable if others do not do so. Second, even if a prudent individual were to believe that her

---

37 To provide an explanation for this number, let us assume that the Foole expects to gain 6 utils if she free rides without being detected, as reflected by Table 3.1, and a disutility of -2 utils if her defective behavior is detected because the sovereign takes her life. Now, if the Foole believes that the likelihood of being detected is 3/8, then the one-shot PD game will be transformed into the assurance game that is depicted by Table 4.1. \[ p = \text{probability that free riding is detected} \rightarrow \frac{3}{8} \]


39 See David Gauthier, The Logic of Leviathan, pp. 85-86.
fellows will lay down their rights of nature and transfer them to the sovereign, and
they generally keep their agreements, she fears that her fellows will sometimes be
shortsighted or weak-willed, and in this sense, foolish. As a consequence, a
prudent individual expects that her fellows will defect from time to time in the
state of nature, which may cost the individual her life. In the state of nature, ‘[h]e
that performeth first, has no assurance the other will performe after; because the
bonds of words are too weak to bridle mens ambition, avarice, anger, and other
passions, without the feare of some coërcive Power’, according to Hobbes.40

Assuming that these two reasons not to trust others in the state of nature are
common knowledge, non-cooperation is usually the rational strategy for prudent
individuals in Hobbes’ extended state of nature if their primary aim is to preserve
their lives, although prudent individuals play an assurance game and not a one-
shot PD game with one another in the state of nature. Moreover, if we take Hobbes’
description of the state of nature seriously, the first reason not to trust other
individuals in the state of nature is usually sufficient to prevent cooperation. The
existence of Fooles in Hobbes’ extended state of nature only aggravates the lack of
trust among individuals in the state of nature. In other words, the problem of
compliance is only a secondary problem in Hobbes’ extended state of nature.

This line of reasoning also clarifies that the assurance dilemma game is
inadequate to model the situation of prudent individuals in Hobbes’ extended state
of nature, because this game captures only the second reason for non-cooperative
behavior in the state of nature, which is short-sightedness. As such, the assurance
dilemma game can model only the interaction of prudent individuals with Fooles.
It cannot, however, model the trust considerations that are expressed by the first
reason for non-cooperative behavior, which arise even if prudent individuals
interact with one another.

In other words, the assurance dilemma game is not general enough to model the
situation of prudent individuals in Hobbes’ state of nature,41 and more
importantly, it cannot model the essential trust considerations that prevent
cooperation in the state of nature. The assurance game, by contrast, can model
both reasons for non-cooperative behavior in the state of nature, because a
prudent individual may not trust that her fellows will cooperate with her because
she expects her fellows to be very cautious in situations of social interaction that
may cost them their lives, as is generally the case in Hobbes’ state of nature. Or, a
prudent individual may not trust that her fellows will cooperate with her because
she expects them to be Fooles.

Following these considerations, the sovereign has two tasks in Hobbes’ extended
state of nature, which is populated by both foolish and prudent individuals. First,

40 Hobbes, *Leviathan*, Part I, Chapter XIV.
41 See also Peter Vanderschraaf, ‘War or Peace? A Dynamical Analysis of Anarchy’, p. 257, who rejects
the assurance dilemma game for his evolutionary analysis of Hobbes’ state of nature for a similar
reason.
the sovereign must make the Foole aware of the negative consequences of strictly non-cooperative behavior in the state of nature, which lowers the Foole’s chances of survival in the natural state and does not allow her to become a member of society. The sovereign must compel the Foole to be prudent. As clarified, the sovereign cannot transform the one-shot PD game or the assurance dilemma game that the Foole mistakenly believes she is playing in the state of nature into an assurance game, because the sovereign does not have the power to do so in the state of nature. The sovereign can only make the Foole aware of the real nature of the game in the state of nature by making clear to her the advantages of society and the necessary steps to establish society.

Second, the sovereign must solve the ‘real’ problem of assurance, as expressed by the first reason for non-cooperative behavior in Hobbes’ state of nature, for both the Foole, who is now aware of the true nature of the game in the state of nature, and all prudent individuals. The sovereign must ensure that only individuals who ratify the social contract are members of society by excluding all other individuals from the cooperative framework. The sovereign’s assurance provides rational individuals, foolish and prudent, with the final incentive to give up their rights of nature and to sign the social contract. The sovereign’s assurance allows rational individuals to leave the state of nature.

VI. THE FOOLE AND THE GRIM STRATEGY
Let us assume that rational individuals follow the logic of Hobbes’ argument and they understand the need to institute an absolute sovereign in order to leave the state of nature. According to Hobbes, there are then two ways to do so, either by acquisition or by institution.\footnote{See Hobbes, \textit{Leviathan}, Part II, Chapter XVII.} In the former case, no game-theoretic explanation of the process of instituting the sovereign must be provided, because a leader who is strong enough simply subdues all other individuals in the state of nature. In the second case, where the inhabitants of the state of nature are assumed to submit themselves voluntarily to a sovereign, the Hobbesian problem of social order arises, which doubts that the behavior of rational self-interested individuals alone can lead to the institution of any social order due to the gap between individual interests and genuinely different group interests. As Jean Hampton has clarified, however, rational individuals can solve the Hobbesian problem of social order.\footnote{See Jean Hampton, \textit{Hobbes and the Social Contract Tradition}, pp. 147-188.} Rational individuals are able to select a leader in Hobbes’ state of nature, because this \textit{leadership-selection problem}, as Hampton calls it, has the structure of a Battle of the Sexes game in which rational individuals can realize the optimal outcome without external help, and rational individuals are able to empower the selected
leader to be the sovereign. As such, an adequate game-theoretic explanation for the institution of the sovereign can be provided.44

Let us assume that the individuals in the state of nature manage to institute an absolute sovereign and they repeatedly interact with one another in the post-natural state. That is, let us extend our analysis from a two-person state of nature to a $n$-person society, where $n$ is assumed to be greater than two, and explicitly discuss a situation of repeated interaction. Then, once society is established, the only remaining task of the sovereign, apart from defending the community against external threats, is to guarantee the citizens’ safety by legislating and enforcing the civil laws that are necessary to ensure peaceful cooperation. The sovereign must do everything in her power to prevent the citizens from falling back into the state of nature, because then she loses her right to exist.45

Assuming that all citizens are rational, the only internal threat to peaceful cooperation after society is established is the Foole. The Foole is tempted to break the agreements made because she does not properly consider the long-term consequences of her non-cooperative behavior. The Foole regards free riding as beneficial, because she thinks that such behavior allows her to secure the benefits of peaceful cooperation without having to pay (all of) the costs for maintaining the cooperative scheme. According to Hobbes, the sovereign must prevent free riding, because it threatens peace. Free riding endangers the security of all citizens who obey the civil laws, and it undermines the sovereign’s credibility of being able to protect the citizens from one another. If the sovereign cannot protect the citizens from each other in the post-natural state, then the individuals are more vulnerable in society than they were in the state of nature, because they forego preemptive action.

The problem of free riding is best expressed by the structure of the one-shot PD game for two reasons. First, although the individuals play a repeated game after society is established, free riders are shortsighted. They consider only their immediate short-term interests, or they overvalue the satisfaction of those interests. As such, the situation of free riders is best modeled by a one-shot game. Second, the incentive for free riding is inherent to the structure of the one-shot PD game, because an individual can gain the most in this game if she defects and her opponent cooperates, and defection is always the safe strategy in a one-shot PD game. As such, a rational individual is always tempted to free ride, and she hopes that she will be able to do so if she faces a situation of social interaction that has the form of a one-shot PD game.

In order to prevent free riding once society is established, the sovereign must transform the one-shot PD game that the shortsighted Foole believes she is

44 See also Gregory Kavka, *Hobbesian Moral and Political Theory*, pp. 179-188, who argues for a two-stage coordination game that allows rational individuals to institute the sovereign, although Kavka’s argument operates with some distance to Hobbes’ text.

playing, with regard to her payoff structure of the game, (at least) into an assurance game by introducing sanctions for defective behavior. In the post-natural state, the sovereign has the legislative and executive powers to do so, and because it is irrational to defect in an assurance game if one’s opponent cooperates, and the problem of assurance is solved after society is established, transforming the alleged one-shot PD game into an assurance game is sufficient to ensure cooperation.

In short, the sovereign must change the game that the Foole believes she is playing in the short run so that the Foole is a fool from her own perspective, if she free rides. In general, the Foole is a fool only if (i) her free-riding behavior is detected and (ii) the negative consequences of such behavior are sufficiently harsh to outweigh the benefits from unilateral defection. In Hobbes’ words, the Foole must be punished for non-cooperative behavior.

[I]f the harm inflicted be lesse than the benefit, or contentment that naturally followeth the crime committed, that harm is not within the definition [of punishment]; and is rather the Price, or Redemption, than the Punishment of a Crime: Because it is of the nature of Punishment, to have for end, the disposing of men to obey the Law; which end (if it be lesse than the benefit of the transgression) it attaineth not, but worketh a contrary effect.\(^{46}\)

I will use the term *punishment* in Hobbes’ strict sense in the following.\(^{47}\) The question that arises here is: How can the sovereign best ensure that the Foole is actually a fool?

In the real world, it is not guaranteed that the sovereign, together with the help of prudent citizens, will detect all instances of free riding. Further, installing an enforcement mechanism that aims to detect all free riders is not necessarily desirable because of high social and economic costs.\(^{48}\) However, if an individual is sure that she can free ride without being detected, then she is not a fool at all when she free rides. Instead, she effectively maximizes her expected utility.

Because there is no guarantee that free riders will always be detected in practice, and the sovereign can neither judge with certainty how much free riding can be tolerated without threatening peaceful cooperation and without undermining the citizens’ trust in her, nor judge what is needed to ‘wake up’ potential Fooles from their shortsightedness, the sovereign must render free riding as unattractive as possible. The sovereign must do her best to deter non-cooperative behavior by, first, instituting a police system that ensures a reasonably high apprehension and

\(^{46}\) Hobbes, *Leviathan*, Part II, Chapter XXVIII.


\(^{48}\) For further conceptual and empirical considerations with regard to this point, see Gregory Kavka, *Hobbesian Moral and Political Theory*, pp. 250-254.
conviction rate, assuming that an enforcement system that aims to detect all free riders is impossible and socially and economically undesirable.49

Second, given the limitations of the first method, the sovereign must threaten the citizens with severe sanctions for free riding if they are detected. As Hobbes suggests, if an individual defects after society is established and her non-cooperative behavior is detected, the sovereign must punish the individual and, in the worst case, exclude her from society or even take her life:

He therefore that breaketh his Covenant, and consequently declareth that he thinks he may with reason do so, cannot be received into any Society, that unite themselves for Peace and Defence [...].

And again, he that having sufficient Security, that others shall observe the same Lawes towards him, observes them not himselfe, seeketh not Peace, but War; & consequently the destruction of his Nature by Violence.50

Given these severe sanctions, an individual is worse off compared to her situation in the state of nature if she free rides and is detected after society is established. Moreover, if the sovereign punishes the Foole for each instance of free riding, the Foole cannot expect to get away with such behavior. Instead, she can be sure that if she free rides and is detected, she will end up with a suboptimal outcome.

In game-theoretic terms, the sovereign plays a grim strategy with the Foole. The sovereign initially cooperates with all members of society after they have agreed to follow the civil laws that she imposes on them in order to secure peace. However, if an individual defects once society is established by violating the civil laws, the sovereign will not cooperate again with the individual by punishing her proportional to the severity of her law violation, and in the most extreme case by taking her life, if the individual’s behavior endangers the internal peace of society according to the sovereign’s judgment.51

Note that it is sometimes argued that the individuals play a grim strategy with one another in the state of nature and not the sovereign with the citizens once society is established.52 I am skeptical about this interpretation because under the fierce conditions described by Hobbes, an individual cannot expect to interact repeatedly with her fellows in the state of nature, and she will not move first in the game of cooperation that establishes society because she thereby risks losing her life. The possibility of retaliation in the next round of the game, as expressed by the grim strategy, cannot solve the problem of assurance in Hobbes’ state of nature

---

49 I assume that the costs for installing and maintaining an optimal enforcement mechanism in this sense are lower than the gains that can be made by the formation of society.
50 Hobbes, *Leviathan*, Part I, Chapter XV.
51 Such punishment is legitimate only if an individual possesses the ability to reason, if she acted voluntarily, and if the civil laws clearly state the prohibited actions along with the associated sanctions. For a further discussion of Hobbes’ theory of punishment, see in particular Dieter Hüning, ‘Hobbes on the Right to Punish’ in The Cambridge Companion to Hobbes’s *Leviathan*, pp. 217-240, pp. 221-229.
52 See Brian Skyrms, ‘The Shadow of the Future’, pp. 12-21, for example.
because it may simply come too late. As such, if we accept Hobbes’ description of the state of nature, as I do in this article, it is implausible to argue that the individuals play a grim strategy with one another in the state of nature.

The game-theoretic interpretation of Hobbes’ response to the Foole after society is established reveals an important feature of Hobbes’ argument, because only the grim strategy can bind rational individuals always to follow the social rules. To clarify, an instrumentally rational individual cannot simply choose to follow a rule, because she has no commitment power. Instead, a rational individual decides in each instance which action is most beneficial for her. As such, in order for a rational individual to strictly follow a rule, rule-guided behavior must pay off for her in each case for which the rule prescribes certain behavior. Or, vice versa, a rational individual ‘follows’ a rule only if the behavior that is prescribed by the rule coincides with the behavior that she expects to be most beneficial for her in the given situation.

If we consider the game of securing peaceful cooperation after society is established as a repeated game where rational individuals can decide in each round of the game what to do, and we assume that rational individuals expect that following the social rules is more beneficial for them in the long run than staying in the state of nature, then an individual can improve her situation, compared to acting in a strictly cooperative manner, only if she can defect from time to time, that is, in the short run, without being punished in Hobbes’ strict sense. If an individual believes that she can do so, then instrumental rationality forces her to defect. In this case, the sovereign cannot prevent free riding.

However, if an individual is not absolutely sure that her non-cooperative behavior will remain undetected and she knows that she will always be punished for such behavior if she is detected, as demanded by the grim strategy, then the Foole must carefully assess whether breaking the social rules is the utility maximizing choice for her. In this case, if the risk of being detected, now or in the future, is sufficiently high, it is irrational to free ride, even for the shortsighted Foole. In this sense, the grim strategy binds rational individuals to follow social rules as closely as possible in a world where not all free riders are necessarily detected. As such, instituting a police system that ensures a reasonably high apprehension and conviction rate and playing a grim strategy with all free riders who are detected minimizes the citizens’ risk of being harmed by their fellows in the post-natural state.

Nevertheless, one may question whether it is justified for the sovereign to exclude the Foole from the cooperative framework for defective behavior and to take her life in the most extreme case, because it may be overall beneficial for society to accept the Foole as a cooperative partner in certain cases, for instance, if the Foole can provide social and economic goods that other individuals cannot.

---

53 I think that one can give a coherent reading of Hobbes’ argument in *Leviathan*, including his reply to the Foole, without introducing the notion of commitment power.
offer. In this case, it may be more beneficial for society, and for each citizen, to risk certain vulnerability instead of entirely excluding the Foole.

For a similar reason, one may argue that although the sovereign may have good reasons to threaten the Foole by playing a grim strategy with her in order to deter free riding as much as possible, in practice the sovereign could be more lenient with free riders as long as such leniency does not undermine her credibility to punish rule violators, because the Foole must only expect to be punished for non-cooperative behavior. As such, if the exemptions made do not significantly lower the Foole’s expectation of being punished, the sovereign can be more forgiving with free riders without risking peaceful cooperation. In light of these considerations, how is the exclusion of the Foole from society in severe cases of defective behavior justified?

According to Hobbes, the task of the sovereign is to guarantee peaceful cooperation, and thus to move the individuals permanently out of the state of nature. This demand becomes clear from Hobbes’ defense of monarchy as the best form of government. Although Hobbes seriously considers other forms of government, he is skeptical that they can guarantee internal peace. Hobbes believes that more democratic forms of government may be able to secure peaceful cooperation under favorable conditions, but this is not sufficient for him. Hobbes demands to institute a government that can secure peace in all circumstances, that is, with near certainty. This demand implies, in turn, that once the sovereign is in place, she must exclude any risks that may jeopardize peaceful cooperation.

Now, although the sovereign has unlimited power over the citizens, she is restricted in her actions by reason, and thus by the laws of nature.⁵⁴ As such, the civil laws made and enforced by the sovereign must be the necessary means to secure peace, according to the sovereign’s judgment. The civil laws must specify the restrictions on the behavior of the citizens that are necessary to ensure peaceful cooperation. As a consequence, if an individual breaks the social rules imposed by the sovereign, the individual directly endangers peace, from the perspective of the sovereign.

To be realistic, not all instances of defective behavior have an immediate negative impact on peaceful cooperation, if they do not directly threaten the lives of other citizens. In particular, many everyday law violations do not pose a direct risk to peace. However, due to incomplete information the sovereign cannot know with certainty when peaceful cooperation is at risk and when the erosion of the commonwealth begins. As a consequence, if the sovereign’s aim is to guarantee peaceful cooperation by avoiding any foreseeable risk to the citizens’ security, she must exclude the Foole from society in severe cases of defective behavior, because the Foole has already shown that she jeopardizes the primary objective of society, which is to secure internal peace.

⁵⁴ See Hobbes, Leviathan, Part II, Chapter XXVIII.
Stated differently, the Foole is irrational from the perspective of the sovereign, because the sovereign has done everything in her power to transform the one-shot PD game that the shortsighted Foole believes she is playing in the post-natural state, with regard to her payoff structure of the game, (at least) into an assurance game by playing a grim strategy with her and by threatening to take her life in the most extreme case. If the Foole still free rides, although she must realize now that such behavior will make her worse off, she is irrational from the perspective of the sovereign and cannot be trusted.

As a consequence, as long as the Foole does not indicate that her reasoning ability has permanently changed, and thus that she is no longer a Foole, the sovereign must expect that the Foole will defect again from time to time. The Foole poses a permanent risk for peaceful cooperation from the perspective of the sovereign and must be excluded from society in severe cases of non-cooperative behavior, if the sovereign’s aim is to guarantee peace with near certainty. If the sovereign does not exclude the Foole from society in severe cases of defective behavior, she acts negligently.

In conclusion, the grim strategy and the exclusion of the Foole from society in severe cases of defective behavior can be justified from the perspective of the sovereign, assuming that the sovereign’s task is to minimize the risk of falling back into the state of nature, because the grim strategy renders strict cooperation as the uniquely best strategy for rational individuals, if the chance of being detected is sufficiently high. This is not to say, however, that the grim strategy necessarily leads to the best possible outcome for each citizen and for society overall. Some citizens, and even society as a whole, may be better off if free riding is tolerated to a certain extent, because the sovereign may, for example, sometimes err in specifying the civil laws that are necessary to ensure peace. The grim strategy enforces only one possible equilibrium outcome. However, it is the strategy that renders free riding as irrational as possible in a world where rational individuals cannot be sure that their defective behavior will not be detected. It is the strategy that has the lowest risk of falling back into the state of nature, and thus has the highest chance of maintaining peace.55

55 I want to thank Richard Bradley, Jerry Gaus, Alex Voorhoeve, and two anonymous reviewers of this journal for helpful discussions of the material addressed in this article and comments on earlier versions.