Hobbes, Harsanyi and the Edge of the Abyss*

GABRIELLA SLOMP *University of Strathclyde*
MANFREDI M. A. LA MANNA *University of St. Andrews*

Introduction

"Out of it [the political state], there is a Dominion of Passions. . . . In it, the dominion of reason."[1] So writes Hobbes in *De Cive*. In view of this remark, it is surprising that the state of nature, that Hobbes suggests to be the dominion of passions, should have been the subject of so many applications of *rational* choice theory. One superficial explanation would be to interpret the above quotation as referring implicitly to the paradox of the Prisoner's Dilemma: individually rational actions lead to an outcome that is less desirable than a feasible alternative. In this article we put forward a different and more radical explanation. We suggest that Hobbes's claim that the state of nature is the realm of irrationality ought to be taken seriously, for it implies that rational-choice arguments cannot be deployed to analyze the behaviour of individual agents in pre-political conditions. We use the very language of rational-choice theory subversively to show that the Hobbesian state of nature can be formalized as a state of affairs in which *individual rational* decision making is altogether impossible.

* We wish to thank the JOURNAL's anonymous referees for their comments. Blame for any errors and omissions must be apportioned evenly to both authors. An earlier version of the article was presented at a meeting of the Rational Choice Theory Study Group, London School of Economics, February 6, 1994. Manfredi M. A. La Manna thanks the Department of Economics at the University of Western Ontario, London, Ontario, for hospitality in the summer term 1993.

1 Thomas Hobbes, *De Cive*, The Clarendon Edition of the Philosophical Works of Thomas Hobbes, Vol. 2, ed. by H. Warrender (Oxford: Clarendon Press, 1983), 130.

Gabriella Slomp, Department of Government, University of Strathclyde, Strathclyde, Scotland, United Kingdom. E-mail: g.slomp@strath.ac.uk
Manfredi M. A. La Manna, Department of Economics, University of St. Andrews, St. Andrews, Scotland, United Kingdom. E-mail: mlm@st-andrews.ac.uk

Canadian Journal of Political Science / Revue canadienne de science politique, XXIX:1 (March/mars 1996). Printed in Canada / Imprimé au Canada

In the context of the recent debate between rational-choice interpreters of Hobbes (for example, Kavka, Hampton, Taylor, McLean[2]) and proponents of the rhetorical-historical interpretation (Neal,[3] among others), we argue that the latter are open to the charge of underestimating the heuristic possibilities of rational-choice theory[4] by identifying it with a narrow range of game-theoretic constructs. On the other hand, we challenge the basic assumption underlying all existing applications of game theory to Hobbes, namely, that the Hobbesian state of nature can be explained as the outcome of *individually rational* actions.

Our game shares with other rational-choice accounts of Hobbes's state of nature the view that rational agents are unable to escape from anarchy. However, unlike some Hobbesian critics (for example, Hampton[5]), we do not interpret this result as an implicit criticism of Hobbes. On the contrary, we see Hobbes's argument as posing a problem to the rational-agent approach, which under some circumstances may fail to suggest *any rational course of action* to individuals. Indeed, we believe that our interpretation of the state of nature lends support to the view that Hobbes's political theory aims not to provide a rational plan for *individual* behaviour in the state of nature, but instead to derive rationally a message for people already living in political associations so as to help them (1) to grasp the irrationality of civil war[6] and (2) to design the conditions for "immortall peace." We also believe that our interpretation provides a satisfactory explanation for a fundamental feature of the Hobbesian state

2 Gregory Kavka, *Hobbesian Moral and Political Theory* (Princeton: Princeton University Press, 1986); Jean Hampton, *Hobbes and the Social Contract Tradition* (Cambridge: Cambridge University Press, 1986); Michael Taylor, *The Possibility of Cooperation* (Cambridge: Cambridge University Press, 1987); and Iain McLean, "The Social Contract in Leviathan and the Prisoner's Dilemma Supergame," *Political Studies* 29 (1981), 339-51.

3 Patrick Neal, "Hobbes and Rational Choice Theory," *Western Political Quarterly* 41 (1988), 635-52.

4 Neal argues that "rational choice theory reaps a good less than Hobbes attempted to sow and serves to obscure more than illuminate his teaching" (ibid., 635).

5 Hampton, *Hobbes and the Social Contract Tradition*.

6 Hobbes compares the state of nature to a state of civil war: "it may be perceived what matter of life there would be, where there were no common power to fear, by the manner of life, which men that have formerly lived under a peaceful government, use to degenerate into, in a civil war" (Thomas Hobbes, *Leviathan*, Vol. 3 of *The English Work of Thomas Hobbes*, ed. by W. Molesworth [London: Scientia Aalen, 1962], 114-15). We should stress that our claim to novelty refers to the way in which we use rational-choice theory to show the irrationality of descent into the state of nature. This reading of Hobbes's theory goes back to some of his contemporaries and two recent examples are Russell Hardin, "Hobbesian Political Order," *Political Theory* 19 (1991), 156-80, and Iain Hampsher-Monk, *A History of Modern Political Thought* (Oxford: Blackwell, 1992), who calls it the "virtual contract" theory.

Abstract. The authors present a new game-theoretic interpretation of Hobbes's state of nature that, unlike existing rational-choice models, questions the *possibility of individually rational decision making*. They provide a general formulation of the two-player two-strategy game applied to the state of nature and derive existing models as special cases. A nonstandard version of Chicken under incomplete information, that interprets "death" as infinitely bad, provides an explanation for important and hitherto unaccounted for claims by Hobbes. The authors suggest that rational choice in Hobbesian political philosophy ought to examine not so much the mechanics of rational *action* in natural conditions, but rather the means whereby citizens already living in civil associations can be persuaded of the irrationality of civil war.

Résumé. Les auteurs proposent une nouvelle interprétation du concept hobbésien d'état de nature. Leur approche, reprise de la théorie des jeux, remet en cause le principe de la rationalité individuelle dans la prise de décision, pourtant admis traditionnellement par les modèles du choix rationnel. Ils donnent une définition générale du jeu à deux joueurs et à deux stratégies dans l'état de nature dont les modèles existants s'avèrent être des cas particuliers. Ils analysent d'une manière originale certains des principaux arguments de Hobbes à l'aide d'une variante du jeu du « Chicken » en environnement incertain, où la « mort » est considérée infiniment mauvaise. Ils suggérent que l'analyse des choix politiques rationnels dans une perspective hobbésienne se doit avant tout de considérer les moyens par lesquels des citoyens vivant déjà dans des associations politiques peuvent être convaincus de l'irrationalité d'une guerre civile, plutôt que de décrire le mécanisme qui conduit à une action rationnelle dans l'état de nature.

of nature that cannot be accounted for by any rational-choice interpretation, namely, that war is "perpetual in its own nature."[7]

In section 1 we examine the assumptions required to analyze the Hobbesian state of nature as a two-player two-strategy non-cooperative game. In section 2 we show how each of the possible equilibria of this game is related to the many game-theoretic interpretations available in the literature. In section 3 we argue that the notion of *mixed strategy*, especially in the version provided by Harsanyi[8] finds a natural application in Hobbes's state of nature and has been surprisingly ignored by his rational-choice interpreters. In section 4 we argue that in Hobbes's state of nature the specific *value* (and not just the mere ranking) of the payoff associated with violent death alters completely the nature of an otherwise standard game, with far-reaching consequences for the feasibility of individual rational choice. Section 5 draws some conclusions on the general implications of our analysis.

1. The State of Nature as a Two-Player Two-Strategy Game

The purpose of this section is to highlight the assumptions required to describe the Hobbesian state of nature as a two-strategy two-player non-cooperative game.[9]

7 Hobbes, *De Cive*, 49.

8 John C. Harsanyi, "Games with Randomly Disturbed Payoffs: A New Rationale for Mixed Strategy Equilibrium Points," *International Journal of Game Theory* 2 (1973), 1-23.

9 The substantive qualitative points we establish in section 4 are unaffected by extending the analysis to N (>2) players. The same does not apply to some of the

IR: *I*nstrumental *R*ationality: Hobbes states that "every man by reasoning seeks out the meanes to the end which he propounds to himselfe."[10]

CK: *C*ommon *K*nowledge: that is, the description of the game is known to both players and each player knows that the other player knows, and so on.[11] Hobbes's reflections on the right of nature and the natural laws come very close to assuming common knowledge explicitly.[12]

B: *B*ipartition: relevant actions by the players can be partitioned into two disjoint sets, namely, actions that lead to conflict and the submission of others ("Fight")—attacking others, dispossessing them, provoking them "by deeds or words"—and actions intended to avoid conflict ("Avoid"). In his description of the state of nature Hobbes gives an account of human behaviour in terms of this bipartition.

E: *E*qual Vulnerability and Dangerousness: all individuals inhabit a very "fragile human frame" so that "even the weakest has strength enough to kill the strongest."[13]

S: *S*elf-preservation: each individual seeks to avoid violent death at the hand of others.

Figure 1 describes a standard bimatrix game:

FIGURE 1

THE GENERIC BIMATRIX GAME

		Column Player	
		Avoid	*Fight*
Row Player	*Avoid*	P, P'	S, D'
	Fight	D, S'	W, W'

games examined in section 2, which rely on interpreting Hobbes's model as a *compound* two-player game.

10 See, for example, Hobbes, *De Cive*, 177; see also Hobbes, *Leviathan*, chap. 5, and Thomas Hobbes, *The Elements of Law Natural and Politic*, ed. by F. Tönnies (2nd ed.; London: F. Cass, 1969), chap. 1.

11 For a formal definition of common knowledge, see Robert Aumann, "Agreeing to Disagree," *Annals of Statistics* 4 (1976), 1236-39; see also Robert Sugden, "Rational Choice: A Survey of Contributions from Economics and Philosophy," *Economic Journal* 101 (1991), 751-85. Assumptions *IR* and *CK* are part of what Neal (in "Hobbes and Rational Choice Theory") calls "E-rationality."

12 According to Hobbes, the content of the right of nature and of the natural laws can be understood by everybody, and everybody can be assumed to understand it (see Hobbes, *Leviathan*, chap. 5, especially 144).

13 Hobbes, *Leviathan*, 110; Hobbes, *Elements of Law*, 70; and Hobbes, *De Cive*, 45.

Each player can choose either to Avoid fighting or to Fight, with the ordered pair of strategies (Row, Column) determining the payoff of each player. If Row player chooses Avoid and Column player Fight, the former obtains the payoff S (where S stands for "Subjection") and the latter the payoff D' (D' for "Dominion"). If both players fight, Row (Column) player has a payoff of W (W')—where W stands for War; whereas if they both avoid fighting, Row (Column) player enjoys a payoff of P (P')—P for "Peace."

The outcome of the game depends on the relative rankings of P vis-à-vis D and S vis-à-vis W for Row player (P' vis-à-vis D' and S' vis-à-vis W' for Column player).

The assumptions listed above imply that for both players mutual avoidance is strictly preferable to unilateral avoidance ($P > S$ and $P' > S'$) and that mutual fighting is strictly inferior to mutual avoidance ($P > W$ and $P' > W'$). This view is expressed most clearly by Hobbes in chapter 13 of *Leviathan*.

What is the ranking of all the other payoffs (namely, "Peace" versus "Dominion" and "Subjection" versus "War") that best accounts for Hobbes's description of the state of nature? It is here that rational-choice interpreters disagree, each pointing out different characteristics of Hobbesian people and buttressing their views with opposing textual references. It is a testimony to the inventiveness of rational-choice theorists that virtually all the possible rankings of the relevant payoffs have been traced back to a suitably selective reading of Hobbes's works.

An analysis of the eight[14] possible rankings of the relevant payoffs is relegated to Appendix A. Here we consider as an example one of the possible outcomes so as to introduce some notation and concepts that will turn out to be useful in what follows.

Consider the case where both players prefer Dominion to Peace and Subjection to War: $P < D$ and $W < S$ and $P' < D'$ and $W' < S'$. Obviously neither player has a dominant (weakly dominant) strategy, that is, a strategy that would be strictly (weakly) preferred irrespective of the other player's choice. If one player fights the other would prefer to avoid fighting and vice versa.

Suppose that Row player and Column player respectively avoid fighting with probability ρ and κ. We can then introduce the notions of a "probability box" and of a *best-response* line. The latter is the graphical answer to the question: "for any given probability that my opponent avoids fighting, with which probability should I avoid fighting so as to achieve the highest expected payoff?" Suppose that Column player fights with certainty ($\kappa = 0$), then Row player is better off by avoiding fighting: the best response to $\kappa = 0$ is $\rho = 1$. Conversely, if Col-

14 We ignore the trivial case of total indifference where both $P = D$ and $S = W$.

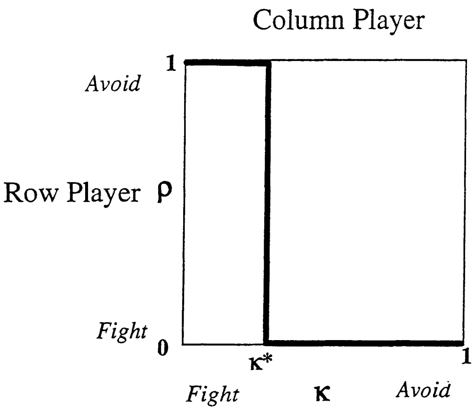
umn player refrains from fighting, $\kappa = 1$, then Row player will fight with certainty, $\rho = 0$. It follows that there must exist a probability that Column player avoids fighting, say κ^*, such that Row player is indifferent between Avoid and Fight (that player attains the same expected payoff, $E\pi$, that is, the same sum of payoffs multiplied by their respective probabilities):

$$E\pi\,(\text{Avoid}) = \kappa^* P + (1 - \kappa^*)\, S \equiv \kappa^* D + (1 - \kappa^*)\, W = E\pi\,(\text{Fight}).$$

Thus we can summarize Row player's behaviour by that player's best-response line in Figure 2.

FIGURE 2

BEST-RESPONSE LINE FOR ROW PLAYER



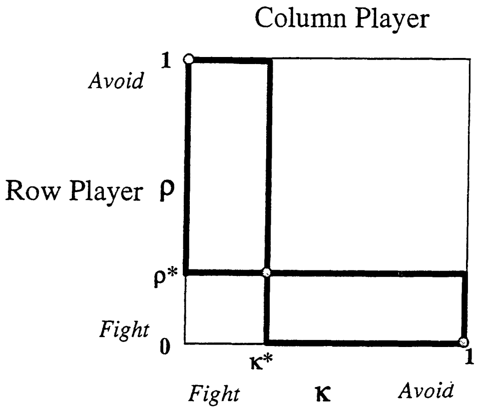
Readers who are uncomfortable with the idea that rational behaviour may call for playing dice with choices of life and death will be reassured to know that the interpretation of mixed strategies offered in section 3 does not rely on players choosing their actions according to the outcome of some randomizing device. If we perform precisely the same experiment with Column player and derive that player's best-response line, then we can draw the two players' best-response lines in the same probability box and ascertain where they intersect (or overlap); see Figure 3. At the intersection *both* players are replying optimally to each other's actions.

The mild requirement that the equilibrium (or equilibria) of the game be where the two best-response lines intersect (or coincide) is, of course, the idea underlying the notion of *Nash equilibrium.*[15]

15 On the relationship between the *CK* assumption and Nash equilibrium, see M. Bacharach, "A Theory of Rational Decision in Games," *Erkenntnis* 27 (1987), 17-55, and Sugden, "Rational Choice."

FIGURE 3

BEST-RESPONSE LINES FOR ROW AND COLUMN PLAYERS



Before turning to the application of the two-player two-strategy game to Hobbes's state of nature, we have to introduce one final assumption:

*FR: F*inite *R*epetition: we assume that the above game is played (and is expected to be played) for a *finite* number of times and not, as assumed in much of the literature on game-theoretic applications to Hobbes,[16] for an *infinite* number of times. The attraction of formulating Hobbes's state of nature as an infinitely repeated game (supergame) is that it opens up the possibility of introducing the notion of retaliation, which turns the mutual avoidance of fighting into a sustainable outcome. We do not believe that the attainment of "good" equilibrium points is open to Hobbesian individuals for the following simple reason: there is strong textual and contextual evidence in support of the view that Hobbesian individuals do not engage in murderous cruelty when they engage in fighting,[17] but aim specifically at the physical submission of their enemies. In the state of nature individuals "use Violence, to make themselves *Masters of other mens persons, wives, childrens, and cattell.*"[18] This implies, of course, that being attacked while avoiding battle becomes an *absorbing* state[19] for the "subdued" player, rendering any retaliation physically unfeasible.

Armed with this simple game-theoretic framework, we can now turn to the analysis of rational-choice interpretations of Hobbes's state of nature.

16 Taylor, *The Possibility of Cooperation*, and McLean, "The Social Contract," are two *loci classici.*

17 See, for example, Hobbes, *Leviathan,* chap. 6, at 44.

18 Ibid., chap. 13, at 88, emphasis added.

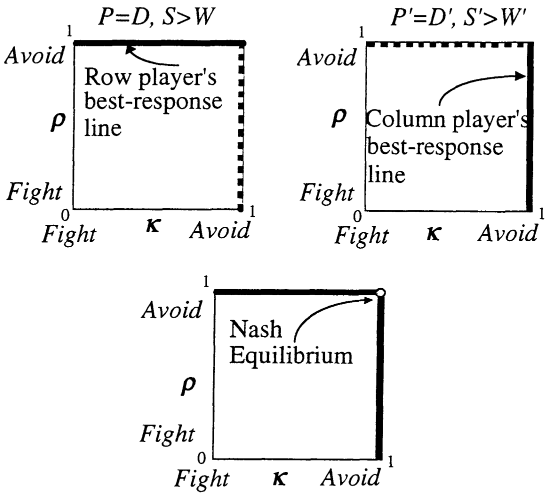
19 That is, a state that, once entered into, cannot be escaped from.

2. Game-Theory Applications to Hobbes's State of Nature

As mentioned above, the assumptions on human nature shared by Hobbes and his rational-choice interpreters are not sufficient to narrow down the range of possible equilibria. All eight possible permutations of the relevant payoffs can be traced to Hobbes's account of the state of nature and, indeed, all the significant cases, namely, Prisoner's Dilemmas, coordination, assurance and Chicken games, have found their champions among Hobbes's rational-choice interpreters.

FIGURE 4

THE BEES AND ANTS GAME[20]



2.1. The Bees and Ants Game

Of the eight payoff rankings listed in Appendix A, only one has not been examined in the literature on Hobbes, namely, the case $P = D$ and $S > W$ (Case D2 in Appendix A). The reason for this omission is obvious. These payoffs, far from leading to the war that characterizes the Hobbesian state of nature, produce the blissful state of peace enjoyed by bees and ants and described in parallel passages of *Leviathan, Elements of Law* and *De Cive*.[21] The key equality $P = D$ is the implication

20 As explained in Appendix A, the dotted lines in Figures 4 and 6 refer to a weakly dominated strategy and can be ignored.

21 "[B]ees and ants live sociably one with another (which are therefore by *Aristotle* numbred amongst Politicall creatures), and yet have no other direction, than their particular judgements and appetites; nor speech, whereby one of them can signifie to another, what he thinks expedient for the common benefit: and therefore some man may perhaps desire to know why mankind cannot do the same" (Hobbes, *Leviathan*, 156; see also Hobbes, *Elements of Law*, 102, and Hobbes, *De Cive*, 87).

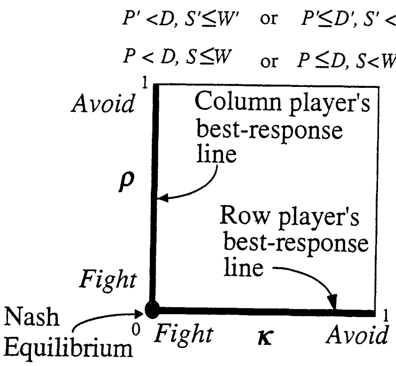
of *non-glory-seeking behaviour*: unlike human beings preoccupied with establishing their "eminence," bees and ants do not derive any benefit from fighting an opponent who refrains from fighting. On the other hand, bees and ants, not unlike humans, are concerned with self-preservation and thus, when confronted by an attacking opponent, prefer subjection to war.

2.2. *Prisoner's Dilemmas*

The payoff rankings $D > P > W \geq S$ (cases A1 and A2) and $D = P > W > S$ (case C2) give rise to the much-studied Prisoner's Dilemma analogue of the Hobbesian state of nature[22] as described in Figure 5.

FIGURE 5

PRISONER'S DILEMMAS



2.3. *Neal's "Coordination" Game*

The payoff rankings $P > D, S = W$ (case B2) have been suggested as an alternative interpretation of Hobbes's state of nature by Neal. He makes the important point that the standard interpretation of the state of nature as a Prisoner's Dilemma does not take into consideration "the substantive payoffs available to human beings in the state of nature."[23] He argues that all strategy combinations other than mutual cooperation

22 The first applications of the Prisoner's Dilemma to Hobbes's theory can be found in David Gauthier, *The Logic of Leviathan* (Oxford: Clarendon Press, 1969), and John Watkins, "Imperfect Rationality," in R. Borger and F. Cioffi, eds., *Explanation in the Behavioural Sciences* (Cambridge: Cambridge University Press, 1970). In the literature on *PD* applications to Hobbes's theory, usually only strict inequalities are considered; however, the same substantive outcome obtains even with weak inequalities.

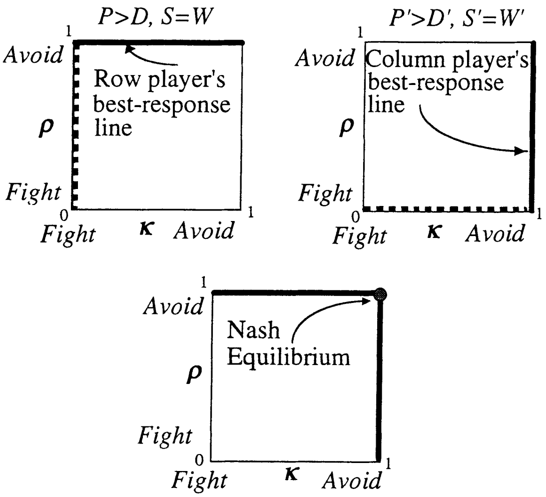
23 Neal, "Hobbes and Rational Choice Theory," 642.

lead to individual payoffs that make individuals as badly off as they can possibly be. More specifically he posits that $D = S = W =$ "Death."

As we shall show in section 4, the insight of considering the specific payoff of Death is crucial to an understanding of Hobbes's theory. However, Neal's statement that the resulting game is one of "coordination" is questionable. If the weakly dominated strategy of Fighting is eliminated, mutual avoidance is the *unique* equilibrium of the game, as shown in Figure 6.

FIGURE 6

MUTUAL AVOIDANCE IN NEAL'S GAME



The same equilibrium would obtain, of course, if Fighting were a strongly dominated strategy (case B1: $P > D$, $S > W$) as would be the case, for example, for the "moderate" (or "temperate") individual mentioned by Hobbes in all his political works.

2.4. Assurance Games

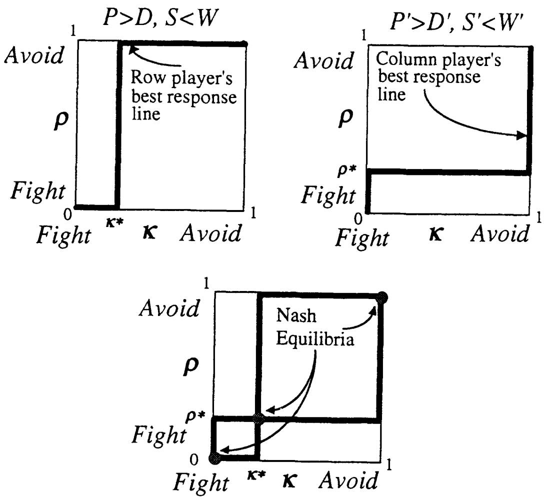
The payoff rankings $P > D$, $S < W$ (case C1) give rise to *assurance games* that figure prominently in the literature on Hobbesian political theory. For example, Kavka, Hampton and Taylor[24] all use games with the above payoffs. The point we want to stress is that coordination games of the type examined here have *three* Nash equilibria, two symmetric equilibria in pure strategies (where both players either fight or avoid with

24 Kavka, *Hobbesian Moral and Political Theory*, especially 182-88; Hampton, *Hobbes and the Social Contract Tradition*, 150-60; and Taylor, *The Possibility of Cooperation*, chaps. 2 and 3.

certainty) and one mixed-strategy equilibrium (where Column [Row] player avoids with probability κ* [ρ*]), as depicted in Figure 7.

FIGURE 7

THE ASSURANCE GAME



2.5. Fowl Games: Chickens, Hawks and Doves

Hobbesian interpreters have somewhat neglected[25] what seems to us the most accurate game-theoretic account of Hobbes's state of nature, namely, a *dis-coordination* game. The payoff rankings of this game, $P < D$, $S > W$ (case D1), will be discussed in section 4, where we put forward our own version of Chicken. As a first step, we have to clarify the notion of *mixed strategy* and its relevance for an explanation of Hobbes's state of nature.

3. Mixed Strategies, Uncertainty and Hobbes

It has been often pointed out that Hobbes's state of nature is marked by uncertainty: "Where there is no commonwealth there is . . . neither property nor community, but *uncertainty*."[26] In particular, he main-

25 Exceptions are McLean, "The Social Contract," and Robert Sugden, *The Economics of Rights, Co-operation and Welfare* (Oxford: Blackwell, 1986), who apply the hawk-dove game originally developed by John Maynard Smith and G. R. Price (in "The Logic of Animal Conflict," *Nature* 246 [1973], 15-18) to Hobbes's state of nature.

26 Hobbes, *Leviathan*, chap. 24, at 233, emphasis added; see also chap. 13.

tains that in natural conditions we cannot form expectations about the behaviour of others: "for though the wicked were fewer than the righteous, yet because *we cannot distinguish them*, there is a necessity of suspecting . . . ever incident to the most honest."[27] Moreover, throughout his works Hobbes repeatedly points out that preferences vary not only across individuals,[28] but also for the same person over time.[29] In view of the above quotations, we believe that the simplest way of introducing uncertainty in a Hobbesian game is to interpret it as a game of *incomplete information*, in the sense that each player, while knowing their own preferences, is uncertain about the precise payoffs that all the other players attach to each outcome.[30]

The connection between mixed-strategy equilibria of the games of complete information sketched in sections 1 and 2 and the pure-strategy equilibrium in games of incomplete information was established over 20 years ago by John Harsanyi.[31]

Harsanyi's ingenious solution was to relax the very strong assumption of complete information about the payoff matrix, and to introduce the idea of randomly fluctuating payoffs. This means, for example, that a typical Hobbesian individual, in deciding whether to fight or avoid battle with an opponent, has not the luxury of knowing for sure if, and how strongly, his enemy prefers Peace to Dominion and Subjection to War. The resulting game of *incomplete information* will have all its equilibrium points in pure strategies (no player will randomize). In other words, players choose their *pure* strategy on the basis of values of the payoffs that are known only to them. *Which* pure strategy is chosen depends on the player's beliefs regarding the (unknown) payoffs of the other player. Moreover, this game of incomplete information has the property that, as the amount of payoff uncertainty approaches zero, the players will use their pure strategies with the same probabilities specified in the mixed-strategy equilibrium of the corresponding game with complete information.

Especially in the Hobbesian state of nature, where individuals "cannot distinguish the wicked from the righteous," it seems to us that

27 Hobbes, *De Cive*, 33, emphasis added.

28 Hobbes, *Elements of Law*, 29; Hobbes, *De Cive*, 74-75, 177; and Hobbes, *Leviathan*, 28-29, 146.

29 "All men [are] not alike affected with the same thing, nor the same man at all times" (Hobbes, *Leviathan*, 146; see also Hobbes, *De Cive*, 74, 177).

30 In other words, player i knows their own *type* t_i and has some beliefs regarding the other players' types, embodied in a probability distribution $p_i(t_{-i})$, where $t_{-i} \equiv (t_1, \ldots, t_{i-1}, t_{i+1}, \ldots, t_N)$ are the types of all other players. More generally one would write $p_i(t_{-i} | t_i)$ to allow for the possibility that player i's belief regarding the other players' types can depend on their own type, t_i. However, in the state of nature as described by Hobbes, it is more likely that players' types are independent.

31 Harsanyi, "Games with Randomly Disturbed Payoffs."

the Harsanyi transformation provides a solid rationale for the interpretation of mixed strategies as a metaphor for the use of pure strategies when payoffs fluctuate randomly.[32]

In view of this remarkable correspondence between the conditions under which mixed strategies can be interpreted as a response to uncertainty concerning the psychological make-up of potential opponents and Hobbes's description of the state of nature as the realm of uncertainty, we are baffled by the way in which many distinguished rational-choice interpreters of Hobbes have treated mixed-strategy equilibria. To mention but a few, Kavka does not examine them, Hampton treats them cursorily and Taylor dismisses them in a footnote.[33]

There are two significant problems with the use of mixed strategies re-interpreted in the manner of Harsanyi. Given that each player's action depends on their beliefs regarding the probability of their opponent being of a certain "type" (for example, more or less intensely glory seeking), the opponent has the incentive to manipulate information about this probability. Moreover, the assumption that all players know the "true" probabilities may be unrealistic, since these beliefs are "social constructs" rather than pre-given.

While valid in general, we would argue that this line of criticism does not apply to Hobbes's state of nature. Undoubtedly, Hobbesian individuals have a clear incentive to mislead others about the probability distribution from which the random component of their payoff is drawn. In forensic terminology, they definitely have *motive*. However, they lack *opportunity* in that in Hobbes's state of nature there are no means of effective communication, no shared language: "private judgement may differ, and beget controversy . . . of what is to be called much, what little . . . what a pound, what a quart."[34] So even though Hobbesian individuals have a clear incentive to dissemble on their own "type," the lack of any common standards against which dissembling can be defined prevents them from deceiving others. This effectively removes the opportunity for any information manipulation and makes the assumption of shared beliefs an innocuous one. The other criticism to the Harsanyi transformation refers to the fact that in the state of nature no social process is available that may generate "commonness" of knowledge. In Hobbes's construct, commonness is not generated by a social process, but by what can be called "reciprocal rationality" (the intersection of assumptions *IR* and *CK*). All individuals know that all others are as efficient in their reasoning capacities as they are, that is,

32 See ibid. for a formal statement and proofs.

33 Kavka, *Hobbesian Moral and Political Theory*; Hampton, *Hobbes and the Social Contract Tradition*; and Taylor, *The Possibility of Cooperation.*

34 Hobbes, *Elements of Law,* 188.

they know the probability distribution of the random component of their opponents' payoffs.

4. Hobbes's State of Nature as the Dominion of Passions

In this section, we will provide our own version of Chicken as applied to the Hobbesian state of nature. In spite of the differences between *Elements of Law, De Cive* and *Leviathan*,[35] our proposed interpretation is based on a set of assumptions common to all three works. The most immediate advantage of our interpretation is that it can account for some important claims made by Hobbes on the state of nature that cannot be explained by other game-theoretic constructs. These include the claim that the state of nature is the "dominion of passions," that war is "perpetual in its own nature" in spite of individuals' equality to kill, that "the nature of war, consisteth not in actual fighting" and that by wishing to remain in the state of nature a rational man "contradicteth himself."

4.1. The Assumptions of the Game

In addition to the standard assumptions *IR, CK, B, FR* and *E* introduced in section 1, our game is predicated on the following assumptions:

II: *I*ncomplete *I*nformation: individuals "cannot distinguish the wicked from the righteous," the glory seeker from the moderate, the intensely from the mildly glory-seeking individual.

$S\infty$: *S*elf-preservation in a strong sense: "violent death at the hands of others" is *infinitely* bad, namely, "of the good things experienced by men *none can outweigh* the greatest of the evil ones, namely, sudden death."[36] Violent death is not simply the worst *ranked* outcome, for in this case it would be possible to attach to it a low enough probability that other desirable alternatives in terms of glory could indeed "outweigh it."

Like many Hobbesian readers, we maintain that this strong view on self-preservation is predominant in all Hobbes's writings.[37] We wish

35 See F. S. McNeilly, *The Anatomy of Leviathan* (London: Macmillan, 1968).

36 Thomas Hobbes, *Thomas White's De Mundo Examined (Anti-White)*, ed. by H. Whitmore Jones (Bradford: Bradford University Press, 1976), 408, emphasis added.

37 "[N]ecessity of nature maketh men . . . to avoid that which is hurtful; but most of all that terrible enemy of nature, death" (Hobbes, *Elements of Law*, 71); "[reason] teaches every man to fly a contre-naturall dissolution, as the greatest mischiefe that can arrive to Nature" (Hobbes, *De Cive*, 27); "for every man is desirous of what is good for him, and shuns what is evill, but chiefly the chiefest of natural evills, which is death" (ibid., 47). See also Hobbes, *Elements of Law*, 71-72, 94; Hobbes, *De Cive*, 47, 53; and Hobbes, *Leviathan*, 329. Only in a few scattered passages of his works does Hobbes put forward a weaker claim on self-

to stress that endorsing $S\infty$ does not imply a vegetative life with no risk taking.[38] What assumption $S\infty$ does rule out is that there can be no trade-offs between the annihilation of one's vital and voluntary motion by being murdered and any benefits in terms of glory.[39]

G: Glory: *some* individuals seek glory, namely, the pleasure of superiority and honour.

Hobbesian critics agree in considering *G* (namely, the claim that only *some* people are glory seekers) as the prevalent view on glory put forward by Hobbes in Book 1 of *Leviathan*, to which the game-theoretic approach is usually applied. Our model would work a fortiori if *all* or *most* people were glory seekers, as suggested in places of *Elements of Law* and *De Cive*, and in Book 2 of *Leviathan*.[40]

4.2. Glory and Chicken

Assumption *G* implies that alongside glory seekers there are "moderate" individuals who, according to Hobbes, do not aim at superiority, but would be "contented with equality."[41] In terms of the game analyzed in Appendix A, the payoff rankings of moderate individuals are $P = D$, $S > W$ (case D2). For a moderate individual "avoid fighting" is a weakly dominant strategy, as it can preserve that person's life whatever other people do, whereas Fight can be life endangering (if the opponent fights back) and can never bring any additional benefit (if the opponent retreats) because of the moderate's lack of vanity.

If we were to assume that *all* people were moderate (and hence drop assumption *G*), the overall result would be a peaceful state of affairs, as described by the bees and ants game analyzed in section 2.1. It is because of *G* that the peaceful state of affairs enjoyed by bees and ants is shattered, as Hobbes does not fail to suggest in three parallel passages of *Leviathan, Elements of Law* and *De Cive*, where,

preservation by suggesting that a few individuals regard life in dishonour or with scorn or without liberty as not worth living (see Hobbes, *Elements of Law*, 39, 86; Hobbes, *De Cive*, 67; and Hobbes, *Leviathan*, 140).

38 Mountain climbing, for example, is fully compatible with $S\infty$.

39 Attaching an infinitely bad payoff to violent death at the hands of others is equivalent to the following two-stage decision process: first, partition the set of social states into the two disjoint subsets of life-endangering actions (actions that may result in violent death) and non-life-endangering actions; second, define preferences (according to glory) over the latter subset; for details see Gabriella Slomp, "Hobbes's Impossibility Theorem" (mimeographed, University of Wales, Swansea, 1995).

40 See, for example, Hobbes, *Elements of Law*, 47; Hobbes, *De Cive*, 43; and Hobbes, *Leviathan*, 156.

41 Hobbes, *Elements of Law*, 70-71; see also Hobbes, *Leviathan*, 112.

in answer to the question why people cannot live "sociably with one another," he replies:

Firstly, that men are continually in competition for honour and dignity, which these creatures are not; and consequently amongst men there ariseth on that ground, Envy and Hatred, and finally Warre; but amongst these not so. Secondly, that amongst these creatures, the Common good differeth not from the Private; and being by nature enclined to their private, they procure thereby the common benefit. But man, whose joy consisteth in comparing himself with other men, can relish nothing but what is eminent.[42]

We now turn to the relevant case of games played by glory seekers. For these individuals the relevant payoff rankings are $P < D$, $S < W$ (as in case D1 in Appendix A). A glory seeker, if faced by an opponent disposed to refrain from fighting, would prefer Fight to Avoid, for this would not only guarantee the attacker's self-preservation, but also yield the pleasure of glory in establishing superiority over the retreating opponent. On the other hand, the overriding concern for preserving one's life implies that, when faced by an opponent potentially inclined to fight, each player would prefer to avoid fighting to the self-destruction of mutual fighting. Because of the assumption of *I*ncomplete *I*nformation, no player can know in advance the psychological make-up of their opponent, or, more specifically, no player can distinguish a glory seeker who attaches a high weight to dominion as opposed to peace (that is, $D \gg P$) from one who has just a mild preference ($D > P$). As a result, in our game payoffs are not fixed but are randomly fluctuating. Hence we can apply the Harsanyi transformation and turn a game with incomplete information into a game with complete but imperfect information.

If each player knew with certainty the payoffs accruing to their opponent, the game described so far would have two asymmetric pure-strategy equilibria (with either player fighting and the other retreating) and one mixed-strategy equilibrium (where each player avoids with the appropriate probability); see Figure 8.

4.3. Chicken and the Infinitely Evil

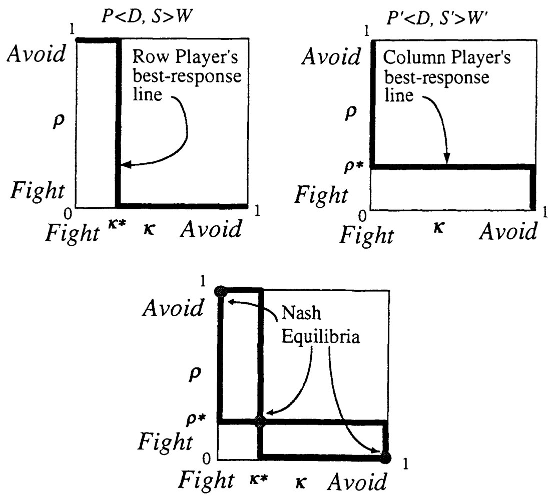
The fact that for Hobbesian people violent death at the hand of others is *infinitely* evil (assumption $S\infty$) has implications not only for the *ranking* of the payoffs associated with Fight and Avoid, but also with the specific *value* of some payoffs. Neal[43] hints at the importance of

42 Ibid., 156; see also Hobbes, *Elements of Law*, 102, and Hobbes, *De Cive*, 87.
43 Neal, "Hobbes and Rational Choice Theory."

lethal payoffs, but, given the structure of his game, they have no effect on his suggested solution.[44] Not so in our game theoretic construct, where the *infinitely* bad payoff associated with "death at the hands of others" ($W = -\infty$) can be shown to change the whole nature of the game and ultimately make *individually* rational behaviour impossible.

FIGURE 8

A CHICKEN GAME



Our argument runs as follows: first we sketch the outcome of the Chicken game on the assumption that players know without error the payoffs enjoyed by their opponents, and we show that this game can never lead to the certainty of a state of peace. Secondly, we confirm the equivalence of the Harsanyi transformation, that is, we show that, in the limit, the solution of our Chicken game when payoffs are known is the same as when the two players are uncertain as to how glory seeking their opponents are (as indicated by their payoff of Dominion). Finally, we argue that the infinitely bad payoff associated to violent death deprives Hobbesian individuals of the very possibility to take *any* rational action.

Consider the Chicken game of section 2.5 and to save on notation let $P = P' = 0,\ D = D' = 1,\ S = S' = -1,\ W = W' = -\varepsilon$. Hence the payoff matrix is displayed in Figure 9.

44 Neal's game (case B2) shares with cases B1 and D2 the property that mutual avoidance is an equilibrium even if $W = -\infty$.

FIGURE 9

A SIMPLIFIED CHICKEN GAME

Row Player \ Column Player	Avoid	Fight
Avoid	0, 0	−1, 1
Fight	1, −1	−ε,−ε

At a mixed-strategy equilibrium, each player will avoid fighting with probability:

$$p* = 1 - \frac{1}{\varepsilon}.$$ [45]

Notice first that the state of peace resulting from mutual avoidance can *never* be an equilibrium because if each player expects the other to avoid, the best response is to fight and this holds *independently* of the payoffs accruing to the players in the event of war.

Now we construct the incomplete information version of the above game. Suppose that Row and Column player are unsure of each other's payoffs associated with Dominion. More specifically, suppose that Row (Column) player's payoff if that player fights and the other player retreats is $1 + t_R$ $(1 + t_C)$, where t_R (t_C) is privately known by Row (Column) player. For simplicity we can assume that t_R and t_C are equally probable of taking on any value between 0 ("mild" glory) and ξ ("intense" glory).[46]

In Appendix B, we show that this game has a unique solution in which each player will avoid fighting if the other player's degree of desire for glory exceeds a threshold value. Moreover, we show that as the amount of uncertainty about the degree of glory shrinks (as ξ approaches 0) the outcome of the game becomes indistinguishable

45 Let p be the probability of the other player avoiding fighting. The expected payoffs from Fight and Avoid are respectively:

$$\pi(F) = 1 \cdot p + (1 - p)(-\varepsilon) \text{ and } \pi(A) = 0 \cdot p + (1 - p)(-1).$$

To find a Nash equilibrium we find the value of p that solves $\pi(F) = \pi(A)$ whence we obtain that each player's optimal probability of avoiding fighting is given by

$$p* = 1 - \frac{1}{\varepsilon}.$$

46 That is, t_R and t_C are independent draws from a uniform distribution on $[0, \xi]$.

from the standard Chicken game described in Figure 10, where each player avoids fighting with probability:

$$p* = 1 - \frac{1}{\varepsilon}.$$

FIGURE 10

CHICKEN UNDER INCOMPLETE INFORMATION

		Column Player	
		Avoid	*Fight*
Row Player	*Avoid*	0, 0	-1, $1+t_C$
	Fight	$1+t_R$,-1	$-\varepsilon,-\varepsilon$

However, the above argument holds only for a *finite* ε. Since in Hobbes's theory the outcome of opposed fight is death and death is infinitely evil ($\varepsilon = \infty$), the argument sketched above would imply that $p^* = 1$, that is, each player's rational choice would invariably be to avoid fighting. However, as shown above, mutual avoidance can never be an equilibrium of this game: even an individual mildly interested in glory cannot resist the temptation of fighting an opponent known to be disposed to avoid fighting.

The implications of this result are far-reaching: each Hobbesian individual, who by assumption *R* always seeks the counsel of reason before performing any politically relevant action, is now unable to make *any* rational choice, and is torn between the unmovable requirement of not being killed by others and the irresistible force of experiencing glory by fighting a retreating opponent.[47] Thus it is the

47 Of course, an infinitely bad payoff violates the assumption of bounded utility and thus it could be argued that it is not surprising that no rational decision making is feasible in Hobbes's characterization of the state of nature. Our response to this line of criticism is as follows. Bounded utility is sufficient, but not necessary, to guarantee the existence of equilibria in mixed strategies in non-cooperative games. Nash Theorem (John Nash, "Equilibrium Points in N-Person Games," *Proceedings of the National Academy of Sciences* 36 [1950], 48-49) states that if mixed strategies are allowed, then any normal-form finite game has at least one equilibrium. This result depends crucially on the payoff functions being *bounded* (for an elegant proof, see Theorem 3.1 in James Friedman, *Game Theory with Applications to Economic* [2nd ed.: Cambridge: MIT Press, 1990]). It is simple to verify that there are cases (Neal's coordination game, B2 and its variant B1, as well as the bees and ants game, D2) that yield mutual avoidance as an equilibrium even if $W = W' = -\infty$.

combination of infinitely bad payoffs with specific assumptions on glory-seeking behaviour that is telling.[48]

5. Concluding Remarks

The result of the game outlined above is that under conditions defined by *IR, CK, B, II, E, FR,* $S\infty$ and *G*, glory seekers can resort only to their *irrational* nature ("concupiscible part") as their inspiration of action, for their reason is mute. Unlike other game-theoretic constructs that either establish the sub-optimality of the individually rational outcome (for example, Prisoner's Dilemmas[49]) or point to the existence of multiple equilibria (see the coordination games in Kavka and Hampton, to name but two), our game shows the impossibility for glory seekers of *individually* rational decision making.

This makes sense of Hobbes's claim, unaccounted for and unexplained by any other rational-choice construct, that outside the political state "there is a dominion of passions." Indeed, Hobbes insists that in wishing to remain in the state of nature, a rational man "contradicteth himself."[50] Our game also provides an explanation for Hobbes's claim that in the state of nature war is "perpetual in its own nature."[51] Any game-theoretic application to Hobbes's state of nature (the Prisoner's Dilemma or a standard Chicken game) that yields war as the outcome of individually rational actions implies that conflict would *not* be perennial, for the simple reason that "equal powers opposed destroy each other."[52] It is the impossibility of reaching a rational plan of action and the consequent reliance on the rule of passions that make occasional battles possible and the *threat* of war a permanent menace in Hobbes's state of nature: "war, consisteth not in battle only . . . the

48 With reference to the payoff rankings listed in Appendix A, cases A1, C1 and C2 (insofar as they assume that $S < W$) are incompatible with $W = -\infty$. Case A2, which subsumes our assumption on glory insofar as it posits that $D > P$, shares with our version of Chicken the non-existence of an equilibrium, as can be easily verified.

49 Interestingly, in his translation of Thucydides' *History*, Hobbes came across what is probabiy the earliest statement of the Prisoner's Dilemma, that he rendered thus: "Everyone supposeth, that his own neglect of the common estate can do little hurt, and that it will be the care of somebody else to look to that for his own good: not observing how by these thoughts of every one in several, the common business is jointly ruined" (Thomas Hobbes, *The History of the Grecian War Written by Thucydides and Translated by Thomas Hobbes,* Vol. 8 of *English Works of Thomas Hobbes* [London: John Bohn, 1839], 147); for an analysis of Thucydides' influence on Hobbes, see Gabriella Slomp, "Hobbes, Thucydides, and the Three Greatest Things," *History of Political Thought* 11 (1990), 565-86.

50 Hobbes, *Elements of Law*, 73; and Hobbes, *De Cive*, 49.

51 Ibid.

52 Hobbes, *Elements of Law*, 34.

nature of war, consisteth not in actual fighting; but in the known disposition thereto.''[53]

The ultimate justification for deploying a game that violates the assumption of bounded utility is not simply that the resulting model accounts for important aspects of Hobbes's theory, but that it suggests a complete reversal of perspective in examining the relationship between the state of nature and the social contract that establishes the political state.

While our Chicken-*cum*-lethal-payoffs game shares with other rational-choice accounts of Hobbes's state of nature the property that rational agents are unable to escape from anarchy, its implications are orthogonal to those of existing Hobbesian rational-choice models. We do not see Hobbes's as a failed attempt to show how rational individuals may exit the state of nature, but as a rational explanation of why citizens cannot allow themselves to fall into a state of civil war. In other words, our game-theoretic interpretation lends support to the view that Hobbes's political theory is aimed not at providing ''natural'' individuals with a blueprint for action, but at persuading people already living in civil associations of the irrationality of civil war.

The shift from individual action to the acknowledgment of certain political arrangements as rational calls for a corresponding shift in perspective, namely, from rational-actor to external-observer analysis. By the latter, we mean the analytical approach of collective-choice theory which, in Hobbes's case, entails each individual taking a stand-back position, shedding the mantle of glory, and acknowledging the irrationality of a return to the state of nature.[54]

Appendix A

In this appendix we examine all the relevant outcomes of the two-strategy two-player non-cooperative game applied to Hobbes's state of nature. The number of permutations of possible equilibria is substantially reduced if it is assumed that the two players have the same preference ordering of *P*eace vs. *D*ominion and of *W*ar vs. *S*ubjection, that is,

$$\text{sign}\,(P - D) = \text{sign}\,(P' - D') \text{ and } \text{sign}\,(W - S) = \text{sign}\,(W' - S').$$

We can distinguish eight possible rankings of the relevant payoffs.[55]

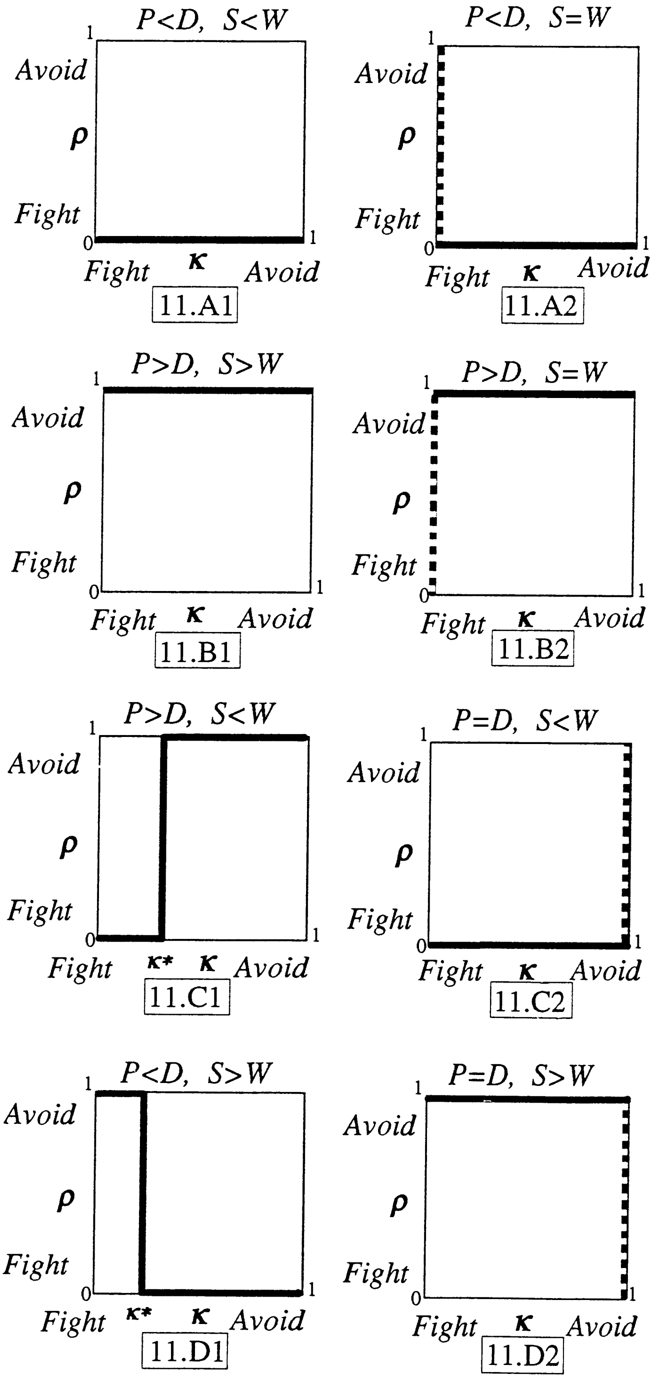
53 Hobbes, *Leviathan*, 113.

54 In a follow-up to this article, one of us provides a new interpretation of Hobbes's political construct from a fully fledged external-observer perspective (see Slomp, ''Hobbes's Impossibility Theorem'').

55 We ignore the trivial case $P = D, S = W$.

FIGURE 11

BEST-RESPONSE LINES FOR ROW PLAYER



Case (A1): $P < D$ and $W < S$: The Fight strategy strictly dominates the Avoid strategy, whatever the probability that Column player avoids fighting, Row player will always prefer fighting. The *best-response* line for Row player is the thick line of Figure 11.A1. One way of reading each of the eight probability boxes in Figure 11 is to imagine Row player choosing the best route to cross the box from left to right, given their preferences Row player[56] will select the "best" probability of avoiding fighting (ρ) for each value of the probability with which that player's opponent avoids fighting ($0 \leq \kappa \leq 1$).

Case (A2): $P < D$ and $W = S$: In this case, Avoid is a *weakly dominated strategy* in the sense that Row player can never be worse off by choosing Fighting instead. The player is indifferent between their two available strategies in the event of Column player choosing to Fight, and strictly prefers Fighting if Column player chooses to avoid fighting. For this class of games[57] we believe that we can reasonably eliminate any weakly dominated strategy. This is why in the best response line of Figure 11.A2 the vertical dotted line (signifying Row player's indifference if Column player chooses Fighting) is assumed inoperative.[58]

Case (B1): $P > D$ and $W > S$; and *Case (B2)*: $P > D$ and $W = S$: These are the reverse of cases (A1) and (A2), with Avoidance being respectively the dominant and weakly dominant strategy for each player. See Figure 11.B1 and 11.B2.

Case (C1): $P > D$ and $W < S$: This is the case already examined as an example in section 1, to which we refer the reader. The best-response for Row player is described by the thick stepwise line of Figure 11.C1.

Case (C2): $P = D$ and $W < S$: In the event of Column player choosing to Avoid (if $\kappa = 1$), Row player will be indifferent between playing either strategy (and indeed any probability mixture of Avoid and Fight); otherwise Row player always avoids, as described in Figure 11.C2.

Case (D1): $P < D$ and $W > S$; and *Case (D2)*: $P = D$ and $W < S$: These are the reverse respectively of cases (C1) and (C2). See Figure 11.D1 and 11.D2.

56 Gender neutrality is a substantial component of Hobbes's theory; on this, see Gabriella Slomp, "Hobbes and the Equality of Women," *Political Studies* 42 (1994), 441-52.

57 For a defence of the elimination of weakly dominated strategies, see, for example, Roger Myerson, *Game Theory: Analysis of Conflict* (Cambridge: Harvard University Press, 1991), Theorem 1.7, 30. The potential problem of eliminating weakly dominated strategies, namely, that the *order* in which the elimination takes place may affect the resulting equilibrium, simply does not arise in two-strategy games such as those considered here.

58 Pursuing the box-crossing analogy sketched a few lines above, we may say that what matters is going from one side of the box to the opposite side, and that wandering around on the zebra crossings on either side of the box is a waste of time.

Appendix B

In the unique pure-strategy Bayesian Nash equilibrium of the game with incomplete information[59] described in Figure 10, Row (Column) player avoids fighting if t_R (t_C) exceeds a threshold value $\bar{\bar{t}}_R$ $(\bar{\bar{t}}_C)$. In this equilibrium each player avoids fighting with probability:

$$\xi - \frac{i}{\xi},\ i = \bar{\bar{t}}_R,\ \bar{\bar{t}}_C.$$

As incomplete information vanishes (as ξ approaches 0) the two players' behaviour at the pure-strategy Bayesian equilibrium will approach their behaviour at the mixed-strategy equilibrium of the complete information version of the game. That is to say, we can now show that

$$\xi - \frac{i}{\xi} \rightarrow 1 - \frac{1}{\varepsilon} \text{ as } \xi \rightarrow 0.$$

Row player's expected payoffs from Avoid and Fight are given respectively by

$$u_R(A) = \frac{\xi - \bar{\bar{t}}_C}{\xi} \cdot 0 + \frac{\bar{\bar{t}}_C}{\xi} \cdot (-1) \text{ and } u_R(F) = \frac{\xi - \bar{\bar{t}}_C}{\xi} \cdot (1 + \bar{\bar{t}}_R) + \frac{\bar{\bar{t}}_C}{\xi} \cdot (-\varepsilon).$$

Therefore Avoid is optimal for Row player iff

$$t_R \geq \frac{\xi - \bar{\bar{t}}_C \varepsilon}{\bar{\bar{t}}_C - \xi} = \bar{\bar{t}}_R.$$

Proceeding in an analogous way for Column player, we obtain

$$\bar{\bar{t}}_R = \bar{\bar{t}}_C = \frac{\xi - \varepsilon + ((\xi - \varepsilon)^2 + 4\xi)^{1/2}}{2}.$$

Applying de l'Hospital rule, we finally obtain that the probability that each player avoids fighting indeed approaches

$$1 - \frac{1}{\varepsilon}$$

as ξ vanishes. Q.E.D.

59 The static Bayesian game in normal form is $\Omega = \{A_A, A_F; T_C, T_R; pr_C, pr_R; u_C, u_R\}$, the action spaces are $A_A = A_F = \{\text{Avoid, Fight}\}$, the type spaces are $T_C = T_R = [0,\xi]$, the beliefs are $pr_C(t_R) = pr_R(t_C) = 1/\xi$, $\forall t_R, t_C$ and the payoffs u_C, u_R are as described in Figure 10.