Economics and Philosophy, **22** (2006) 243–279 Copyright © Cambridge University Press
doi:10.1017/S0266267106000897

WAR OR PEACE?: A DYNAMICAL ANALYSIS OF ANARCHY*

PETER VANDERSCHRAAF

I propose a dynamical analysis of interaction in anarchy, and argue that this kind of dynamical analysis is a more promising route to predicting the outcome of anarchy than the more traditional *a priori* analyses of anarchy in the literature. I criticize previous *a priori* analyses of anarchy on the grounds that these analyses assume that the individuals in anarchy share a unique set of preferences over the possible outcomes of war, peace, exploiting others and suffering exploitation. Following Hobbes' classic analysis of anarchy, I maintain that typically in anarchy some *moderate* individuals will most desire mutual cooperation while other *dominators* will most desire to exploit others' cooperation. I argue that once one allows for different types of individuals in anarchy, any *a priori* analysis of anarchy requires unrealistic assumptions regarding the agents' common knowledge of their situation. However, this move also suggests a dynamical analysis of anarchy, one that assumes no common knowledge. In the *Variable Anticipation threshold* model developed here, individuals modify their behavior as they learn from repeated interactions. I present specific instances of this model where the individuals in anarchy converge to different equilibria corresponding to either peace or war, depending on the initial conditions. I show that individuals are liable to converge to Hobbes' war of all against all even if only a small percentage of are dominators. The presence of only a few "nasty" individuals gradually drives all, including those inclined to be "nicer", to imitate the "nasty" conduct of these few. This dynamic analysis suggests that the Hobbesian war in anarchy is indeed inevitable in most realistic circumstances.

You have the same propension, that I have, in favor of what is contiguous above what is remote. You are, therefore, naturally carry'd to commit acts of injustice as well as I. Your example both pushes me forward in this

* Thanks to Luc Bovens, Sharon Lloyd, Brian Skyrms, Susanne Sreedhar and an anonymous referee for many helpful comments of early versions of this essay.

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

> way by imitation, and also affords me a new reason for any breach of equity, by showing me, that I shou'd be the cully of my integrity, if I alone shou'd impose on myself a severe restraint amidst the licentiousness of others. (David Hume, *A Treatise of Human Nature*)

INTRODUCTION

Hobbes' description of the natural condition of mankind or *State of Nature* characterizes anarchy.[1] Hobbes and his fellow Hobbesians argue that anarchy inevitably leads to a war of all against all. They also argue that the rational choices of the individuals in anarchy are the direct causes of this war. Hobbesians conclude that rational individuals must live in civil society, the unique alternative to the Hobbesian State of Nature, in order to avert the war of all against all. Others deny the Hobbesian conclusion, claiming that people who lack a government are not so prone to act aggressively as Hobbesians suppose. Locke argues that people need not fall into conflict in the State of Nature he analyzes, of which anarchy is a special case. Locke even suggests that peace prevails in the Lockean State of Nature when all act rationally. If Locke and his fellow Lockeans are right, the outcome of anarchy is indeterminate, and rational people do not necessarily need to establish civil society in order to secure peace.

In our time, philosophers and social scientists use game theory to develop more precise Hobbesian and Lockean analyses of interaction in anarchy.[2] If, as some Hobbesians believe, everyone in a state of anarchy strictly prefers exploiting others' cooperation over reciprocating cooperation, then all interactions are Prisoners' Dilemmas. If everyone strictly prefers to reciprocate cooperation over exploiting cooperation, as some Lockeans believe, then all interactions in anarchy have an Assurance game structure. The unique Nash equilibrium[3] of the Prisoners' Dilemma corresponds to war, while Assurance has Nash equilibria corresponding to both war and peace. So one might suppose that to resolve the dispute between the Hobbesians and the Lockeans, one need only show that the individuals in anarchy have Prisoners' Dilemma preferences or that they have Assurance preferences. More generally, one might conclude that the

[1] See especially *Leviathan* (ch. 13) and *De Cive* (ch. 1). Below I will give the chapter and paragraph number when citing a passage in Hobbes' *Leviathan* or *De Cive*, and section number when citing a passage in Locke's *Second Treatise of Government*. For instance, *Leviathan* 3:4 refers to the 4th paragraph of ch. 3 of *Leviathan*, and *Second Treatise* §19 refers to the 19th section of the *Second Treatise of Government*.

[2] Some of the most important works in this tradition include Lewis (1969), Hampton (1986), Kavka (1986), Sugden (1986), Binmore (1994) and Jiborn (1999).

[3] A strategy profile $\boldsymbol{s}$ of a game is a *Nash equilibrium* (1950, 1951) if each agent gains her highest expected payoff by following the strategy that is her part of $\boldsymbol{s}$ on condition that all follow their parts of $\boldsymbol{s}$. $\boldsymbol{s}$ is a *strict* Nash equilibrium if $\boldsymbol{s}$ defines each agent's unique expected payoff maximizing strategy given that all follow $\boldsymbol{s}$.

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

key to analyzing anarchy is to identify the game that most accurately reflects the preferences of all of the individuals in anarchy.

However, in a large population the preferences of the individuals might not be characterized by a single game. Some in the population are likely to prefer to exploit cooperation, while others are likely to prefer to reciprocate cooperation. In this paper I will propose a new game-theoretic model of interaction in anarchy, one I believe reflects assumptions regarding human psychology that both Hobbes and Locke would accept and that their followers should accept. Unlike other game-theoretic models of anarchy in the literature, the *Variable Anticipation Threshold* model developed in this paper allows individuals or *parties* to have a range of heterogeneous preferences. Previous game-theoretic models of anarchy are static, in the sense that the philosophers and social scientists who use them assume that one can predict the outcome of anarchy via *a priori* analysis of a particular game. I maintain that once one allows for different types of individuals in anarchy, any *a priori* analysis of anarchy will rest upon unrealistic assumptions regarding the agents' common knowledge[4] of their situation. However, incorporating alternative types of agents into a model of anarchy suggests a dynamical analysis, one that assumes no common knowledge in anarchy. In the Variable Anticipation threshold model, individuals modify their behavior dynamically as they learn from repeated interactions. I present specific instances of this model in which the dynamical systems that characterize the beliefs of the individuals in anarchy converge to different equilibria corresponding to either peace or war, depending on the initial conditions. One striking finding presented here is that individuals are liable to converge to Hobbes' war of all against all even if only a small percentage of them actually prefer to exploit others' cooperation. The presence of only a few "nasty" individuals gradually drives all, including those inclined to be "nicer," to imitate the "nasty" conduct of these few. This dynamic analysis suggests that the Hobbesian war in anarchy is indeed inevitable in most realistic circumstances.

The remainder of this paper is organized as follows: Section 1 sets the stage by reviewing Hobbes' and Locke's alternate analyses of anarchy and the Prisoners' Dilemma and Assurance models of State of Nature interaction based upon their divergent conclusions. Section 2 evaluates two *hybrid models* of the Hobbesian State of Nature that distinguish between different types of agents in anarchy. The structures of these hybrid models foreshadow the heterogeneous preference structure of the dynamic Variable Anticipation Threshold model introduced in this paper.

[4] Lewis (1969) presented the first analysis of common knowledge. A proposition A is *Lewis-common knowledge* among a group of agents if each agent knows that all know A and knows that all can infer the consequences of this mutual knowledge (Lewis 1969: 56–7).

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

Kavka (1986, 1989) proposes an *Assurance Dilemma* model of interaction in anarchy that combines elements of the Prisoners' Dilemma and Assurance games. Kavka uses the Assurance Dilemma to give a neo-Hobbesian *a priori* argument that war is the inevitable outcome of anarchy. I argue that Kavka's model, while an important step forward, is insufficiently general given Kavka's own analysis. I maintain that a π-*Assurance* game of incomplete information is a better hybrid model of the Hobbesian State of Nature than Kavka's Assurance Dilemma. One's rational choice in π-Assurance depends on the likelihood of meeting a given type of counterpart party. If one adopts a π-Assurance model of anarchy, then war is not inevitable or even necessarily likely. In the end, I maintain that this hybrid model is still insufficiently general and also rests upon implausible common knowledge assumptions. Nevertheless, the π-Assurance model helps to motivate a dynamical model of interaction in anarchy. Section 3 introduces the Variable Anticipation Threshold model. In this model, each party has a personal payoff function reflecting how much she prefers peace over war and how much she is tempted to exploit others rather than cooperate with them. Over time, each follows a strategy that is rational given what she has learned from past encounters. In this section I discuss the results of computer simulations of interaction in anarchy that converge to the equilibrium of peace or war, depending on whether or not some of the agents strictly prefer to exploit others. Section 4 discusses some of the advantages of analyzing anarchy dynamically using variable payoffs, and closes with suggestions for future research.

1. TWO ACCOUNTS OF ANARCHY

In *Leviathan*, Thomas Hobbes declares that people are in a State of Nature when they "live without a common Power to keep them all in awe . . ." (*Leviathan* 13:8). The Hobbesian State of Nature is a state of anarchy. Hobbes famously claims that in this state, "there is no way for any man to secure himselfe, so reasonable, as Anticipation . . ." (*Leviathan* 13:4). What exactly is Anticipation? Hobbes gives an official definition: For one to Anticipate is "by force or wiles to master the persons of all men he can."[5] But this account of Anticipation is implausible if taken literally. Given Hobbes' explicit claims that adult people have roughly equal mental and physical powers (*Leviathan* 13:3), an individual in a Hobbesian State of Nature with respect to others will seldom, if ever, be able to fully master any of these others by his own force or wiles for even a brief time.[6] Hobbes needs a more general definition of Anticipation. And, in fact, he suggests a more

[5] This is Hobbes' only description of Anticipation in *Leviathan*. The term "anticipate" appears in only one other passage in *Leviathan* (11:9).

[6] Sharon Lloyd suggested an interesting exception to me, namely, that an especially charismatic individual might be able to persuade others to follow his bidding without

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

general definition in *De Cive*, where he says that to Anticipate is to attack another preemptively in an attempt to get the better of the other (Preface 12). Hobbes defines power formally as means by which one can obtain some future goods. Along with one's *natural* powers of body and mind, one may acquire *instrumental* powers, such as external goods (*Leviathan* 10:1–2). In the Hobbesian State of Nature, the instrumental powers others possess are a primary source of the power Hobbes claims people restlessly seek in order to live well (*Leviathan* 11:2). So Hobbes can define Anticipation against others as the seizure of *some* of their power. Often one Anticipates against another by seizing some of the goods she possesses. One can also Anticipate against another by occupying some of her physical and mental powers in a conflict. For the remainder of this paper, I will employ this more general definition of Anticipation.

As is well known, Hobbes concludes that individuals in the State of Nature do *not* make themselves more secure when they all do what is allegedly "so reasonable." For when individuals Anticipate against each other, they fall into a conflict that includes costly defensive posturing and occasional violent fighting. Such individuals are in a state of war, a war that profits no one. And this war is a consequence of each individual following the course of action Hobbes claims to be the most reasonable. Hence Hobbes' notorious description of life in the State of Nature as "solitary, poore, nasty, brutish, and short" (*Leviathan* 13: 8, 9).

John Locke claims that people in the State of Nature he analyzes might arrive at a very different outcome.

> And here we have the plain *difference between the State of Nature, and the State of War*, which however some Men have confounded, are as far distant, as a State of Peace, Good Will, Mutual Assistance, and Preservation, and a State of Enmity, Malice, Violence and Mutual Destruction are one from another. Men living together according to reason, without a common Superior of Earth, with Authority to judge between them, is properly the State of Nature. But force, or a declared design of force upon the Person of another, where there is no common superior on Earth to appeal to for relief, is the *State of War*. (*Second treatise* Section 19)

question, perhaps for religious reasons. However, I believe that even if such exceptional individuals exist in a Hobbesian State of Nature, in order for any one of them to maintain a permanent following he would have to be able to control what his followers know *and* to suppress both communication among his followers and information from the outside. As Hobbes himself suggests (*Leviathan* 13:9), one of the disadvantages of anarchy is that it is difficult if not impossible to generate mutual or common knowledge or otherwise regulate the flow of information among a group of people. So groups of fanatical disciples of a leader would be rare and inherently unstable in the Hobbesian State of Nature. Still, Lloyd's suggestion may prove valuable in studying processes of social reform, in which a population makes a transition from one stable state to another.

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

Locke maintains that philosophers like Hobbes are wrong to identify the State of Nature with the state of war. In fact, here Locke suggests that when people act according to reason, the State of Nature is a state of *peace*. Yet soon afterwards, Locke speaks of the State of Nature as a state of war, after all (*Second treatise* §21). Some of Locke's readers accuse Locke of blatantly contradicting himself (Macpherson 1962; Jenkins 1967). On a more charitable reading, Locke is simply arguing that people in a State of Nature might end up *either* at peace or at war, or even in some intermediate state between peace and war (Simmons 1989; 1993: ch. 1).

What accounts for Hobbes' and Locke's divergent conclusions? Locke and Hobbes propose differing accounts of the State of Nature, but this is no answer to the question because they agree that people in anarchy are in a State of Nature. In *Leviathan* 13:8, Hobbes *defines* his State of Nature as a state of anarchy. According to Locke, people are in a State of Nature with respect to each other when they are not ruled by a government they have consented to obey (*Second treatise* §14, §15). So for Locke, anarchy is a sufficient but not a necessary condition for a State of Nature to exist between parties.[7] Since Hobbes and Locke agree that anarchy characterizes a State of Nature, perhaps they disagree over what rational people in anarchy do because they disagree over what these people are like. This is a common interpretation of this dispute between Hobbes and Locke, and is well summarized by a pair of game-theoretic models of interaction in anarchy proposed as summaries of Hobbes' and Locke's arguments.

Many authors claim that Hobbes' analysis of the State of Nature foreshadows the reasoning of parties engaged in a Prisoners' Dilemma.[8] As in any Prisoners' Dilemma, in the Hobbesian State of Nature, each party must choose between "cooperative" and "individualist" options, and all prefer what results if all are "cooperative" over what occurs if all are "individualists." Suppose that in the Hobbesian State of Nature, each individual can either *cooperate* with (*C*) or *Anticipate* against (*A*) those he encounters. To cooperate is to leave the other in peace. To Anticipate is to treat the other as an enemy by seizing some of the power available to the other when she is left in peace. Note that an Anticipator exploits a cooperator by seizing some of the cooperator's power without resistance, and still seizes some of another Anticipator's power by occupying the other's energies in the ensuing conflict. A cooperator avoids the costs of defensive posturing and conflict, but also becomes vulnerable to exploitation. When both parties cooperate, they are at *peace* with each other. When both Anticipate, they at *war*. This corresponds to Hobbes'

[7] For fine critical studies of Hobbes' and Locke's accounts of the State of Nature, see Kavka (1983; 1986: ch. 3) and Simmons (1989; 1993: ch. 1).

[8] Some of these authors include Barry (1965: 253–4), Rawls (1971: 269), Gauthier (1979: 554), Taylor (1976: ch. 6), Curley (1994: *xxiv–xxv*) and Jiborn (1999: sect. 4.4).

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

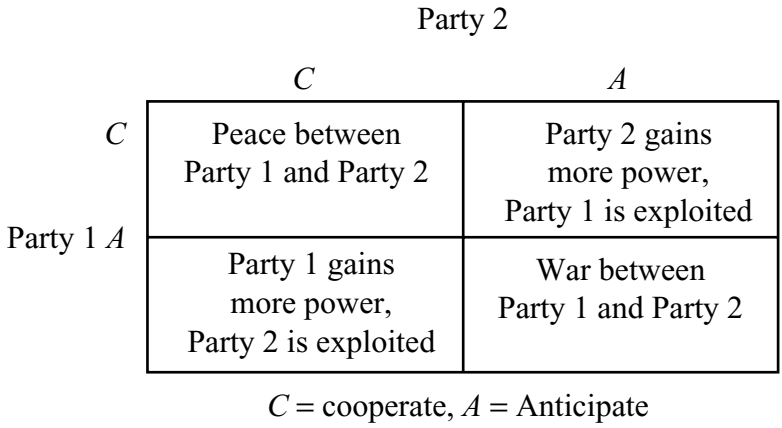


FIGURE 1.1. Two parties in a Hobbesian State of Nature.

accounts of war as a state in which all know that all are predisposed to fight, and peace as a state where all trust each other not to fight (*Leviathan* 13:8). Figure 1.1 summarizes the outcome of a pairwise encounter in the Hobbesian State of Nature.

The key reason so many attribute Prisoners' Dilemma reasoning to Hobbes is that Hobbes explicitly asserts that Anticipation is everyone's most reasonable option in the State of Nature, even though each party prefers the peace that results if all cooperate over the war that results if all Anticipate. Hobbes' assertion leads some of his readers to surmise that Hobbes is presenting an informal dominance argument for Anticipation. An agent's choice is (Bayesian) rational if this choice maximizes her expected payoff, based upon her preferences and her beliefs regarding her situation. A (Bayesian) rational agent consistently chooses so as to maximize expected payoff. The Hobbesian State of Nature indeed has a Prisoners' Dilemma structure if the "reasonable" Anticipation is one's *strictly dominant* choice in a State of Nature encounter, that is, Anticipation is one's unique rational choice whatever the others choose to do. In a pairwise State of Nature encounter, *A* strictly dominates *C* if one assumes the following: One's best outcome occurs if one Anticipates and the other cooperates, since then one enhances his security by gaining control over more resources. The second best outcome occurs if both cooperate, since then one keeps what he has at no cost. The third best outcome occurs if both Anticipate, for then one suffers the costs of a conflict but keeps the remainder of his resources. But one's worst outcome occurs if one cooperates unilaterally. If one cooperates with an Anticipator, then one loses so great a share of one's original resources that the loss exceeds the costs of conflict. Figure 1.2 summarizes the preferences of a pair of such

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

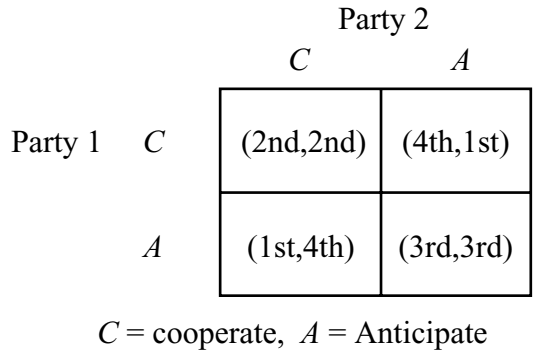


FIGURE 1.2. Parties' preference rankings in a Prisoners' Dilemma.

parties in a State of Nature encounter. From these preferences, one can infer that the parties' payoffs given the various possible outcomes characterize a Prisoners' Dilemma. According to the logic of the Prisoners' Dilemma, by following their strictly dominant strategies the two parties settle into the unique Nash equilibrium (*A*, *A*), the outcome of war. One can extend this logic to a larger population of parties in anarchy. Assuming the gains of Anticipation, the losses of exploitation and the costs of conflict increase monotonically with the number of parties one encounters, rational parties will all Anticipate. From these premises, sometimes attributed to Hobbes, we have arrived at the war of all against all.

As noted above, on a sympathetic reading Locke is arguing that the outcome of anarchy is indeterminate. In several passages, Locke states that one should leave another in peace and in possession of his goods unless he first acts, or overtly threatens to act, aggressively without provocation (*Second Treatise* § 16, § 19). But if another party strikes first, Locke allows that one may justly punish this aggressor until one has reason to believe that the aggressor will change and treat others peaceably in the future:

> in the State of Nature, for want of positive Laws, and Judges with Authority to appeal to, *the State of War once begun, continues*, with a right to the innocent Party, to destroy the other whenever he can, until the aggressor offers Peace, and desires reconciliation on such Terms, as may repair any wrongs he has already done, and secure the innocent for the future . . . . (*Second Treatise* § 20)

Locke would say that one should Anticipate against another only if one expects the other to Anticipate, and one should expect this only if the other overtly threatens to Anticipate or has Anticipated before. When parties Anticipate against each other, they are at war, but Locke claims that even in anarchy warring parties can make peace. Locke's discussion also suggests that a rational individual *prefers* to Anticipate against another only when

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

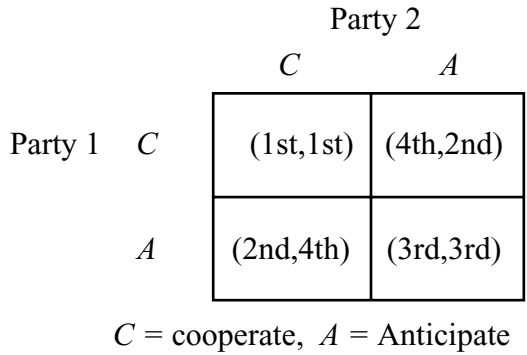


FIGURE 1.3. Parties' preference rankings in an Assurance game.

she expects the other to Anticipate. Figure 1.3 summarizes the preferences of rational individuals according to this interpretation of Locke.

The preferences of Figure 1.3 characterize a coordination game known as *Assurance* (Sen 1967, 1973).[9] In this game, no strategy is strictly dominant. To cooperate is the better strategy, provided that one expects the other to cooperate. To Anticipate is to forgo the chance of realizing the outcome that is best for both, in exchange for being certain not to suffer one's own worst outcome. The (*A, A*) outcome of war remains an equilibrium in this Assurance game. But now the (*C, C*) outcome of peace is also an equilibrium, and (*C, C*) is optimal. If anarchy really has an Assurance structure, then war is not inevitable after all. Indeed, as the (*C, C*) equilibrium is optimal, some authors suggest that this is the equilibrium rational individuals will select.[10] By extending the logic of the Assurance game over all of the parties in anarchy, the outcome where all cooperate is an optimal equilibrium that all will follow if each party expects the others to reciprocate cooperation. We have arrived at peace from Lockean premises.

Philosophers use the Prisoners' Dilemma or the Assurance game in their attempts to complete the Hobbesian project of predicting the outcome of anarchy as the result of agents' rational decisions. To be sure, neither of

[9] A coordination game (Schelling 1960: 84; Lewis 1969: 14) is a game where if an agent can improve her own payoff by changing her strategy, then when she does so she improves every agents' payoff, so that the interests of the agents coincide.

Kavka (1989: 5–6) is the only other author I know of who suggests that Locke's State of Nature can be modeled as an Assurance game based upon Figure 1.3 preferences. Kavka also proposes Assurance as a model of a Hobbesian State of Nature in which all parties are *moderates*, who only prefer to Anticipate in self-defense. See Section 2 and Section 3 for a more detailed discussion of moderates in anarchy.

[10] See, for instance, Hampton (1986: sect. 2.2).

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

these games fully summarizes the predicament of individuals in anarchy. Few actual encounters between people can be described completely by a payoff matrix alone. Indeed, as I shall argue below, the analysis of interaction in anarchy changes dramatically when one explicitly considers how parties might interact repeatedly over time with neighbors. In Section 3 these factors will be modeled explicitly with iterated games embedded in a spatial structure. But first, I will consider a crucial part of Hobbes' own treatment of the preferences of individuals in anarchy that cannot be adequately captured by either the Assurance game or the Prisoners' Dilemma.

2. HYBRID MODELS OF ANARCHY

The Prisoners' Dilemma and the Assurance models of anarchy reviewed in Section 1 yield different results because they are developed from different sets of assumed preferences. If these two games accurately summarize the key elements of Hobbes' and Locke's analyses of anarchy, then the dispute between Hobbesians and Lockeans is a dispute over *psychology*. However, I suspect that neither of the game-theoretic models of anarchy considered in Section 1 accurately reflect human psychology. For the Prisoners' Dilemma and Assurance models of interaction in anarchy presuppose that parties have identical preferences over the alternative outcomes. It is admittedly possible that all parties in a Hobbesian State of Nature share the same set of preferences. So the Hobbesian State of Nature could have a Prisoners' Dilemma or an Assurance structure. But these are endpoint cases. It is also possible that the preferences of parties in a Hobbesian State of Nature *vary* to some extent. The likelihood that such variance is present in a State of Nature grows as the number of parties in this State of Nature grows. In this section I will examine two hybrid models of anarchy that incorporate heterogeneous preferences across parties.

Gregory Kavka (1986, 1989) developed the first contemporary Hobbesian analysis of anarchy that incorporates different *types* of parties. Kavka's novel but little known *Assurance Dilemma* model of anarchy is a precursor to the Variable Anticipation Threshold model of anarchy introduced in this paper. As noted in Section 1, the Hobbesians who attribute an informal Prisoners' Dilemma analysis of anarchy to Hobbes rely upon Hobbes' claim that Anticipation is one's most reasonable option. However, for Anticipation to be each individual's best option, it is not necessary that Anticipation be each individual's strictly dominant option. In fact, Hobbes may have understood this informally. Consider what Hobbes says in *Leviathan* immediately after recommending Anticipation as the most reasonable course of action in the Hobbesian State of Nature:

> And this is no more than his own preservation requireth, and is generally allowed. Also, because there be some, that taking pleasure in contemplating

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

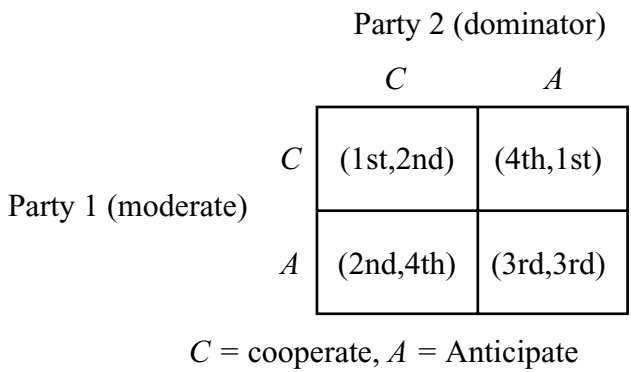


FIGURE 2.1. Parties' preference rankings in an Assurance Dilemma.

> their own power in acts of conquest, which they pursue farther than their security requires; if others, that otherwise would be glad to be at ease within modest bounds, should not by invasion increase their power, they would not be able, long time, by standing only on their defence, to subsist. And by consequence, such augmentation of dominion over men, being necessary to a man's conservation, it ought to be allowed him. (*Leviathan* 13:4)[11]

Here Hobbes identifies two types of parties in the State of Nature. Parties of only one of these two types prefer to Anticipate against cooperators. I will follow Kavka's terminology and refer to parties of this aggressive type as *dominators*. The more "modest" parties are *moderates* (Kavka 1986: 97). Both types most fear exploitation and prefer peace over war. However, a moderate most prefers mutual cooperation, while a dominator most prefers to exploit cooperators. Figure 2.1 summarizes the preferences of Party 1, a moderate, and Party 2, a dominator, matched in a pairwise State of Nature encounter.

The Figure 1 preferences characterize a game Kavka (1989) dubs the *Assurance Dilemma*. This game has a determinate outcome if both know their preferences and at least the moderate knows that they are engaged in an Assurance Dilemma and that they are both rational. For *A* is Party 2's strictly dominant choice, and Party 1, knowing this, knows that the possible outcomes of the game are (*C, A*) and (*A, A*), depending on her choice. Since Party 1 prefers (*A, A*) over (*C, A*), she chooses *A*. Hence (*A, A*) is the unique *rationalizable* strategy profile (Bernheim 1984; Pearce 1984) of the Assurance Dilemma.[12] Kavka calls this game the Assurance *Dilemma*

[11] Similar texts appear in *Leviathan* (11:2) and *De Cive* (1:4).

[12] Kavka (1989: 9–10) gives a similar argument that (*A, A*) is the determinate solution of the Assurance Dilemma. In his argument, Kavka does not state the necessary assumptions regarding the parties' knowledge of their situation. Also, Kavka was evidently unaware

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

because (*A*, *A*) is the unique Nash equilibrium of this game even though only one of the parties prefers to deviate from the optimal (*C*, *C*) outcome.

The Assurance Dilemma is a hybrid of the Prisoners' Dilemma and Assurance. Kavka maintains that the Assurance Dilemma is

> the best simple representation of Hobbes's analysis of the problem of anarchy. Since it relies on less cynical assumptions about human nature than the standard prisoner's dilemma view, we may call this the *Kinder-Gentler* Interpretation of Hobbes's state of nature. (Kavka 1989: 10)[13]

The Assurance Dilemma model incorporates more "kinder, gentler" preferences than those of the Prisoners' Dilemma. Still, if encounters in the Hobbesian State of Nature have this preference structure, one can predict that rational parties with sufficient mutual knowledge of their situation will go to war. Hobbes' own texts indicate that he attributes "kinder, gentler" preferences to people than do many of his followers. If only all in the State of Nature were moderates, then peace would be an equilibrium. Alas, this is not likely to be the case. Dominators are likely to be present in the State of Nature, and these dominators will start a spiral of expected Anticipation that leads all to go to war (Kavka 1986: sect. 3.2, sect. 3.3; 1989). The Assurance Dilemma illustrates a simple case of how this spiral unfolds.

Kavka's Assurance Dilemma model is an important step forward in the analysis of the Hobbesian State of Nature, but it has its shortcomings. The Assurance Dilemma model summarizes a compelling argument for why mutual Anticipation is inevitable when a rational moderate and a rational dominator meet and the moderate recognizes the dominator as such. But how often does this happen? When dominators meet, their encounter is a Prisoners' Dilemma. When moderates meet, their encounter is an Assurance game. If moderates can identify the dominators in their midst, then of course they can identify the other moderates. Such moderates might adopt the *contingency strategy* (Lewis 1969; Aumann 1974) of cooperating with moderates and Anticipating against dominators. The premises necessary to enable one to predict that parties in an Assurance Dilemma will surely follow (*A*, *A*) on account of their rationality also imply that when moderates are present in a Hobbesian State of Nature, the outcome need not be Hobbes' war of all against all. Indeed, if the

that his argument is an informal rationalizability argument. A strategy profile of a game is *rationalizable* if this profile remains after stepwise deleting strictly dominated strategies of the stepwise reduced game. A game is *dominance solvable* if it has a unique rationalizable strategy profile, as does the Assurance Dilemma. Players will follow a rationalizable strategy profile if they have common knowledge of their rationality and the game's payoff structure (Brandenburger and Dekel 1988; Vanderschraaf 2001: ch. 2).

[13] I have followed the exact punctuation, capitalization and emphases in Kavka's unpublished manuscript.

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

moderates in anarchy consistently cooperate with the other moderates and Anticipate against the dominators, the outcome is the intermediate state between peace and war that Simmons (1989; 1993: ch. 1) argues Locke predicted!

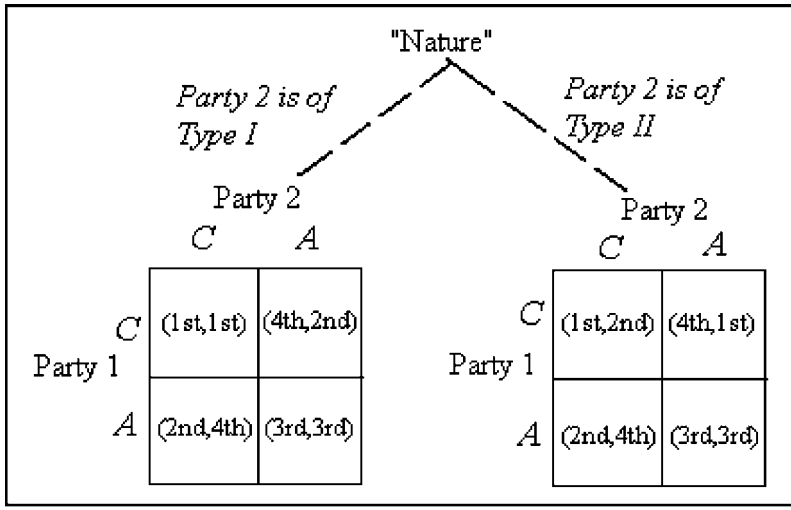
Kavka may not have fully recognized this problem for his Assurance Dilemma model. But he was evidently aware that not all encounters in anarchy will be Assurance Dilemmas. Kavka argues that the moderate parties in a Hobbesian State of Nature will reason *as if* they were in an Assurance Dilemma because these parties in fact *cannot* distinguish the dominators from the moderates. Given that the moderates are uncertain as to whether or not the other parties they meet are dominators, Kavka argues their best choice is to "play it safe" and Anticipate. Consequently, these cautious moderates together with the dominators generate more expected Anticipation in an ever-increasing spiral (Kavka 1986: 104–7; 1989). Again, Hobbes himself gives a similar argument:

> we cannot tell the good and the bad apart, hence even if there were fewer evil men then good men, good, decent people would still be saddled with the constant need to watch, distrust, anticipate and get the better of others, and to protect themselves by all possible means. (*De Cive*: Preface 12)

Hobbes, and Kavka after him, maintain that the moderates' uncertainty as to whether or not other parties they encounter are dominators ultimately leads the moderates to avoid ever running the risk of being exploited. Following their reasoning, a moderate will treat any other party she encounters as if she were certain that this other party is a dominator.[14] Hence, from the perspective of a moderate, anarchy does have an Assurance Dilemma structure in which one's only rational alternative is to Anticipate.

[14] An anonymous referee suggested that Kavka may be assuming that in the Hobbesian State of Nature parties should apply some cautious principle of decision such as maximin. Kavka does not say explicitly which decision rule rational parties apply in this context. However, elsewhere Kavka recommends a principle of *disaster avoidance* in situations where one can suffer a disastrous outcome and one has only ordinal information regarding the likelihoods of alternative outcomes. Kavka's principle of disaster avoidance recommends choosing an alternative action that minimizes one's chances of suffering any disastrous outcome (1986: 142–4, 200–10). Kavka endorses the disaster avoidance principle in arguing for the rationality of keeping promises in certain circumstances, and attributes disaster avoidance reasoning to Hobbes in the response to the Foole in *Leviathan* 15 (Kavka 1995). So one might be tempted to read Kavka's discussion of the Assurance Dilemma model of anarchy as an application of disaster avoidance, where the moderate chooses *A* so as to avoid any chance of suffering the disaster of being exploited. However, I reject this interpretation of Kavka's discussion because presumably war is itself a disastrous outcome. I believe Kavka would have argued that one cannot follow the disaster avoidance principle in anarchy because one cannot avoid the disaster of war. Below I also maintain that the

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press



C = cooperate, *A* = Anticipate, Party 1 is a moderate
Type *I* = moderate, Type *II* = dominator

FIGURE 2.2. Parties' preference rankings in a Hybrid of Assurance and Assurance Dilemma.

Hobbes, and later Kavka, conclude that the parties' uncertainty regarding party types they meet is a key reason the Hobbesian State of Nature converges to war. However, this uncertainty suggests an alternate hybrid game-theoretic model of the Hobbesian State of Nature that neither Kavka nor any of Hobbes' other readers consider. As noted above, when a moderate encounters another party, the *actual* structure of this encounter is either Assurance or Assurance Dilemma, depending upon the other's type. Figure 2.2 summarizes the situation a moderate faces when she meets another party. Moderate Party 1, who is uncertain as to Party 2's type, considers the possibilities as if the type of party she meets is selected for her by "Nature." If Party 2 is of Type *I*, he is also a moderate and the game is Assurance. If Party 2 is of Type *II*, he is a dominator and the game is Assurance Dilemma. In other words, the moderate's predicament has the structure of a *game of incomplete information* (Harsanyi 1967, 1968*a*, 1968*b*), where each individual in anarchy belongs to one of two *payoff types*, the moderate-type and the dominator-type.[15] I will call

Assurance Dilemma model tacitly assumes that agents should base their decisions upon unrealistically pessimistic probabilities (16–17).

[15] I first proposed modeling the Hobbesian State of Nature as a game of incomplete information in a referee report of an early version of Dodds and Shoemaker (2002). In

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

this game π-*Assurance*, where π is Party 1's probability that Party 2 is of Type *I*. π-Assurance is a probabilistic hybrid of the Assurance and the Assurance Dilemma games. What should moderate Party 1 do in this game? The answer to this question is indeterminate. For either cooperation or Anticipation could be Party 1's better option, depending upon the specific payoffs Party 1's preferences generate and the probabilities Party 1 ascribes over Party 2's possible types.[16] Note that Kavka's Assurance Dilemma is a special case of π-Assurance where $\pi = 0$. Indeed, this is my main reason for rejecting the Assurance Dilemma as a hybrid game-theoretic model of anarchy. The Assurance Dilemma model presupposes that a moderate's uncertainty over party types should lead her to act as if she is sure to meet only Type *II* parties in her encounters. But why should moderate parties be so pessimistic? Even if Party 1 cannot distinguish the dominators from the other moderates, she might believe that she will seldom encounter a dominator. Consequently, her value for π will be relatively high. If π is sufficiently high, then cooperation can be Party 1's best response given what she expects Party 2's to choose if Party 2 is of Type *I*.[17] The π-Assurance model accommodates this possibility, and

the final version of their essay, Dodds and Shoemaker consider this possibility, but argue that in general a moderate should proceed as if he were in an Assurance Dilemma because parties in general will lack common knowledge of their types (2002: 357–60). However, their argument misses the point, since Harsanyi proposed his formal model of games of incomplete information for the express purpose of analyzing games where players do not know their counterparts' payoffs and hence have uncertainty over their types.

[16] A dominator is also faced with a game of incomplete information. However, the analysis of this game is trivial. If the dominator's counterpart is a moderate, then they are in an Assurance Dilemma, while if the dominator's counterpart is also a dominator, then they are in a Prisoners' Dilemma. So *A* is the dominator's strictly dominant strategy in this game of incomplete information since *A* is the dominator's unique best strategy given any probability distribution he has over his counterpart's types.

[17] Indeed, Kavka admits as much when he points out that in a Hobbesian State of Nature with both a low population density and a low percentage of dominators, cooperation is probably a better security strategy than Anticipation (1986: 175). However, Kavka evidently believes that such conditions would seldom if ever obtain in a Hobbesian State of Nature.

To compute the conditions under which C_1 is Party 1's best alternative, we need to examine Party 1's cardinal payoffs in π-Assurance. Let Party 1's payoffs be defined by the matrix

		Party 2	
		C_2	A_2
Party 1	C_1	x	0
	A_1	$1-\alpha$	1

C = cooperate, A = Anticipate, $x > 1-\alpha, \alpha \geq 0, k = i, j$

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

indeed accommodates any possible beliefs a moderate party might have regarding her counterpart's payoff types and alternative strategies.

The π-Assurance game can have equilibria where both parties cooperate as well as the equilibrium where both Anticipate. This game also exposes certain difficulties that arise when one tries to predict war or peace in anarchy via *a priori* analysis alone. In order to identify the equilibria of the π-Assurance game, one needs to know the payoffs for both parties at each possible outcome and the moderate Party 1's probability distribution over Party 2's payoff types. Even if one has this much information, this does not imply that one can now identify some outcome of the game as the "solution" the parties will follow. One cannot predict which equilibrium of the game the parties will follow, or even that they follow an equilibrium at all, unless one assumes that they have common knowledge of their rationality, the complete payoff structure, and their probabilities over each other's strategies and payoff types. Where would all this common knowledge come from? Beyond all this, the game of incomplete information derived from the Figure 2.2 preferences characterizes a pairwise encounter between a *particular* pair of parties in anarchy. Whatever one might ultimately conclude by analyzing this game of incomplete information need not characterize the behavior of *all* of the parties in anarchy. For there are no good reasons to suppose that the moderates or the dominators in anarchy share a single payoff function over alternative outcomes because they share a preference set over these outcomes. And there are no good reasons to suppose that the parties in anarchy share a unique probability distribution over the types and the strategies of their counterparts.[18]

let P_1, E_1 and u_1 denote Party 1's probability distribution, expectation operator and payoff function, respectively, and let M denote the state where Party 2 is of Type I. Then $P_1(M) = \pi$, and

$$E_1(u_1(C_1)) = \pi(x \cdot P_1(C_2|M) + 0 \cdot P_1(A_2|M)) + (1-\pi)(x \cdot P_1(C_2|M^c) + 0 \cdot P_1(A_2|M^c))$$
$$= \pi \cdot (x \cdot P_1(C_2|M)).$$

Similarly

$$E_1(u_1(D_1)) = \pi(1+\alpha) \cdot P_1(C_2|M) + \pi(1 - P_1(C_2|M)) + (1-\pi).$$

We have $E_1(u_1(C_1)) > E_1(u_1(D_1))$ when

$$\pi \cdot P_1(C_2|M) > \frac{1}{x-\alpha}.$$

[18] In other words, I deny that the parties in anarchy would have a *common prior* probability distribution over the possible states that characterize their situation. A common prior is assumed in some of the literature on the foundations of game theory, but I concur with many game theorists that it is doubtful that the common prior assumption obtains for actual human agents engaged in a game.

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

		Party j: C	Party j: A
Party i	C	(x_i, x_j)	$(0, 1+\alpha_j)$
	A	$(1+\alpha_i, 0)$	$(1, 1)$

$C =$ cooperate, $A =$ Anticipate, $x_k > 1, \alpha_k \geq 0, k = i, j$

FIGURE 3.1. Hobbesian state of nature interaction.

In short, I believe a more flexible game-theoretic model is necessary, one in which payoffs and probabilities vary across the individuals in anarchy over ranges, rather than across discrete types only. I believe that one can develop such a model by relaxing some of the assumptions of the π-Assurance model and by considering repeated interactions in the Hobbesian State of Nature. This opens the door to a dynamical analysis of anarchy based upon how parties learn over time. I propose such a model in the following section.

3. A VARIABLE ANTICIPATION THRESHOLD MODEL

This section introduces the *Variable Anticipation Threshold* model of interaction in anarchy.[19] A population of n parties interact repeatedly over discrete time periods. At each interaction period, each Party i can meet one or more *partner* parties. When parties meet, each either cooperates (C) or Anticipates (A). The possible outcomes of an interaction between partners are defined by the 2-party payoff matrix of Figure 3.1. The Figure 3.1 matrix characterizes a family of 2×2 games whose payoffs vary according to the parameters $x_i > 1$, $\alpha_i \geq 0$, $i \in \{1, \ldots, n\}$. Most of the game-theoretic literature on interactions between members of a large population assumes

[19] The Variable Anticipation Threshold model presented here is a special kind of agent-based model in which agents enter into a strategic form game when they interact. An anonymous referee alerted me to the work of Epstein and Axtell (1996), who investigate other agent-based models in which populations are heterogeneous with respect to characteristics such as resource holdings. Epstein and Axtell use their models to explore social processes such as the emergence of trade as well as conflicts between groups. The Variable Anticipation Threshold model complements their agent-based models, as I use this model to investigate the prospects for cooperation or war in anarchy.

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

all parties have identical payoffs. As discussed above in Section 2, this assumption is unlikely to be compatible with the preferences of all in an actual group of people in anarchy. The hybrid models of Section 2 employ two different payoff types, but these models are still insufficiently general because they do not reflect the relative intensity of preferences that different individuals might have over alternative outcomes. In the Variable Anticipation Threshold-model, the variable parameters reflect differing intensities of preferences via the variable payoffs. One can interpret the parameters of an interaction defined by the Figure 3.1 matrix as follows: $x_i - 1$ reflects how much Party i prefers peace over war. α_i is Party i's "temptation parameter," as it determines how much Party i desires the gains of exploiting another party. If $x_i > 1 + \alpha_i$, Party i is a moderate, if $x_i = 1 + \alpha_i$, Party i is a *weak dominator*, and if $1 + \alpha_i > x_i$, Party i is a *strict dominator*.[20] If two partners are both strict dominators, they play a Prisoners' Dilemma. If both are moderates, they play an Assurance game. If one is a moderate and the other a strict dominator, they play an Assurance Dilemma. So for a moderate, the payoffs of an encounter with a given partner in anarchy are like those of a π-Assurance game where the moderate is uncertain as to whether or not her partner is also a moderate. However, as we shall see, the Variable Anticipation Threshold model does not assume all of the common knowledge needed for an *a priori* analysis of the strategies parties follow in the π-Assurance model.

In the Variable Anticipation Threshold model, there can be up to n distinct payoff functions, one for each party. At each time period, each party ordinarily chooses a strategy that maximizes expected payoff.[21] $u_i : \{C, A\}^2 \to \mathbb{R}$ denotes Party i's payoff function, $\lambda_i(t)$ denotes Party i's probability that a partner Anticipates at period t, and $E_i^t(\cdot)$ denotes Party i's expectation operator at period t. Note that Party i's probabilities and expected payoffs can vary over time. We have

$$E_i^t(u_i(A)) = (1 - \lambda_i(t)) \cdot (1 + \alpha_i) + \lambda_i(t) = 1 + \alpha_i - \lambda_i(t)\alpha_i$$

and

$$E_i^t(u_i(C)) = (1 - \lambda_i(t)) \cdot x_i = x_i - \lambda_i(t)x_i.$$

[20] If $x_i = 1 + \alpha_i$, then Anticipation is Party i's *weakly dominant strategy* since Party i always does at least as well by following A rather than C and does strictly better if Party j follows A.

Note that the symmetry of the payoffs in the Figure 3.1 matrix when $x_i = x_j$ and $\alpha_i = \alpha_j = 0$ does not imply that agents' payoffs are fully comparable according to a fixed objective scale. The payoffs are scaled in this manner to simplify the exposition, and reflect the relative intensity of each agents' preferences over the various alternative outcomes.

[21] Below we will consider an exception, namely, when a party occasionally deviates from following a best response by mistake or because she is experimenting (pp. 23–5).

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

$\lambda_i^* = \max\{0, \frac{x_i - 1 - \alpha_i}{x_i - \alpha_i}\}$ is Party i's *Anticipation threshold*, since $E_i^t(u_i(A)) > E_i^t(u_i(C))$ exactly when $\lambda_i(t) > \lambda_i^*$. Note that if Party i is a weak or strict dominator, $\lambda_i^* = 0$.

The epistemic assumptions of the Variable Anticipation Threshold model are quite weak compared to the mutual knowledge assumptions implicit in Kavka's Assurance Dilemma model and the common knowledge assumptions of the π-Assurance model. Each party in the Variable Anticipation Threshold model is rational, and knows her own payoffs. But this model does not assume a party knows the payoffs of any of her partners or even that her partners are necessarily payoff maximizers like herself. If the payoff functions and the rationality of the parties were common knowledge, then moderates could adopt the contingency strategy of cooperating only with other moderates. Similarly, if the identities of parties who Anticipate against cooperators were common knowledge, then parties could adopt conditionally cooperative strategies similar to "tit for tat" or "trigger" that characterize Nash equilibria of an indefinitely repeated Prisoners' Dilemma played out by a fixed pair of players.[22] However, there are no mechanisms in the Variable Anticipation-threshold model for generating any common knowledge among the parties. Indeed, creating such mechanisms would be just the sort of cooperative endeavor Hobbes expressly denies is possible while people remain in anarchy (*Leviathan* 13:9). The Variable Anticipation Threshold model does not assume that parties can identify which of their partners are dominators or have a history of exploiting others. This is in line with Hobbes' belief that the parties in anarchy are unable to distinguish between the types of partner parties they encounter.

The strategies the parties in the Variable Anticipation model follow depend crucially upon how they form their probabilities over their partners' strategies. Intuitively, one would expect a party's probability that a current partner will Anticipate to increase monotonically with the number of his past partners who Anticipated. Game theorists model this type of inductive learning with *fictitious play processes* (Brown 1951).[23] In a fictitious play process, at each period each agent forms a prediction of the

[22] In an indefinitely repeated Prisoners' Dilemma, a player who follows "tit for tat" cooperates in the first period and then at each tth period where $t > 1$ imitates the counterpart's behavior in the $t - 1$st period. A player who follows "trigger" cooperates at each period so long as the counterpart has never Anticipated, and if the counterpart Anticipates for the first time at period t' the "trigger" agent Anticipates for each period $t > t'$.

[23] Brown introduced the first fictitious play process as a method of computing Nash equilibria in certain games. Fictitious play gets its name because according to Brown's original interpretation, players mentally simulate each other's actions and update their beliefs over a sequence of imaginary plays. More recently, game theorists have used fictitious play processes as models of how agents can learn to follow equilibria in games as the

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

others' strategies based on historical frequency of past strategies, and then follows a strategy that maximizes expected payoff given the prediction. Laboratory experiments where subjects repeatedly play 2×2 coordination games, including the Assurance game, show that *weighted* fictitious play is a good approximation of their learning processes (Cheung and Friedman 1997; Nyarko and Schotter 2002). The Variable Anticipation-threshold model incorporates a generalization of weighted fictitious play. At each period t, if Party i has at least one partner she updates her probability $\lambda_i(t)$ according to the following rule:

$$(3.1) \quad \lambda_i(t+1) = \frac{\sum_{k=0}^{t} \theta_i^k A_i(t-k)}{\sum_{k=0}^{t} \theta_i^k}$$

where $A_i(0) = \lambda_i(0)$ is Party i's prior probability that a partner Anticipates, $A_i(m)$ is the fraction of Party i's partners who Anticipate at period $m > 0$, and θ_i where $0 \leq \theta_i \leq 1$ is Party i's weighting factor by which she discounts the past. If Party i has no partners at period t then $\lambda_i(t+1) = \lambda_i(t)$. This updating rule generalizes standard fictitious play for 2×2 games, which assumes that each inductive learner is matched in the game with exactly one other player at each period. If $\theta_i = 1$, then Party i does not discount the past at all and (3.1) becomes a special case of the *Dirichlet rule* for inductive updating (Skyrms 1991). If $\theta_i = 0$, then Party i discounts all of the past completely save for the most recent period, and (3.1) becomes a special case of the *Cournot best response* dynamic (Camerer 2003). Experimental evidence suggests that human learners are neither so forgetful as are Cournot updaters nor so insensitive to the recent past as are Dirichlet updaters. In their coordination game experiments, Cheung and Friedman found median values for θ_i across subjects between 0.25 and 0.50 (Cheung and Friedman 1997; Camerer 2003: 284). The specific cases of the Variable Anticipation Threshold model presented in this paper use distributions of the discount factors that are compatible with these experimental results.

To summarize thus far, the Variable Anticipation Threshold model contains three main sets of parameters, namely, the Anticipation thresholds λ_i^*, $i \in \{1, \ldots, n\}$ that are derived from the family of payoff functions, and the priors and discount factors $\lambda_i(0)$, θ_i, $i \in \{1, \ldots, n\}$ that determine how the parties learn from experience. The probabilities $\lambda_i(t)$, $i \in \{1, \ldots, n\}$ form a dynamical system that evolves as a function of the discount factors, the priors and the strategies followed by each party's partners. To complete the Variable Anticipation Threshold model, an account of how parties meet their partners is needed. Most of the game-theoretic literature assumes

result of *actual* repeated plays. This is the interpretation of fictitious play processes I adopt in this paper.

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press



White parties are moderates that follow C. Dark gray parties are moderates that follow A.

Light gray cells are unoccupied.

FIGURE 3.2. Initial state of a population of all moderates who update beliefs according to weighted fictitious play.

either that each player in a game interacts exclusively with a fixed set of counterpart players or that at each time period players are matched with new counterparts selected completely at random from a large population. Neither of these extremes is a good approximation of how individuals in anarchy are likely to interact. The Variable Anticipation Threshold model defines each party's partners as those parties that are sufficiently near her, and allows parties to migrate over time and encounter new partners that were a relatively short distance away. More specifically, in the Variable Anticipation Threshold model parties occupy cells in a two-dimensional lattice. At time period t, Party i's partners are the parties in the cells immediately adjacent to Party i's cell in the lattice. At period $t + 1$, Party i may move if possible to a new cell in the lattice, particularly in order to move away from partners who Anticipated at period t.

To illustrate how one can use the Variable Anticipation Threshold model to analyze the emergence of war or peace in anarchy, I will summarize the results of several computer experiments. More detailed descriptions of these and other experiments appear in the Appendix. Figure 3.2 summarizes the initial state of a computer simulation of the interactions among a population of 1256 parties in a 50×50 2-dimensional lattice. In this simulation, the parties' payoff function parameters were

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

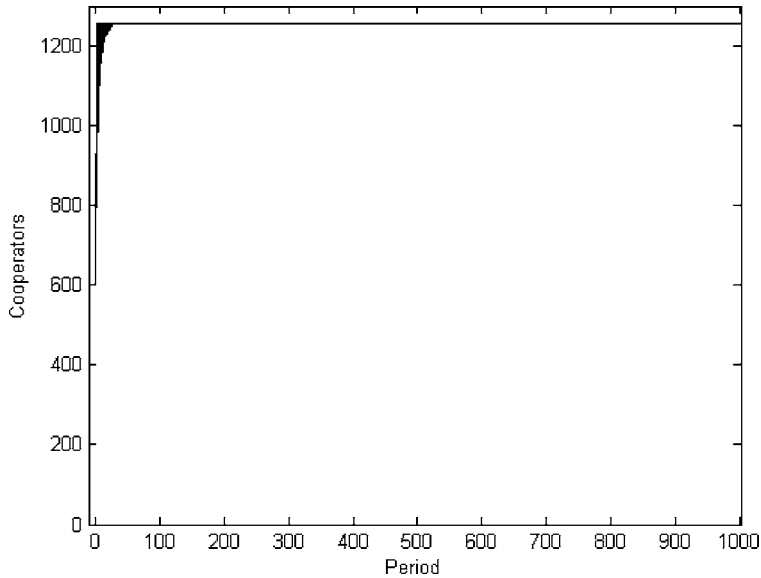


FIGURE 3.3. Frequency of *C*-followers in population of all moderates who update beliefs according to weighted fictitious play.

distributed so that $x_i > 1 + \alpha_i$ for all $i \in \{1, \ldots, n\}$, that is, every party was a moderate, and the priors were distributed so that initially approximately half of the parties Anticipated. The fictitious play weights were distributed with a median value 0.375, the midpoint of the median values found in experimental studies with human subjects. At period $t + 1$, if any of Party i's neighbors Anticipated at period t, Party i would migrate to a neighboring unoccupied cell if possible. Otherwise, Party i would migrate with probability 0.25 to a new cell if Party i neighbored this cell at period t and this cell was unoccupied at period t. These migration rules reflect the idea that parties will definitely flee from Anticipators if they can, and otherwise will move at random. The system was allowed to evolve for 1000 periods or *generations*. As the graph in Figure 3.3 shows, the system rapidly converged to a state where all parties follow *C*, even though slightly more than half of them began by following *A*. This system converged to the equilibrium of peace because every party was a moderate. Each Party i's Anticipation threshold λ_i^* was bounded away from 0, and consequently the initially high dispersion of Anticipators was rapidly overwhelmed by the spread of the *C* strategy as the parties learned through their experience with their partners.

A second experiment tests the stability of the all-*C* equilibrium in a population of all moderates. In this simulation, 1257 parties occupied sites

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

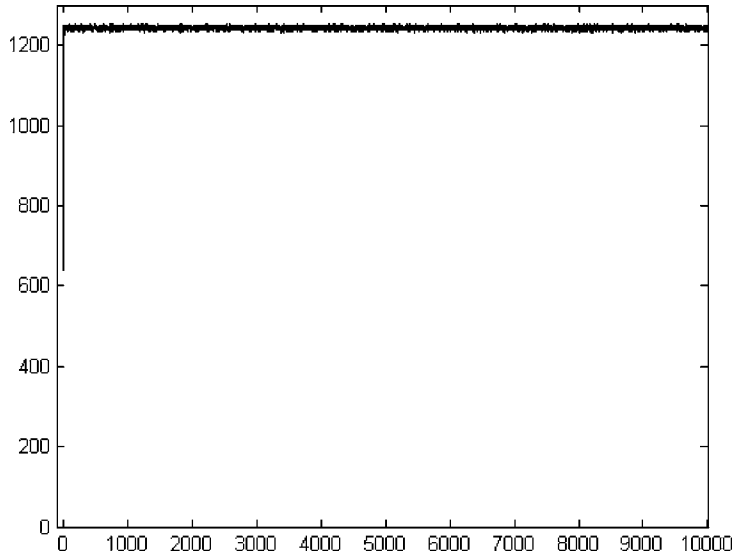
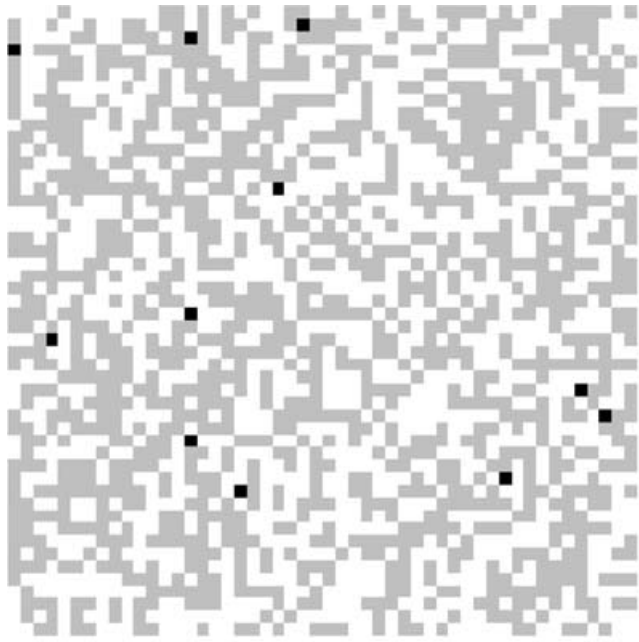


FIGURE 3.4. Frequency of *C*-followers in population of all moderates who update beliefs according to weighted fictitious play with mutation at rate $\epsilon = 0.2$.

in a 50×50 2-dimensional lattice, and their payoffs again were distributed so that all were moderates. Once more the priors were distributed so that initially approximately half Anticipated. Again the migration rules had each party move if possible when a neighbor Anticipated, and otherwise each party would move if possible with probability 0.25. However, in this simulation parties were somewhat prone to deviate from their best responses given their beliefs. More specifically, at each generation 2% of the parties followed either *C* or *A* at random instead of following their best responses to their beliefs formed by weighted fictitious play. One can interpret these random *mutations* as mistakes or experiments. This system was allowed to evolve for 10,000 generations. Figure 3.4 illustrates the frequency of *C*-followers in the system. Even though the system was continually bombarded by a relatively high 2% rate of mutation, the parties rapidly gravitated towards the all-*C* equilibrium and remained near this equilibrium throughout the duration of the simulation. This and the previous example illustrate a general phenomenon: *When all in the population are moderates, universal cooperation is an attracting equilibrium of the dynamical system*. A population of rational individuals in anarchy can converge to a state of peace, provided that all of them are moderates whose best outcome is the outcome of reciprocal cooperation.

What happens when dominators are present in the population? Figure 3.5 summarizes the initial state of a population of 1256 parties in

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

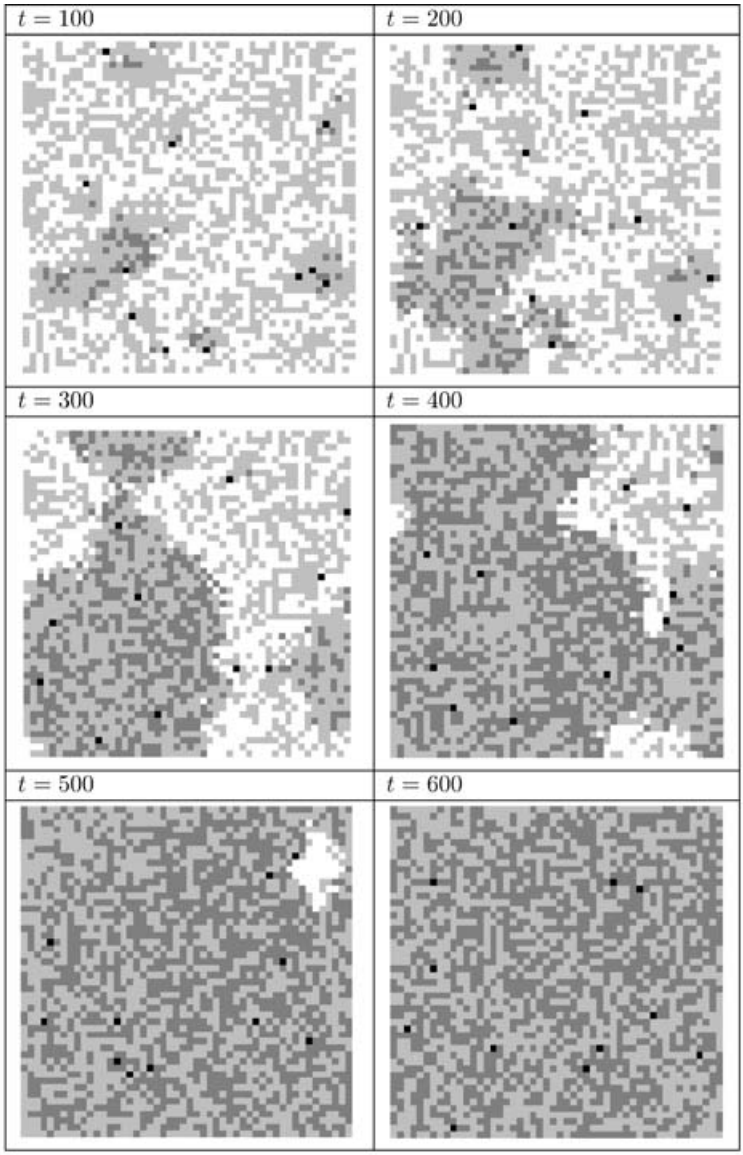


White parties are moderates that follow C. Black parties are dominators.
Light gray cells are unoccupied.

FIGURE 3.5. Initial state of a mixed population who update beliefs according to weighted fictitious play.

a two-dimensional lattice. This time, the priors were all set at $\lambda_i(0) = 0$, that is, each party initially had a zero probability for the event that any partner Anticipates. Once again, at each successive time period, a party moved if possible if a neighbor Anticipated in the previous time period and otherwise moved with probability 0.25 if possible. However, in this population the payoff functions were such that approximately 0.88% of the parties were strict dominators. Initially, each moderate followed C given his prior beliefs, while each dominator followed A, his strictly dominant strategy. This system was allowed to evolve for 1000 generations. Even though nearly all of the parties began by cooperating, the A-strategy gradually spread through and overtook the population. Figure 3.6 shows the state of this population at various generations. Figure 3.7 summarizes the proportion of cooperators in the population over time. In fewer than 600 generations, the entire system had converged to the all-A equilibrium even though fewer than 1% of the parties were dominators.

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press



White parties are moderates that follow C. Dark gray parties are moderates that follow A. Black gray parties are dominators. Light gray cells are unoccupied.

FIGURE 3.6. States of a mixed population who update beliefs according to weighted fictitious play over time.

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

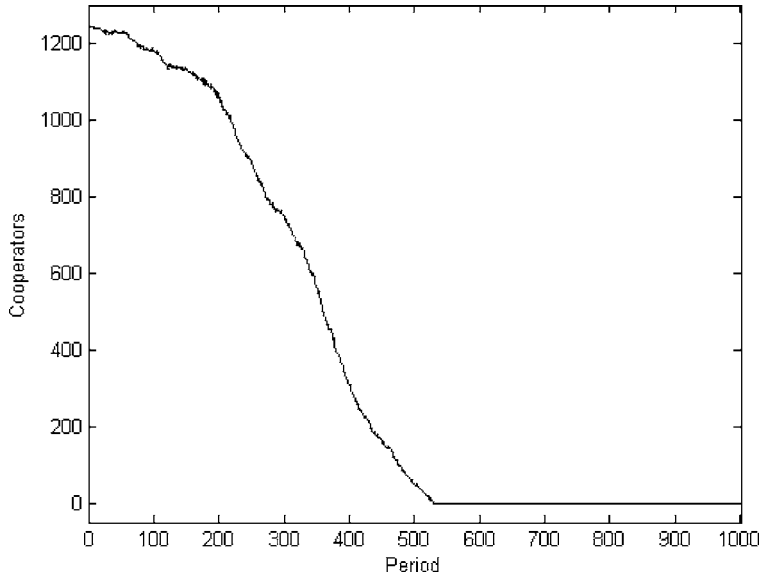


FIGURE 3.7. Frequency of *C*-followers in a mixed population who update beliefs according to weighted fictitious play.

In a fourth simulation, 1268 parties in a two-dimensional lattice again began with priors all set at $\lambda_i(0) = 0$. These parties followed the same migration rules as in the previous experiments. Approximately 0.71% of the parties in this population were strict dominators. In this population at every generation parties mutated at the rate of 2%. Figure 3.8 summarizes the proportion of cooperators in the population over 10,000 generations. The system rapidly gravitated to the all-*A* equilibrium of war and remained near this equilibrium despite the high rate of mutation. These latter two examples illustrate another general phenomenon: *When some in the population are strict dominators, universal Anticipation is an attracting equilibrium of the dynamical system, while universal cooperation is never an equilibrium*.

A variety of additional experiments test the sensitivity of the Variable Anticipation model to variations in the parameters. The properties of the Variable Anticipation model proved surprisingly robust to changes in the parameters that govern population size, density, movement, mutation and partnership. The results of all of these experiments are summarized by Tables A.1–A.4 in the Appendix. In every simulation where all the moderates in the population initially followed *C*, when the payoffs were distributed so that a small percentage of the population were strict dominators, the system converged fairly rapidly to the all-*A*

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

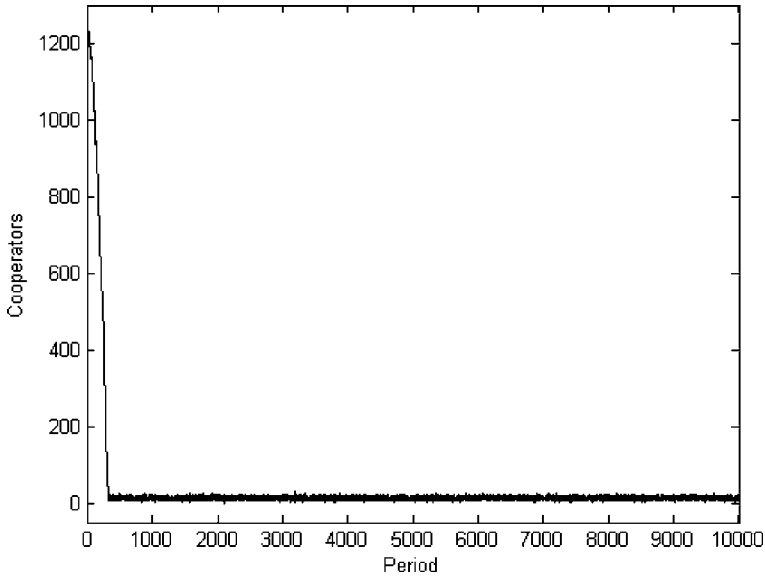


FIGURE 3.8. Frequency of *C*-followers in a mixed population who update beliefs according to weighted fictitious play with mutation at rate $\epsilon = 0.2$.

equilibrium even if the dominators received no "help" from moderates who occasionally mutated to *A*. When parties could mutate, populations with dominators present would always remain near the all-*A* equilibrium even if the system was "bombarded" steadily by mutation rates as high as $\epsilon = 0.02$. On the other hand, in each simulation where the payoffs were distributed so that the parties' Anticipation thresholds were all bounded away from 0, so that all were moderates, when parties deliberately moved away from Anticipators the system would always remain near the all-*C* equilibrium even with the relatively high mutation rate of $\epsilon = 0.02$. So for a wide range of model parameters, the presence of dominators are both a sufficient and a necessary condition to destabilize the all-*C* state and drive the system to the all-*A* equilibrium of war.

In general, in the populations where all are moderates, cooperation can spread throughout the system even if a large percentage of the population begin by Anticipating and parties mutate at a high rate. All-*C* is an attracting equilibrium of such systems because Anticipation is no party's best strategy in the long run. Every party eventually learns to follow *C* excect for the generations in which she mutates because she encounters sufficiently many cooperators over time to convert her to follow *C* as her best response. If a party initially Anticipates or

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

Anticipates because of a mutation, her neighbors will try to move away from her but they will not persist in doing so because this party will not persist in Anticipating. When all parties are moderates, all can learn to cooperate. On the other hand, in the populations containing some strict dominators, the moderates gradually convert to Anticipation because of the activity of these strict dominators. While strict dominators might form only a tiny part of the population in such systems, they always Anticipate unless they mutate because for them Anticipation is strictly dominant. Parties might try to move away from these dominators, but they cannot avoid them in the long run, so the parties near the strict dominators suffer *persistent* Anticipation. Since parties cannot distinguish persistent Anticipators from others in the system, they cannot adopt contingency strategies that treat persistent Anticipators differently from the others. The moderate parties that begin by cooperating convert to Anticipation when sufficiently many of their neighbors persistently Anticipate. At first, moderates convert to Anticipation only when they are in the proximity of strict dominators whom they cannot flee. Gradually the consistent Anticipation of the strict dominators creates "clusters" of neighboring parties who have converted to Anticipation because their updated $\lambda_i(t)$ probabilities now exceed their Anticipation thresholds. These clusters of Anticipators in effect form a critical mass of converts to Anticipation, since they grow and produce a "contagion" of Anticipation that overtakes the entire population. In populations containing any strict dominators, the all-*C* state of peace cannot be an equilibrium of the corresponding dynamical system. Indeed, given the distribution of Anticipation thresholds in this population, the only equilibrium of this system is the all-*A* equilibrium of war. A population of rational individuals in anarchy might converge to Hobbes' war of all against all even if only a few of them are strict dominators and all of the moderates cooperate at first. In effect, a few "bad apples" can spoil the entire "barrel"![24]

4. CONCLUSION

One tradition in political philosophy tries to deduce what sort of society emerges as the result of rational choices in particular circumstances. Hobbes is a giant of this tradition, and his analysis of anarchy is a cornerstone of his rational choice theory of the social contract. But as we have seen, given the variable preferences across people that Hobbes explicitly accepts, the war of all against all in anarchy need not inevitably follow from the individuals' *a priori* rational decisions. A second tradition

[24] Thanks to Brian Skyrms for suggesting this metaphor.

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

in political philosophy analyzes how society might evolve as its members modify their behavior in light of experience. Hume was one of the first to propose this kind of dynamical analysis of society. Hume argues that injustice breeds more injustice, as people tend to follow the lead of those they encounter who commit acts of injustice. In this essay, I have used the Variable Anticipation threshold model to explore how the interactions of a population in anarchy might evolve. Special cases of this model show that when individuals learn from experience and modify their behavior accordingly, the presence of a few dominators can lead others to imitate their example, as Hume predicted, so that they all ultimately end up at war, as Hobbes predicted. The key elements of the Variable Anticipation Threshold model are the variable payoffs that determine differential Anticipation thresholds across the population, the spatial structure and rules that determine the movement of individuals, and the learning rules that underwrite the dynamical system of the individuals' expectations. This kind of dynamical model has a number of advantages over the static models of anarchy that are more common in the literature. The variable payoffs evidently reflect actual human psychology, since in real human societies some, but not all, similarly situated individuals exploit others. A variable payoffs model is faithful to the texts of classical political philosophers such as Hobbes and Locke, who allow for the possibility that some in anarchy will act like moderates and some like dominators. The existing static models of anarchy either incorporate unrealistically pessimistic assumptions regarding human nature, as does the Assurance Dilemma model, or require unrealistically strong assumptions regarding parties' knowledge and then still yield indeterminate conclusions, as does the π-Assurance model. A dynamical model can account for how Anticipation spreads in anarchy as individuals learn from their interactions. In the Variable Anticipation Threshold model, when dominators are present the population converges to the equilibrium of the Hobbesian war simply because the individual parties migrate across a spatial structure and update their beliefs according to a simple inductive rule. Future work should explore the behavior of parties who migrate according to more complex movement rules and who update their beliefs according to more sophisticated learning rules. Yet the simple Variable Anticipation Threshold model of anarchy suggests that the route to vindicating Hobbes' claim that anarchy inevitably leads to war is not Hobbes' own *a priori* approach but the dynamical approach Hume proposed informally.

The Variable Anticipation threshold model also suggests a promising route to resolving the dispute between the Hobbesians and the Lockeans. The presence of criminals in every civil society is strong evidence that some dominators will be present in any sizable population. Hobbes explicitly acknowledges this, and nothing Locke says suggests he would deny this.

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

Again, I maintain that Locke and Hobbes are not really at odds over their accounts of human psychology. I believe the real crux of the dispute between the Lockeans and Hobbesians is the set of assumptions one makes regarding common knowledge in anarchy, not psychology. If all are moderates and this is common knowledge, then it is also common knowledge that they can sustain the peace of universal cooperation. If moderates and dominators are present in the population and the identities of the moderates and the dominators are common knowledge, then it is common knowledge that parties can follow an equilibrium where each moderate follows the contingency strategy of cooperating exactly with the other moderates. If one assumes this much common knowledge, then the Lockeans are right. But the Hobbesians would deny that people in anarchy can acquire this common knowledge. If individuals do not know enough about others to enable them to distinguish between moderates and dominators, then they cannot follow contingency strategies that depend on party type. Such individuals must rely upon what they learn over time from direct experience interacting with others. The Variable Anticipation Threshold model shows that war can be inevitable when individuals in anarchy base their strategies upon inductive learning and even a few dominators are present. When individuals cannot identify the types of parties they encounter, a small number of dominators can spark the war of all against all that Hobbes predicted. Of course, common knowledge and complete lack of knowledge of party types are endpoint cases. Future work should address intermediate cases where the agents in anarchy have some ability to distinguish between the types of parties they meet that fall short of their having common knowledge of party types. Such work may shed light upon the precise conditions that make peace in anarchy or civil society possible.

APPENDIX: COMPUTER SIMULATIONS

The simulations explore how behavior evolves in particular instances of the Variable Anticipation Threshold model, where parties who neighbor each other in a 2-dimensional lattice play games with payoff functions that conform with the Figure 3.1 matrix.[25] The lattice has m^2 sites, where the user defines m, each of which has an index (j, k), $j, k \in \{1, \ldots, m\}$. In all simulations summarized in this paper, $m = 50$, which defines a lattice with 2,500 sites.[26] Topologically the lattice is equivalent to a torus, and

[25] The simulations were programed and run in MatLab 7.0.4. The MatLab programs used to run these simulations are available from the author upon request.

[26] I explored several simulations with larger lattices such as the 10,000 site lattice defined when $m = 100$. However, these simulations require far more computer resources and are harder to interpret visually than the simulations summarized here where $m = 50$. In these larger simulations I attempted, I obtained results similar to those where $m = 50$.

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

can be mapped onto a square whose edges "wrap around." The program simulates a sequence of M generations of play, or *run* of length M, between neighboring agents in the lattice who update their beliefs according to the weighted fictitious play formula (3.1). At the start of a run, a set of agents are randomly distributed in the lattice as follows: For each (j, k), $j, k \in \{1, \ldots, m\}$, a value $U_1(j, k)$ is sampled from the uniform distribution over [0, 1] and site (j, k) is initially occupied if $U_1(j, k) < \rho$, where the user chooses the density $\rho \in [0, 1]$. Hence there are $n \leq m^2$ agents in the system, and $n = m^2$ if $\rho = 1$. X_1, X_2 and X_3 denote independent normally distributed random variables with mean 0 and variance 1, and ν_α, ν_β, ν_θ and μ_θ are constants chosen by the user. For each Party i, α_i was sampled from the random variable $Y = \nu_\alpha \cdot \text{abs}(X_1)$, $x_i = 2 + \beta_i$ where β_i was sampled from the random variable $Z = \nu_\beta \cdot X_2$, and θ_i was sampled from the random variable $W = \nu_\theta X_3 + \mu_\theta$. The user can also set the prior $\lambda_i(0)$ and the initial strategy $s_i(0)$ for each party according to the formula

$$s_i(0) = \begin{cases} C & \text{if } U_2(j_i, k_i) \leq \sigma \\ A & \text{if } U_2(j_i, k_i) > \sigma \end{cases}$$

where $U_2(j_i, k_i)$ is sampled from the uniform distribution over [0, 1], (j_i, k_i) is Party i's site at $t = 0$ and $\sigma \in [0, 1]$ is chosen by the user.

If Party i occupies site (j_i, k_i) at a given period t, then the set

$$\mathcal{M}^t(j_i, k_i) = \{\{j, k\} : |j_i - j| = 1 \mod m \text{ and/or } |k_i - k| = 1 \mod m\}$$

is Party i's *Moore-8 neighborhood*, and the set

$$\mathcal{V}^t(j_i, k_i) = \{\{j, k\} : |j_i - j| = 1 \mod m \text{ or } |k_i - k| = 1 \mod m\}$$

is Party i's *von Neumann neighborhood*. The user can choose to have each Party i play all of the other parties that lie in either Agent's Moore-8 neighborhood or Party i's von Neumann neighborhood.

In each run, parties in the lattice can migrate according to either of two rules chosen by the user. According to the first rule, at the end of each period t, if a party in $\mathcal{M}^t(j_i, k_i)$ (or $\mathcal{V}^t(j_i, k_i)$, depending upon the user's choice) followed A and at least one cell in $\mathcal{M}^t(j_i, k_i)$ (or $\mathcal{V}^t(j_i, k_i)$) is unoccupied then at period $t + 1$ Party i migrates into of these unoccupied cells. Otherwise, if a cell in $\mathcal{M}^t(j_i, k_i)$ (or $\mathcal{V}^t(j_i, k_i)$) is unoccupied then Party i migrates into this cell with probability $\frac{\eta}{8}$(or $\frac{\eta}{4}$) where η is chosen by the user. The second migration rule is simply the second part of the first rule. Parties that migrate according to the second rule simply move about in the lattice at random. Parties that migrate according to the first rule deliberately move away from Anticipators and otherwise move at random. In the special case where $\rho = 1$, every cell in the lattice is occupied and the network of interactions is uniform and fixed. The fixed network games of this special case are similar to the network games analyzed in Nowak and May (1992), Nowak, May and Bonhoeffer (1994), and Vanderschraaf and

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

Alexander (2005), although in these earlier studies all the agents in the uniformly connected graph have the same payoff matrix.

In the simulations summarized in this paper, the fictitious play weights were distributed normally with a median value $\mu_\theta = 0.375$ and a standard deviation $\nu_\theta = 0.125$. These parameters ensured that the fictitious play weights of the parties were concentrated along the interval of fictitious play weights compatible with the experimental studies of coordination games played by human subjects (Cheung and Friedman 1997; Nyarko and Schotter 2002). A variant of the model uses *perturbed* weighted fictitious play, where with some probability ϵ, an party *mutates* or chooses a strategy at random instead of best responding given current beliefs. One can interpret a mutation as a mistake or an experiment. Mutations are assumed to be independent and identically distributed across players. One can use the perturbed fictitious play to test the stability of equilibria, in a manner analogous to the stability analyses of equilibria for updating rules such as Cournot best response dynamics explored in works such as Ellison (1993, 2000) and Young (1998).

Tables A.1–A.4 summarize the results of a variety of simulations of interaction in the Hobbesian State of Nature. Run 1.1 is the first simulation of a population of all moderates discussed in Section 3. A population of 1256 moderates interacted in the 50^2 sites of the lattice. In this simulation, $\nu_\alpha = 0.20$ and $\nu_\beta = 0.10$. Consequently, in this system the parties' payoff function parameters were distributed so that $x_i > 1 + \alpha_i$ for all $i \in \{1, \ldots, 1256\}$. The initial strategies were distributed so that initially 47.93% of the parties cooperated and 52.07% Anticipated. Parties migrated according to the rule where each party moved if possible when any of its Moore-8 neighbors Anticipated and otherwise migrated at random if possible with $\eta = 0.25$. By the end of the rum of $M = 1000$ generations, 0.0% of the parties Anticipated. Run 1.3 is the first simulation of a mixed population discussed in Section 3. A mixed population of 1256 interacted in the 50^2 sites of the lattice. This time the priors were all set at $\lambda_i(0)$. Once again, at each successive time period parties actively moved away from Anticipators if possible and otherwise moved at random with $\eta = 0.25$. However, here $\nu_\alpha = 0.35$ and $\nu_\beta = 0.10$, so in this population the payoff functions were distributed so that $1 + \alpha_i > x_i$ for approximately 0.88% if the $i \in \{1, \ldots, 1256\}$. That is, approximately 0.88% of the parties were strict dominators. At the start of the run of $M = 1000$ generations, only 0.88% of the population Anticipated, namely, the strict dominators. By the end of this run, 100.0% of the parties in this system Anticipated.

Tables A.1 and A.2 summarize simulations where parties moved according to the rule where parties deliberately fled Anticipators when possible. Tables A.3 and A.4 summarize simulations where parties moved at random only. Tables A.1 and A.3 (Tables A.2 and A.4) summarize simulations where parties interact with counterparts who lie in their

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

Run	M	ρ	n	ν_α	ν_β	σ	ϵ	η	D	$A(1)$	$A(M)$
1.1	1000	0.50	1256	0.20	0.1	0.5	0.0	0.25	0.0	52.07	0.0
1.2	10,000	0.50	1257	0.20	0.1	0.5	0.02	0.25	0.0	49.09	0.072
1.3	1000	0.50	1256	0.35	0.1	1.0	0.0	0.25	0.88	0.88	100.0
1.4	10,000	0.50	1268	0.35	0.1	1.0	0.02	0.25	0.71	1.03	98.97
1.5	10,000	0.50	1198	0.20	0.1	1.0	0.05	0.25	0.0	2.25	3.01
1.6	1000	0.25	608	0.35	0.1	1.0	0.0	0.25	0.49	0.49	100.0
1.7	1000	0.75	1859	0.35	0.1	1.0	0.0	0.25	0.75	0.75	100.0
1.8	1000	0.90	2231	0.35	0.1	1.0	0.0	0.25	0.77	0.77	100.0
1.9	1000	1.00	2500	0.35	0.1	1.0	0.0	0.25	0.56	0.56	0.60
1.10	1000	1.00	2500	0.45	0.1	1.0	0.0	0.25	3.44	3.44	100.0
1.11	1000	0.5	1255	0.35	0.1	1.0	0.01	0.0	0.72	1.04	99.36
1.12	1000	0.5	1225	0.35	0.1	1.0	0.01	0.5	0.65	0.90	99.27
1.13	1000	0.5	1258	0.35	0.1	1.0	0.01	1.0	0.87	1.51	99.60
1.14	1000	0.5	1261	0.30	0.2	1.0	0.01	0.25	0.71	1.59	99.44
1.15	1000	0.5	1268	0.25	0.3	1.0	0.01	0.25	0.79	1.50	99.53

$A(1)$ = percentage of parties that follow A in period 1.
$A(2)$ = percentage of parties that follow A in period M.
D = percentage of dominators in the population.

TABLE A.1. Simulations where parties interact with neighbors in $\mathcal{M}^t(j_i, k_i)$ (parties deliberately move away from anticipators if possible).

Run	M	ρ	n	ν_α	ν_β	σ	ϵ	η	D	$A(1)$	$A(M)$
2.1	1000	0.5	1221	0.20	0.1	0.5	0.0	0.25	0.0	50.78	0.0
2.2	10,000	0.5	1233	0.20	0.1	0.5	0.02	0.25	0.0	50.28	2.27
2.3	1000	0.5	1271	0.35	0.1	1.0	0.0	0.25	0.55	0.55	100.0
2.4	10,000	0.5	1230	0.20	0.1	1.0	0.02	0.25	0.89	1.06	99.11
2.5	10,000	0.5	1267	0.20	0.1	1.0	0.02	0.25	0.0	1.18	1.34
2.6	1000	0.25	610	0.35	0.1	1.0	0.0	0.25	0.49	0.49	100.0
2.7	1000	0.75	1862	0.35	0.1	1.0	0.0	0.25	0.70	0.70	100.0
2.8	1000	0.90	2235	0.35	0.1	1.0	0.0	0.25	0.67	0.67	100.0
2.9	1000	1.0	2500	0.35	0.1	1.0	0.0	0.25	0.56	0.68	0.68
2.10	1000	1.0	2500	0.45	0.1	1.0	0.0	0.25	3.64	3.64	100.0
2.11	1000	0.5	1221	0.35	0.1	1.0	0.01	0.0	0.82	1.31	99.67
2.12	1000	0.5	1226	0.35	0.1	1.0	0.01	0.5	0.49	0.82	99.51
2.13	1000	0.5	1244	0.35	0.1	1.0	0.01	1.0	0.88	1.37	99.44
2.14	1000	0.5	1260	0.30	0.2	1.0	0.01	0.25	0.71	0.95	99.37
2.15	1000	0.5	1255	0.25	0.3	1.0	0.01	0.25	1.59	1.91	99.52

$A(1)$ = percentage of parties that follow A in period 1.
$A(2)$ = percentage of parties that follow A in period M.
D = percentage of dominators in the population.

TABLE A.2. Simulations where parties interact with neighbors in $\mathcal{V}^t(j_i, k_i)$ (parties deliberately move away from anticipators if possible).

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

Run	M	ρ	n	ν_α	ν_β	σ	ϵ	η	D	$A(1)$	$A(M)$
3.1	1000	0.50	1184	0.20	0.1	0.5	0.0	0.25	0.0	49.07	0.0
3.2	10,000	0.50	1283	0.20	0.1	0.5	0.02	0.25	0.0	48.56	1.40
3.3	1000	0.50	1244	0.35	0.1	1.0	0.0	0.25	0.40	0.40	100.0
3.4	10,000	0.50	1264	0.35	0.1	1.0	0.02	0.25	1.03	1.66	98.66
3.5	10,000	0.50	1267	0.20	0.1	1.0	0.05	0.25	0.0	1.82	2.76
3.6	1000	0.25	614	0.35	0.1	1.0	0.0	0.25	0.65	0.65	100.0
3.7	1000	0.75	1891	0.35	0.1	1.0	0.0	0.25	0.53	0.53	100.0
3.8	1000	0.90	2238	0.35	0.1	1.0	0.0	0.25	0.54	0.54	100.0
3.9	1000	1.00	2500	0.35	0.1	1.0	0.0	0.25	0.68	0.68	0.68
3.10	1000	1.00	2500	0.45	0.1	1.0	0.0	0.25	3.00	3.00	100.0
3.11	1000	0.5	1280	0.35	0.1	1.0	0.01	0.0	0.86	1.09	14.14
3.12	1000	0.5	1229	0.35	0.1	1.0	0.01	0.5	0.49	0.90	99.84
3.13	1000	0.5	1292	0.35	0.1	1.0	0.01	1.0	0.54	1.01	99.61
3.14	1000	0.5	1240	0.30	0.2	1.0	0.01	0.25	0.73	1.37	99.27
3.15	1000	0.5	1271	0.25	0.3	1.0	0.01	0.25	1.18	1.81	99.37

$A(1)$ = percentage of parties that follow A in period 1.
$A(2)$ = percentage of parties that follow A in period M.
D = percentage of dominators in the population.

TABLE A.3. Simulations where parties interact with neighbors in $\mathcal{M}^t(j_i, k_i)$ (parties move at random).

Run	M	ρ	n	ν_α	ν_β	σ	ϵ	η	D	$A(1)$	$A(M)$
4.1	1000	0.50	1253	0.20	0.1	0.5	0.0	0.25	0.0	49.40	0.0
4.2	10,000	0.50	1310	0.20	0.1	0.5	0.02	0.25	0.0	50.23	98.63
4.3	1000	0.50	1259	0.35	0.1	1.0	0.0	0.25	0.64	0.64	100.0
4.4	10,000	0.50	1252	0.35	0.1	1.0	0.02	0.25	0.32	0.96	99.20
4.5	1000	0.50	1226	0.20	0.1	1.0	0.02	0.25	0.0	0.65	79.45
4.6	1000	0.25	626	0.35	0.1	1.0	0.0	0.25	1.28	1.28	100.0
4.7	1000	0.75	1894	0.35	0.1	1.0	0.0	0.25	0.53	0.53	100.0
4.8	1000	0.90	2272	0.35	0.1	1.0	0.0	0.25	0.44	0.44	100.0
4.9	1000	1.00	2500	0.35	0.1	1.0	0.0	0.25	0.60	0.60	0.64
4.10	1000	1.00	2500	0.45	0.1	1.0	0.0	0.25	3.40	3.40	11.64
4.11	1000	0.5	1218	0.35	0.1	1.0	0.01	0.0	0.82	1.15	42.86
4.12	1000	0.5	1282	0.35	0.1	1.0	0.01	0.5	0.16	0.39	99.45
4.13	1000	0.5	1241	0.35	0.1	1.0	0.01	1.0	0.81	1.13	99.92
4.14	1000	0.5	1275	0.30	0.2	1.0	0.01	0.25	0.55	0.94	99.29
4.15	1000	0.5	1207	0.25	0.3	1.0	0.01	0.25	1.16	1.66	99.50

$A(1)$ = percentage of parties that follow A in period 1.
$A(2)$ = percentage of parties that follow A in period M.
D = percentage of dominators in the population.

TABLE A.4. Simulations where parties interact with neighbors in $\mathcal{V}^t(j_i, k_i)$ (parties move at random).

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

Moore-8 neighborhoods (von Neumann neighborhoods). With only a single exception (the mutation rates in $t = 5$), in each run $1.t$, $2.t$, $3.t$ and $4.t$, $t \in \{1, \ldots, 15\}$, the other parameters of the simulation are the same for each t in order to compare the properties of the dynamics across movement rules and interaction rules. In general, in these simulations the presence of dominators was always a sufficient condition to produce a contagion of Anticipation, and dominators were needed in order to produce this contagion in all but the 4.5 simulation discussed in the following paragraph.

The runs where $t = 5$ test the stability of the all-*C* equilibrium when no dominators are present in the population. In these runs all are moderates and all start by following *C* except for mutants. In both runs where parties interact with partners in their Moore-8 neighborhoods, the system remains near the all-*C* equilibrium over 10,000 generations even at a mutation rate of $\epsilon = 0.05$! In run 2.5, where parties interacted with partners in their von Neumann neighborhoods and tried to move away from Anticipators, they remained near the all-*C* equilibrium over 10,000 generations at the relatively high mutation rate of $\epsilon = 0.02$. But in run 4.5, where parties interacted with partners in their von Neumann neighborhoods and moved at random only, with $\epsilon = 0.2$ the system approached the all-*A* equilibrium after than fewer than 1000 generations. In this case, the all-*C* equilibrium could not withstand the high "bombardment" of mutants for long. These runs show that the stability of the all-*C* equilibrium is somewhat sensitive to rules governing partnership and movement. When all are moderates, the all-*C* equilibrium is quite stable when parties have the relatively many partners who lie in their Moore-8 neighborhoods or actively flee Anticipators. But when they have the relatively few partners who lie in their von Neumann neighborhoods and they move at random only, the all-*C* equilibrium is less stable in the face of random mutation.

Note that in runs where $t = 6$, 7 or 8 where population density ρ was varied across values less than or greater than $\rho = 0.5$, the system always converged to the all-*A* equilibrium where the percentage of Anticipators was $A(M) = 100.0$ even though in general the percentage of dominators was $D < 1.0$. So the attracting power of the all-*A* equilibrium is fairly robust to changes in population density, so long as some movement is possible. In the runs where $t = 9$, $\rho = 1.0$ and the system would remain near the all-*C* equilibrium even though some dominators were present. This is not surprising when one takes into account the fact that in these systems parties could not migrate, so that the few dominators are "trapped" in the lattice and cannot spread their aggressive behavior. On the other hand, in several of the runs where $t = 10$, the all-*A* equilibrium took over the system because $\nu_\alpha = 0.45$ even though $\rho = 1.0$ once more. So it is possible in a fixed network of parties with no migration for the all-*A* equilibrium to take over even when all moderates initially cooperate, so long as there

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

are sufficiently many strict dominators. The runs where $t = 11$, 12 and 13 suggest that if dominators are present, the attracting power of the all-A equilibrium is robust with respect to changes in the probability η of random movement. In the runs where $t = 14$ and 15, the values of ν_α and ν_β are varied. In these runs strict dominators are again present and the systems all quickly move to near the all-A equilibrium. Again, it is the fact that dominators are in the system and not the specific parameter values that introduce the dominators that drives the system near all-A. (In the runs where $t = 11 - 15$ the systems do not converge exactly to all-A because the mutation rate is set at $\epsilon = 0.01$.)

REFERENCES

Aumann, Robert. 1974. Subjectivity and correlation in randomized strategies. *Journal of Mathematical Economics* 1:67–96.

Axelrod, Robert. 1984. *The evolution of cooperation*. Basic Books.

Barry, Brian. (1965) 1990. *Political argument: a reissue with a new introduction*. University of California Press.

Bernheim, B. Douglas. 1984. Rationalizable strategic behavior. *Econometrica* 52:1007–28.

Binmore, Ken. 1994. *Playing fair*. MIT Press.

Brandenburger, Adam and Dekel, Eddie. 1988. The role of common knowledge assumptions in game theory, in *The economics of missing markets, information and games*, ed. Frank Hahn. Clarendon Press:46–61.

Brown, G. W. 1951. Iterative solutions of games by fictitious play. In *Activity analysis of production and allocation*, ed. T. C. Koopmans. Wiley.

Camerer, Colin. 2003. *Behavioral game theory: experiments in strategic interaction*. Russell Sage Foundation, and Princeton University Press.

Cheung, Yin-Wong and Friedman, Daniel. 1997. Individual learning in normal form games: some laboratory results. *Games and Economic Behavior* 19:46–76.

Curley, Edwin. 1994. Introduction to Hobbes' *Leviathan*. In *Leviathan*, ed. Edwin Curley. Hackett Publishing Company; *viii–xlvii*.

Dodds, Graham and Shoemaker, David. 2002. Why we *can't* all just get along: human variety and game theory in Hobbes's State of Nature. *Southern Journal of Philosophy* 40:345–74.

Ellison, Glenn. 1993. Learning, local interaction and cooperation. *Econometrica* 61:1047–71.

Ellison, Glenn. 2000. Basins of attraction and long run equilibria. *Review of Economic Studies* 67:17–45.

Epstein, Joshua and Axtell, Robert. 1996. *Growing artificial societies: social science from the bottom up*. Brookings Institute Press.

Gauthier, David. 1979. Thomas Hobbes: moral theorist. *Journal of Philosophy* 76:547–59.

Hampton, Jean. 1986. *Hobbes and the social contract tradition*. Cambridge University Press.

Harsanyi, John. 1967. Games with incomplete information played by "Bayesian" players, I: The basic model. *Management Science* 14:159–82.

Harsanyi, J. 1968a. Games with incomplete information played by "Bayesian" players, II: Bayesian equilibrium points. *Management Science* 14:320–4.

Harsanyi, J. 1968b. Games with incomplete information played by "Bayesian" players, III: The basic probability distribution of the game. *Management Science* 14:486–502.

Hobbes, Thomas. (1640) 1998. *De Cive (On the Citizen)*, trans. and ed. Richard Tuck and Michael Silverthorne. Cambridge University Press.

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press

Hobbes, Thomas. (1651) 1991. *Leviathan*, ed. Richard Tuck. Cambridge University Press.
Hume, David. (1740) 2000. *A treatise of human nature*, ed. David Fate Norton and Mary J. Norton. Oxford University Press.
Hume, David. (1777) 1998. *An enquiry concerning the principles of morals: a critical edition*, ed. Tom Beauchamp. Clarendon Press.
Jenkins, J. J. 1967. Locke and Natural rights. *Philosophy* 152–3.
Jiborn, Magnus. 1999. *Voluntary coercion: collective action and the social contract*. Doctoral Thesis, Department of Philosophy, Lund University.
Kavka, Gregory. 1983. Hobbes's War of all against all. *Ethics* 93:291–310.
Kavka, Gregory. 1986. *Hobbesian moral and political theory*. Princeton University Press.
Kavka, Gregory. 1989. Political contractarianism. unpublished manuscript.
Kavka, Gregory. 1995. The rationality of rule-following: Hobbes's dispute with the foole. *Law and Philosophy* 14:5–34.
Lewis, David. 1969. *Convention: a philosophical study*. Harvard University Press.
Locke, John. (1690) 1988. *Two treatises of government*, ed. Peter Laslett. Cambridge University Press.
Macpherson, C. B. 1962. *The political theory of possessive individualism*. Oxford University Press.
Nash, John. 1950. Equilibrium points in *n*-person games. *Proceedings of the National Academy of Sciences of the United States* 36:48–49.
Nash, John. 1951. Non-cooperative games. *Annals of Mathematics* 54:286–95.
Nowak, Martin A. and May, Robert M. 1992. volutionary games and spatial chaos. *Nature* 359:826–9.
Nowak, Martin A., Bonhoeffer, S. and May, Robert M. 1994. Spatial games and the maintenance of cooperation. *Proceedings of the National Academy of Sciences of the USA*, 91:4877–81.
Nyarko, Yaw and Schotter, Andrew. 2002. An experimental study of belief learning using elicited beliefs. *Econometrica*, 70:971–1005.
Pearce, David. 1984. Rationalizable strategic behavior and the problem of perfection. *Econometrica* 52:1029–50.
Rawls, John. (1971) 1999. *A theory of justice, revised edition*. Harvard University Press.
Schelling, Thomas. 1960. *The strategy of conflict*. Harvard University Press.
Schelling, Thomas. 1978. *Micromotives and macrobehavior*. W. W. Norton and Company.
Sen, Amartya. 1967. Isolation, assurance and the social rate of discount. *Quarterly Journal of Economics* 81.
Sen, Amartya, 1973. *On economic inequality*. Clarendon Press.
Simmons, A. John. 1989. Locke's state of nature. *Political Theory* 3:449–70.
Simmons, A. John. 1993. *On the edge of anarchy: Locke, consent and the limits of society*. Princeton University Press.
Skyrms, Brian. 1991. Inductive deliberation, admissible acts, and perfect equilibrium. In *Foundation of decision theory*, ed. M. Bacharach and S. Hurley. Blackwell: 220–41.
Vanderschraaf, Peter. 2001. *Learning and coordination*. Routledge.
Vanderschraaf, Peter and Alexander, J. McKenzie. 2003. Follow the leader: local interactions with influence neighborhoods. *Philosophy of Science*: 72:86–113.
Young, H. Peyton. 1998. *Individual strategy and social structure*. Princeton University Press.

https://doi.org/10.1017/S0266267106000897 Published online by Cambridge University Press