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

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. 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.