Graduate School of
Business Administration

KOBE
UNIVERSITY

ROKKO KOBE JAPAN

2014-9

Leadership in the Prisoner's Dilemma With Inequity-Averse Preferences

Koji Abe Hajime Kobayashi Hideo Suehiro

Discussion Paper Series

Leadership in the Prisoner's Dilemma with Inequity-Averse Preferences*

Koji Abe† *Yokohama National University*
Hajime Kobayashi‡ *Kansai University*
Hideo Suehiro§ *Kobe University*

May 9, 2014

Abstract

We consider the economic consequences of fairness concerns under the freedom to choose the timing of moves by developing a new economic theory of leadership. We study the prisoner's dilemma in which players are endowed with Fehr and Schmidt preferences with inequity aversion as their private information and then choose cooperation or defection once at one of two timings that they prefer. In this model, we consider an equilibrium in which a leader–follower relationship endogenously emerges as a consequence of players' heterogeneous inequity aversions. We present three results. First, we provide a sufficient condition for the existence of a leadership equilibrium. Then, we present a comparative statics analysis of the equilibrium. Finally, we investigate who takes the leadership, depending on the game parameters. We provide a characterization of the equilibrium leadership patterns. These results also hold when agents can choose the timing of moves from more than two timings.

JEL Classification: C72, D03, D82

Keywords: Leadership, Endogenous Timing, Prisoner's Dilemma, Inequity Aversion

*We are grateful to Eduard Faingold, Ernst Fehr, Benjamin Hermalin, Hideshi Itoh, Shinsuke Kambe, Takashi Kunimoto, Eiichi Miyagawa, Yasuyuki Miyahara, Daniele Nosenzo, Hideo Owan, Larry Samuelson, Klaus Schmidt, and Stephanie Wang for their helpful comments. We thank conference and seminar participants at NASM 2012, EEA/ESEM 2012, GAMES 2012, a seminar at the University of München, the GCOE workshop at Osaka University, the Kyoto–Hitotsubashi Game Theory Workshop 2013, and the mini–conference on economics of leadership at Hitotsubashi University. Koji Abe, Hajime Kobayashi, and Hideo Suehiro gratefully acknowledge financial support from the Japan Society for the Promotion of Science under JSPS KAKENHI Grant Numbers 24730166, 24530207, and 24530198, respectively.

†Faculty of International Social Sciences, Yokohama National University, 79-4 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan; abe@ynu.ac.jp.

‡Faculty of Economics, Kansai University, 3-3-35, Yamate, Suita, Osaka, 564-8680, Japan; khajime@kansai-u.ac.jp.

§Graduate School of Business Administration, Kobe University, 2-1, Rokkodai, Nada, Kobe, Hyogo, 657-8501, Japan; suehiro@kobe-u.ac.jp.

1 Introduction

1.1 Introduction

A considerable body of evidence by the experimental research in recent decades indicates that many individuals are concerned with fairness (Fehr and Schmidt (2006), Cooper and Kagel (2013)). This pushes economists towards examining how fairness concerns affect the consequences of economic transactions. The research so far has focused on the implications of the fact that fairness-concerned individuals may choose different alternatives than self-interested individuals in exogenous move games.[1] This paper addresses another issue that fairness concerns may influence economic outcomes for the reason that fairness-concerned individuals may choose to take actions at different timings than self-interested individuals when they have the freedom to choose when to move.[2]

Economists recognize that fairness concerns can affect economic transactions through time. For example, conditional cooperation is one of the central ideas that fairness concerns could cause individuals to behave differently from self-interest (Gächter (2007)). Timing is an essential element in the attitude paraphrased as "if you cooperate, I will cooperate." Researchers have found that a considerable percentage of people exhibit the conditional cooperation attitude when placed in an "after you" position (Fischbacher, Gächter and Fehr (2001), Herrmann and Thöni (2009)).

However, when agents have the freedom to choose when to move, the conditional cooperation is in vain unless the "after you" position is realized as an equilibrium play. Including conditional cooperation, it remains unexplored how fairness concerns affect economic transactions by effectively changing individual behaviors taken within a span of time under the freedom to choose when to move.

We develop a theory of one such mechanism. In particular, we consider the issue of time and social dilemmas. We show that if agents are concerned with fairness, the freedom to choose the timing of moves could resolve certain social dilemmas. If individuals are concerned with fairness, conditional cooperation has the potential to realize cooperation when they move sequentially. The issue is how to have agents move sequentially under the freedom to choose the timing of moves and how to have an earlier mover commit to cooperation, which is necessary for later cooperation by conditional cooperators. We show that when agents are divergent in their fairness concerns and there are enough conditional cooperators, they sort themselves into earlier and later movers by their own will, and earlier movers commit to cooperation so that later

[1] Several alternative approaches to modeling fairness concerns have been proposed. One of the promising models is that of Fehr and Schmidt (1999), which we employ in this paper. Studies of choice based on this model under exogenously given sequences of moves extend to various fields, including (1) ultimatum games (Fehr and Schmidt (1999)), (2) moral hazard (Fehr and Schmidt (2004), Itoh (2004), Demougin and Fluet (2006), Demougin, Fluet and Helm (2006), Dur and Glazer (2008), Rey-Biel (2008), Englmaier and Wambach (2010), Neilson and Stowe (2010)), (3) partnerships (Bartling and von Siemens (2010a)), (4) teams (Li (2009)), (5) tournaments (Grund and Sliwka (2005), Bartling and von Siemens (2010b), Dubey, Geanakoplos, and Haimanko (forthcoming)), (6) contract design (Fehr, Klein, and Schmidt (2007)), (7) ownership (Fehr , Kremhelmer, and Schmidt (2008)), and (8) adverse selection (Desiraju and Sappington (2007) and Rasch, Wambach, and Wiener (2012)).

[2] Time is an essential ingredient of many economic transactions. In many natural settings, agents move without an order of moves being formally provided and the consequences of economic transactions are critically influenced by the choice of timings of moves by the agents in equilibrium. Under the classical assumption of self-interested agents, economists have explored how agents guide themselves to move within a span of time and how the freedom to choose timings of moves affects the efficiency of transactions. Examples are the English auction (Milgrom and Weber (1982)), bargaining (Perry and Reny (1993)), and joint projects (Marx and Matthews (2000)).

cooperation is realized by later movers who are conditional cooperators. This happens because agents with different fairness concerns have different degrees of incentives to induce conditional cooperation from others. Thus, fairness concerns play a double role in resolving social dilemmas: the role of generating conditional cooperation and the role of realizing the sequence of moves necessary for conditional cooperation to come into effect through leadership.

In our analysis, we specifically consider the prisoner's dilemma as a prototype social dilemma. We hypothesize that a player has Fehr and Schmidt (1999) preferences with inequity aversion over the outcomes of the prisoner's dilemma. Furthermore, we assume a variety of preferences among players, under which players may differ in their sensitivities to inequity measured by an envy parameter α and a guilt parameter β. We consider a Bayesian model of the prisoner's dilemma with endogenous moves. A player is endowed with his type, which is defined by his envy parameter and guilt parameter (α, β). A type is a realization of a continuous random variable and it is his private information. Each player must choose C (cooperation) or D (defection). There are two timings for moves: timing 1 and timing 2. Each player makes his choice between C and D once at one of the two timings and which timing he selects is at his discretion.

We examine a particular Bayesian strategy, which we will call a three-mode strategy. It divides player's types into three categories: leader types, defector types, and conditional cooperator types. A leader type chooses C at timing 1, whereas a defector type and a conditional cooperator type wait.[3] A defector type and a conditional cooperator type differ in their responses to their opponent's behaviors at timing 1. The former chooses D irrespective of his opponent's behaviors, while the latter responds to C with C (and with D otherwise.)

If a three-mode strategy prevails as an equilibrium and a leader type is matched with a conditional cooperator type, then the leader type commits to C at timing 1 and then the conditional cooperator type responds with C after postponing his choice until timing 2. The cooperative outcome (C, C) is realized along an equilibrium path by having players move sequentially by their own will. In this equilibrium, the on-path behavior of the leader type is a leadership behavior (that is, a behavior of taking the leadership).

We show three results. First, we present a sufficient condition over a set of payoff parameters of a prisoner's dilemma for a three-mode strategy to be a sequential equilibrium of the game. The condition has a simple and straightforward interpretation; a pure materialist who feels no envy and no guilt ($\alpha = \beta = 0$) has an incentive to lead by choosing C at timing 1 as long as he expects that his opponent postpones his choice until timing 2 with certainty. For any prisoner's dilemma that supports this incentive, there exists a sequential equilibrium in a three-mode strategy, in which the pure materialist himself may or may not become a leader type.

Second, we examine a comparative statics. Suppose that a payoff from the cooperative outcome becomes higher or a payoff from the defection outcome (D, D) becomes lower. Then, we can say that the cooperation becomes less difficult, because the incentive to choose D over C against an opponent's choice of C is weaker under a higher cooperative payoff; the incentive to choose D over C against an opponent's choice of D is weaker under a lower defection outcome payoff; and the degree of the Pareto improvement from the defection outcome to the cooperative outcome is larger in either case. We show that the leadership behavior is more likely to emerge in a prisoner's dilemma

[3] As we mention in the Discussion section, a defector type does not need to wait in general. Here, we consider the simplest case to illustrate our equilibrium strategy.

with less difficulty of cooperation than in a prisoner's dilemma with more difficulty. Formally, the maximum value of the probabilities that the prior assigns to the leader type in all the sequential equilibria in a three-mode strategy is higher in the prisoner's dilemma with less difficulty of cooperation than in the prisoner's dilemma with more difficulty.[4] The same is true for the minimum value of the leader-type probabilities under the sufficient condition for the existence of a three-mode equilibrium.

Third, we characterize our equilibrium in terms of who takes the leadership. We introduce the notion of an incentive to lead and define a way to compare the strength of the incentive to lead across types. A type (α, β) who has the strongest incentive to lead is a representative type in the set of leader types in a given prisoner's dilemma. We identify the type who has the strongest incentive to lead, and we show how the type varies according to a change in the parameters of the prisoner's dilemma. The pure materialist ($\alpha = \beta = 0$) has the strongest incentive to lead in a prisoner's dilemma with a higher difficulty of cooperation, while a particular type who has modest fairness concerns (modestly high α and β) has the strongest incentive to lead in a prisoner's dilemma with a lower difficulty of cooperation. Finally, we study how the whole set of leader types vary according to a change in the parameters of the prisoner's dilemma. We show that, as the cooperation becomes less difficult, the set of leader types in a corresponding prisoner's dilemma consists of more fairness–concerned types.

Furthermore, we show that these three results also hold when agents can choose the timing of moves from more than two timings. The mechanism of leadership elucidated by our analysis is also valid and our conclusion on the resolution of social dilemmas remains unchanged. Our theory of leadership applies to a prisoner's dilemma with many timings in general.

The main contribution of this paper is to consider the issue of time and social dilemmas by taking into account the fact that many individuals are concerned with fairness. Our theory is a first attempt to provide a mechanism by which fairness concerns affect the consequences of economic transactions under the freedom to choose the timing of moves.

Our theory advances our understanding of the role that fairness concerns play in dynamic transactions. For example, recent experimental results indicate that some social dilemmas could be resolved when individuals have the freedom to choose when to move. Arbak and Villeval (2013) conduct experiments of two-stage voluntary contribution games with a similar time structure to our game, in which each subject is requested to make a choice of contribution level at one of two timings that he prefers.[5] The data display that some of the subjects move with high levels of contributions in advance of the others and some of those subjects who postpone their choices to timing 2 respond with high levels of contributions.[6] In light of the importance for economists

[4]We show that there may exist multiple equilibria in a three-mode strategy, depending on a set of payoff parameters.

[5]The game of Arbak and Villeval (2013) is a standard voluntary contribution game, in which payoffs are symmetric across agents and zero contribution is the dominant strategy. This game involves a social dilemma similar to the prisoner's dilemma. Nosenzo and Sefton (2011), Rivas and Sutter (2011) and Préget, Nguyen, and Willinger (2012) also conduct similar experiments. The results of Rivas and Sutter (2011) and Préget, Nguyen, and Willinger (2012) are less relevant to this paper because they use the repeated-game setting for their experiments. The experiments of Arbak and Villeval (2013) and Nosenzo and Sefton (2011) are implemented in the stranger setting and so the subjects in the experiments face a social dilemma game in the same setting as this paper. However, Nosenzo and Sefton (2011) differ from Arbak and Villeval (2013) in that their game is a version of the Varian (1994) game, in which payoffs are asymmetric across agents and there exists strategic substitution effects.

[6]See Figure 1 and Table 3 in Arbak and Villeval (2013).

to understand how human beings can and do overcome social dilemmas, their findings urge economists to develop a theory that explains the role of the freedom to choose the timing of moves in resolving social dilemmas. Our theory provides a possible theoretical interpretation for what forces resolve the dilemma in their experiments because their findings are compatible with our equilibrium prediction for the prisoner's dilemma that leading by cooperation occurs with positive probabilities and responding to cooperation with cooperation occurs with positive probabilities, depending on players' types.

Our approach to the issue of time and social dilemmas has several features to be noted. First, we employ the Fehr and Schmidt (1999) model for fairness concerns. This model, which has proved powerful in explaining various experimental results, is one of the simplest models proposed to express other regarding preferences. The tractability of this model enables us to develop not only the condition for equilibrium existence (the first result) but also two fundamental results: the comparative statics on leadership probabilities (the second result) and the characterization of equilibrium in terms of leader's types (the third result). These results can be used to induce a set of testable predictions or hypotheses based on our theory for the purpose of testing the theory in future experiments. These results may also provide theoretical insights when one considers the issue of time and other forms of social dilemmas.[7]

Second, we introduce incomplete information into fairness concerns by agents. If a player knows that his opponent is endowed with the fairness concerns that make the opponent a conditional cooperator, it is obviously in the interest of the player to lead by taking C at timing 1 and induce the opponent to take C at timing 2. Therefore, under complete information of fairness concerns by agents, there is no obstacle for them to resolve a social dilemma by leadership.[8] In reality, the way in which people respond to leadership by others is diverse and it is often observed that a leadership is betrayed by a follower with uncooperative behaviors. This makes it difficult for people to resolve a social dilemma by leadership. How this obstacle is overcome by agents is the central issue that must be studied to understand the mechanism by which fairness concerns work in resolving a social dilemma under the freedom to choose the timing of moves. Hence, it is essential for our theory to consider incomplete information about fairness concerns by agents.

There are a few papers that discuss incomplete information under the Fehr and Schmidt (1999) preferences. Fehr and Schmidt (1999), (2004), Fehr, Klein, and Schmidt (2007), and Fehr, Kremhelmer, and Schmidt (2008) argue that their experimental findings on the ultimatum, the multitask, the contract design, and the ownership games can be explained if they assume that subjects are endowed with one of two or four kinds of parameters in the Fehr and Schmidt (1999) preferences with particular probabilities, and that a set of parameters for a subject is his private information.[9]

In contrast to these experimental studies, we examine the whole set of continuous densities over the domain of the Fehr and Schmidt (1999) preferences. Then, we develop our three results in terms of a set of payoffs and a density of the preferences. As Fehr, Kremhelmer, and Schmidt (2008) state, the heterogeneity of fairness concerns is a well–established fact and thus *the question is not whether theory should incorporate*

[7]For example, if we confine ourselves to prisoner's dilemmas, we can show that very different behaviors are realized in simultaneous moves and exogenous sequence moves. Our theory helps us understand why and how they differ. See our Discussion.

[8]We will elaborate on this point in the Discussion.

[9]They conduct this exercise of choosing particular distributions because their purpose is to show that the theory based on Fehr and Schmidt (1999) preferences helps us interpret and better understand the observed data patterns. See footnote 16 in Fehr, Klein, and Schmidt (2007).

heterogeneous social preferences but which distribution of preferences theory should assume.[10] In this respect, this paper is one of the first attempts to study games played by Fehr–Schmidt players under general incomplete information and provide robust results by associating them with conditions on type distribution.

Finally, our theory investigates a three-mode strategy that assigns three different behavioral modes to the corresponding three classes of Fehr and Schmidt (1999) preferences. The Fehr and Schmidt (1999) model is a very simple model with a two-dimensional type space: envy parameter and guilt parameter. Fehr and Schmidt (2004), Fehr, Klein, and Schmidt (2007), and Fehr, Kremhelmer, and Schmidt (2008) successfully illustrate that even such a simple model is enough to explain the experimental data for fairly complex games, such as the multitask, the contract design, and the ownership with two behavioral modes corresponding to two classes of Fehr and Schmidt (1999) preferences.

In our game, however, it can be shown that cooperation by leadership cannot be supported in equilibrium if we confine a Bayesian strategy to a class that admits only two behavioral modes. Our first result shows that the Fehr and Schmidt (1999) model successfully generates three behavioral modes, which support the cooperation in equilibrium. This indicates that in spite of its limited type space, the Fehr and Schmidt (1999) model may be powerful and useful enough to explain diverse behaviors in complex games.[11]

The rest of this paper is organized as follows. We review the related literature in Subsection 1.2. Then, we begin our analysis in Section 2 with an illustrative numerical example to demonstrate the idea of the leadership mechanism driven by fairness concerns. Section 3 presents the formal model of the prisoner's dilemma played through two timings by players with inequity-averse preferences. Section 4 explains the notion of a three-mode equilibrium in this game. Section 5 states the first result on the existence of a three-mode equilibrium. Section 6 states the second result on the comparative statics of the three-mode equilibrium. Section 7 states the third result on the characterization of leadership patterns with respect to fairness concerns. Section 8 shows that the three-results developed in a prisoner's dilemma with two timings also hold in a prisoner's dilemma with many timings in general. Section 9 discusses some issues that we excluded from our analysis. Appendix A supplements Section 5 with respect to the exact bound on the game parameters for the existence of a three-mode equilibrium. Furthermore, when this exact bound applies, the results in Sections 6 and 7 become much sharper. Appendix B supplements Section 7 by showing that a certain class of leader patterns is not supported in equilibrium by players with some distribution of fairness concerns. Appendix C contains proofs.

1.2 Literature

There are many empirical studies showing that observed behaviors are inconsistent with the classical assumption that all agents are self-interested. An agreed fact is that many people are concerned with fairness and they are heterogeneous in fairness concerns. For example, by conducting an Afriat–Diewert–Varian-type revealed preference test on social outcomes, Andreoni and Miller (2002) verify it and in addition demonstrate that most of the subjects, whether identified as self-interested or not, are consistent with the

[10] See footnote 16 in Fehr, Kremhelmer, and Schmidt (2008).

[11] An example of diverse behaviors is the behaviors observed in the ultimatum experiments. Fehr and Schmidt (1999) argue that four classes of the Fehr and Schmidt (1999) preferences is enough to explain the data.

standard revealed preference paradigm.[12] Such empirical studies have provoked the development of various economics models that accommodate observed other-regarding behaviors. Using those models, many researchers have studied how fairness concerns affect economic outcomes in different environments. See Camerer (2003), Fehr and Schmidt (2006), and Cooper and Kagel (2013) for a detailed survey of the empirical and theoretical studies.

Among the models for studying fairness concerns, the outcome-based social preference models are most closely related to this paper. In the models, an agent with particular fairness characteristics is modeled as a preference relation over social outcomes. An altruism model (Andreoni and Miller (2002)) assumes that an agent's utility depends on the relevant agents' material payoffs profile and is monotonically increasing in the material payoff of the other agent.[13] An envy model (Bolton (1991)) assumes that an agent's utility depends on his material payoff and his relative payoff to the other player's payoff and that his utility is strictly increasing in his relative payoff when it is smaller than 1 and otherwise constant in his relative payoff.[14] These models can capture agents' unconditional reciprocal behavior and unconditional spiteful behavior, respectively, but not simultaneously. In contrast to these models, the Fehr and Schmidt (1999) model of inequity aversion that we employ in this paper models the two-dimensional aspect of social preferences.[15] In particular, the two kinds of sources of inequity aversions (envy and guilt) are treated separately. This enables us to explain leadership by three different behavioral modes that human beings exhibit in a given prisoner's dilemma.[16]

Other than the outcome-based models of social preferences, there are two dominant models of fairness concerns in the literature. In models of intention-based social preferences, a social interaction is modeled by using psychological game theory (Geanakoplos, Pearce, and Stacchetti (1989), Battigalli and Dufwenberg (2009)), and an agent with particular fairness characteristics is modeled as a preference relation over social outcomes and (higher-order) beliefs about behavior induced by the supposed strategy profile to capture concerns with agents' intention of the play. For applications of intention-based approach, see Rabin (1993), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006), Segal and Sobel (2007), Charness and Dufwenberg (2006), and Battigalli and Dufwenberg (2007).

In models of interdependent social preferences, an agent with particular fairness characteristics is modeled as a collection of preference relations over social outcomes, where interdependency arises in the sense that one of the preference relations is adopted depending on the fairness characteristics of everyone involved (Levine (1998)). This preference interdependency captures agents' personalities such as "if you are a good

[12] Of their 142 subjects, there is only one subject who had serious violations of revealed preference.

[13] See also Andreoni (1989), Andreoni (1990) for a related warm-glow model in a public-goods economy. Rotemberg and Saloner (1993) also study a leadership style using a model of preferences that parameterize a degree of altruism.

[14] See Kirchsteiger (1994) for a related model.

[15] The ERC model (Bolton and Ockenfels (2000)) assumes that an agent's utility depends on his material payoff and his payoff relative to the other player's payoff as in an envy model. In contrast to an envy model, however, the utility has a flexibility in its partial derivative with respect to the relative payoff and hence can capture both reciprocity and spitefulness behaviors. The Fehr and Schmidt (1999) and ERC model yield qualitatively similar results for the two-player cases but not in general. See Fehr and Schmidt (2006) for a related discussion.

[16] One can consider more complex models than the Fehr and Schmidt (1999) model by combining it with other models. For example, Charness and Rabin (2002) and Erlei (2008) develop such hybrid models. We employ the Fehr and Schmidt (1999) model because the model incorporates the minimum set of elements in social preferences that we need to explain the leadership mechanism.

person, then I will treat you kindly." See Rotemberg (2008, 2011) and Gul and Pesendorfer (2010) for applications of interdependent social preference models.[17]

Adopting these alternative models of other-regarding concerns may offer different kinds of leadership mechanisms from the one investigated in this paper.[18] Although this may provide interesting research topics, we employ the Fehr and Schmidt (1999) model for two reasons. First, although this model is simple, it successfully explains the dynamic behaviors in prisoner's dilemmas. In this respect, the Fehr and Schmidt (1999) model is certainly a promising candidate model when one considers the issue of time and social dilemmas by taking fairness concerns into account. Second, the Fehr and Schmidt (1999) model is tractable. In particular, this model has a simple and tractable utility functional form.[19] This fact enables us to develop our analyses in several directions, such as equilibrium existence, comparative statics, and equilibrium characterization in terms of fairness characteristics. Thus, the tractability of the Fehr and Schmidt (1999) model leads us to a testable theory on time and social dilemmas.

There are also other kinds of other-regarding models in team theory frameworks, which may be labeled as action-based other-regarding models, because those models directly assume that an agent feels psychological disutility from the relevant agents' actions. Among action-based other-regarding models, the one most related to ours is a model of social pressure. In addition to the cost of effort, a member of a team feels a disutility when he takes an action in a particular direction.[20] Using a pressure function, Huck and Rey-Biel (2006) model a conformity pressure on effort choices. That is, a member feels disutility from the difference between his effort level and his opponent's effort level. They study the relations between leadership and degrees of conformity inclination in a two-member team and show that endogenous leadership occurs in a team with at least one conformist under a complete information setup. Then, they briefly discuss a private information model with two possible degrees of conformity inclination (conformist and nonconformist) and claim that there are sequential equilibria with endogenous leadership. In contrast to Huck and Rey-Biel (2006), we employ an outcome-based social preference model, study endogenous leadership under a general incomplete information setup, and develop our analyses in several directions in addition to equilibrium existence.

Finally, other than the current paper, there are some attempts to explain leadership

[17] Gul and Pesendorfer (2010) also provide a foundation for this model in terms of economic primitives.

[18] For example, one may consider formalizing betrayal aversion, which is experimentally found by Bohnet and Zeckhauser (2004) and Bohnet, Greig, Herrmann, and Zeckhauser (2008). As we show in Section 2, how an agent thinks about his own betrayal (his D to opponent's C) and the risk of being betrayed (opponent's D to his C) is important for the endogenous leadership. The Fehr and Schmidt (1999) model creates a variety of thoughts about it, and this generates the endogenous leadership with a positive probability in equilibrium. However, as Bohnet and Zeckhauser (2004) and Bohnet, Greig, Herrmann, and Zeckhauser (2008) state, betrayal aversion may include elements beyond mere outcome-based other-regarding preferences. Formalizing this concept by using the alternative models in the text may be useful for studying other leadership mechanisms.

[19] As a result, this model has generated many economic applications as already listed in footnote 1 and has been well studied in terms of its behavioral basis as in Neilson (2006), Sandbu (2008), and Rohde (2010). See also Saito (2012) for a recent axiomatic development of the Fehr and Schmidt (1999) model, which proposes an extension of the Fehr and Schmidt (1999) model that accommodates the ex ante fairness problem presented by Fudenberg and Levine (2012).

[20] A seminal work on social pressure is Kandel and Lazear (1992). They modeled peer pressure. In addition to the cost of effort, a member of a team feels disutility from peer pressure, which may be social in the sense that it depends on other members' effort, such as the member feels substantial disutility when only he shirks. They introduced a general peer-pressure function and studied several kinds of peer pressure.

in social dilemma situations using a model of inequity aversion.[21] Duffy and Muñoz-García (2011) also employ the model of Fehr and Schmidt (1999) preferences with inequity aversion as in our theory and develop a Bayesian model of the prisoner's dilemma. However, they study prisoner's dilemmas with exogenous sequential moves for their study of leadership.[22] In a later section in this paper, we analyze the same model as theirs for the purpose of comparison with our endogenous timing game.

2 An illustrative example

To develop basic intuition about the relation between a diversity of inequity aversions and endogenous leadership, we illustrate by a degenerate numerical example the idea of a mechanism by which in equilibrium one player with a particular inequity aversion moves first with C and the other player with a different inequity aversion chooses to wait and responds with C to the first mover's choice of C.

i / j	C	D
C	$2, 2$	$0, 3$
D	$3, 0$	$1, 1$

Table 1: Example of a prisoner's dilemma

Consider the prisoner's dilemma in Table 1. Suppose that there are three possible types of inequity aversions: Materialist, Envy, and Envy-and-Guilt.[23] Each type evaluates each material payoff profile in Table 1 as in Table 2.

i / j	C	D
C	2	0
D	3	1

Materialist

i / j	C	D
C	2	-3
D	3	1

Envy

i / j	C	D
C	2	-3
D	0	1

Envy-and-Guilt

Table 2: Utilities in the prisoner's dilemma in Table 1

The Materialist type corresponds to a self-interested agent. His utility from an outcome of a pair of choices is the material payoff that he receives in the outcome. The other types differ from Materialist when there is an inequality in material payoffs. When there is a disadvantageous inequality, the Envy and Envy-and-Guilt types feel envy, which causes a disutility of the payoff difference in addition to a utility of own material payoffs. This happens in their (C, D) cells, in which their utilities are $0 + (-3) = -3$. When there is an advantageous inequality, the Envy-and-Guilt type feels guilt, which causes a disutility of the payoff difference. This happens in his (D, C) cell, in which his utility is $3 + (-3) = 0$. We assume that a player is a Materialist type, an Envy type, and an Envy-and-Guilt type with probabilities 0.2, 0.2, and 0.6, respectively.

Now consider a situation in which two players meet at random and play a game as in Table 1 without knowledge of the opponent's type, where players can decide the timing

[21] Nosenzo and Sefton (2011) consider a version of the Varian (1994) game. They embed the Varian (1994) game into the same two-period structure as in our game. However, they assume that agents have common inequity aversion parameters and that the parameter values are common knowledge. Santos-Pinto (2008) also considers an endogenous timing duopoly market with inequity-averse firms under the same assumptions.

[22] Their main concern is to study signaling motives about fairness in a twice-repeated simultaneous-move or sequential-move prisoner's dilemma game.

[23] They correspond to $(\alpha, \beta) = (0, 0)$, $(1, 0)$, and $(1, 1)$ in the Fehr and Schmidt (1999) model.

of their choice voluntarily from the possible two timings. In this situation, we can show that endogenous leadership happens with a positive probability in equilibrium. The key is how each type thinks about his own betrayal (his D to his opponent's C) and a risk of being betrayed (his opponent's D to his C). The diversity of inequity aversions creates diverse ways of thinking about it; that is, it creates different degrees of incentives to induce conditional cooperation under a risk of being betrayed and/or different degrees of incentives to cooperate after observing the opponent's cooperation. Then, such diverse sensitivities of cooperation naturally generate diverse ways of dynamic decision making, from which endogenous leadership results.

In what follows, we demonstrate that the following Bayesian strategy constitutes a sequential equilibrium. The Materialist type behaves as a leader, which means that he chooses C at timing 1. The Envy type behaves as a defector, which means that he chooses D at timing 2 irrespective of his observation. The Envy-and-Guilt type behaves as a conditional cooperator, which means that he moves at timing 2 and chooses C (resp., D) if his opponent chooses C at timing 1 (resp., otherwise).

Let us verify that no type has an incentive to mimic the other types' behaviors.[24] First, we check if the Materialist type behaves as a leader. He is a benchmark case of this example in that he is free from envy and guilt; that is, he feels neither disutility of disadvantageous inequality from being betrayed nor disutility of advantageous inequality from his own betrayal. Given the induced population of the three behavior modes (leader: 0.2, defector: 0.2, conditional cooperator: 0.6), the Materialist type enjoys expected utility 1.6 ($= 0.2 \times 2 + 0.2 \times 0 + 0.6 \times 2$) if he behaves as a leader, 1.4 ($= 0.2 \times 3 + 0.2 \times 1 + 0.6 \times 1$) if he behaves as a defector, and 1.2 ($= 0.2 \times 2 + 0.2 \times 1 + 0.6 \times 1$) if he behaves as a conditional cooperator. Hence, the Materialist type optimally behaves as a leader. This happens because the chance of playing with the Envy-and-Guilt type, who behaves as a conditional cooperator, is large enough so that the benefit of inducing the Envy-and-Guilt type to take C exceeds the cost of giving up D against the Materialist and Envy types.

Second, we check if the Envy type behaves as a defector. Since leading behavior may create a disadvantageous inequality under the presence of defectors, envy discourages players from the leadership behavior under the betrayal risk. To be precise, the expected utility earned by the Envy type behaving as a leader reduces from 1.6 to 1.0 ($= 0.2 \times 2 + 0.2 \times (-3) + 0.6 \times 2$). This makes the defector mode optimal for the Envy type.

Third, we check if the Envy-and-Guilt type behaves as a conditional cooperator. In addition to the abovementioned effect of envy, guilt plays a role. Since defecting behavior may create an advantageous inequality in the presence of leaders, guilt discourages players from the betraying behavior. To be precise, the expected utility earned by the Envy-and-Guilt type behaving as a defector reduces from 1.4 to 0.8 ($= 0.2 \times 0 + 0.2 \times 1 + 0.6 \times 1$). This makes the conditional cooperator mode optimal for the Envy-and-Guilt type.

Thus, we establish that the abovementioned Bayesian strategy indeed constitutes a sequential equilibrium under the specific type distribution. Now suppose that a player of the Materialist type and a player of the Envy-and-Guilt type are paired and play the prisoner's dilemma. Then, the Materialist type chooses C at timing 1 and the Envy-and-Guilt type, who moves at timing 2, reacts with C. Thus, a successful leadership is endogenously realized with a positive probability in equilibrium.

Note that it is critical for the endogenous leadership that all of the three behavior modes prevail with positive probabilities. First, the leading behavior taken by the

[24] The reader may verify that he has no incentive for any other deviation.

Materialist type is obviously indispensable. Furthermore, the conditional cooperation by the Envy-and-Guilt type is not triggered without the leading behavior. Next, a conditional cooperator is also indispensable. To see this, suppose that no conditional cooperator exists. Then, the Materialist type would not behave as a leader because then behaving as a leader has no effect of inducing a second mover to take C and simply means giving up the dominant choice of D.[25] Finally, defectors must also exist. If there is no defector, there is no risk of being betrayed. Therefore, the Envy-and-Guilt type would not wait and see, but rather would behave as a leader. Then, there would be no type who behaves as a conditional cooperator. In this way, the endogenous leadership cannot afford to lose any of the three behavior modes.

In the above degenerate example, we assume a particular combination of three types of inequity aversions with a particular distribution and consider a particular leadership pattern by these types. In the general model that we consider in this paper, the degree of incentive for a type to induce conditional cooperation is determined in equilibrium jointly with a distribution of the three behavior modes. As we will show, there exist three leadership patterns, depending on the parameters.

3 The model

We consider a version of the prisoner's dilemma game in which players choose the timing of moves under incomplete information about the preferences of players. This game is called PD hereafter and defined as follows.

A prisoner's dilemma is a symmetric game given by Table 3. The parameters a, b, c, d are the material payoffs. They are assumed to be $b > a > d > c$. This payoff structure exhibits social dilemma, because D is the payoff-maximizing choice for any given belief about the opponent's choice although (D, D) is Pareto inferior to (C, C).[26]

i / j	C	D
C	a, a	c, b
D	b, c	d, d

Table 3: Prisoner's dilemma

We hypothesize that each player i has Fehr–Schmidt inequity-averse preferences over the outcomes described by the pairs of material payoffs in Table 3. The preferences are represented by a utility function

$$u_{(\alpha_i,\beta_i)}(x_i, x_j) = x_i - \alpha_i \max\{x_j - x_i, 0\} - \beta_i \max\{x_i - x_j, 0\} \tag{1}$$

where x_i and x_j are material payoffs to players i and j, respectively. The first term represents the direct utility from the material payoff x_i. The second term captures the utility loss from disadvantageous inequality when x_i is less than x_j. The parameter $\alpha_i \geq 0$ may be interpreted as the envy of player i. The third term captures the utility loss from advantageous inequality when x_i is larger than x_j. The parameter $\beta_i \geq 0$ may be interpreted as the sense of guilt of player i.

[25] In the general model which we consider in this paper, the effect of no conditional cooperator is that then behaving as a conditional cooperator dominates behaving as a leader.

[26] We do not include the restriction $2a > b + c$ commonly imposed in repeated prisoner's dilemma studies, which guarantees that (C, C) is value-maximizing efficient, because it is irrelevant to our analysis.

From the pairs of material payoffs (x_i, x_j) in Table 3, we have a utility representation of the prisoner's dilemma played by players with Fehr–Schmidt preferences. This is shown in Table 4.

i / j	C	D
C	a, a	$c - \alpha_i(b-c), b - \beta_j(b-c)$
D	$b - \beta_i(b-c), c - \alpha_j(b-c)$	d, d

Table 4: Utility representation of the prisoner's dilemma

We assume that Table 3 is common knowledge among the players, but Table 4 is not. Player i knows the inequality aversion parameters α_i and β_i in his utility function, but he does not know his opponent's parameters α_j and β_j.

We introduce a common prior assumption about the belief that a player initially holds about his opponent's utility function. We call the pair (α, β) in the utility function (1) the type of the player. The type is a realization of the continuous random variable $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in the space of possible types given by

$$T = \{(\alpha, \beta) | 0 \leq \alpha \leq \bar{\alpha}, 0 \leq \beta \leq 1, \beta \leq \alpha\}. \tag{2}$$

The parameter $\bar{\alpha}$ is the upper bound of the envy parameter and it can be either finite or infinite. The upper bound of the guilty parameter is assumed to be 1 because a guilty parameter larger than 1 would mean that more material payoffs are undesirable when a player is receiving more than his opponent. The envy parameter α is assumed to be no less than the guilty parameter β because inequality in material payoffs matters more when a player receives less than his opponent.[27] We assume that a type (α, β) is realized according to a density $f(\alpha, \beta)$ with the full support over T. We also assume that realizations are independent across players. All of these assumptions about initial beliefs are assumed to be common knowledge.

Under incomplete information about their utility functions, the players play the prisoner's dilemma in Table 3 in the following sequence. There are two timings: 1 and 2. At timing 1, the players choose either C, D, or $\emptyset$ independently and simultaneously, where $\emptyset$ denotes postponing their choices until timing 2 rather than choosing C or D. At timing 2, a player has a move when and only when he chooses $\emptyset$ at timing 1. Before he moves, he is informed of his opponent's choice at timing 1 and then he must choose either C or D. When both players choose $\emptyset$ at timing 1, their choices at timing 2 are made independently and simultaneously. This is the end of a play. Each player has made a choice over C or D either at timing 1 or 2 once and only once. A player receives the material payoff in Table 3 corresponding to the pair of their choices.

A PD thus defined is determined by two sets of parameters. One is the material payoffs (a, b, c, d) in the underlying prisoner's dilemma. The other is a characteristic (f, T) of players' preferences. We fix T throughout the paper and write $PD((a, b, c, d), f)$ when the parameters of the PD need to be explicit.

[27] We follow the standard formalization of the parameter space that Fehr and Schmidt (1999) proposed. Several subsequent studies consider the possibility of expanding this parameter space (for example, status seeking modeled by $\beta < 0$). However, we focus on the issue of inequity aversion captured by this type space.

4 The three mode equilibrium

4.1 The strategy

A player has four information sets in PD. One is the information set for choice at timing 1. The other three are the information sets for choice at timing 2 corresponding to his opponent's choice at timing 1 being either C, D, or $\emptyset$. A (pure) strategy is a complete plan that assigns to each of these information sets an action available at the information set. Formally, it is a quadruplet $s = (a_1, a_C, a_D, a_\emptyset)$ where $a_1 \in \{C, D, \emptyset\}$ is the prescribed choice at timing 1, and $a_C, a_D, a_\emptyset \in \{C, D\}$ are the prescribed choices at timing 2 when his opponent's choices at timing 1 are C, D, and $\emptyset$, respectively. The (pure) strategy space is $S = \{C, D, \emptyset\} \times \{C, D\} \times \{C, D\} \times \{C, D\}$.

A Bayesian strategy is a mapping $\boldsymbol{s} : T \to S$. It assigns to each type $(\alpha, \beta) \in T$ a strategy

$$\boldsymbol{s}(\alpha, \beta) = (\boldsymbol{a}_1(\alpha, \beta), \boldsymbol{a}_C(\alpha, \beta), \boldsymbol{a}_D(\alpha, \beta), \boldsymbol{a}_\emptyset(\alpha, \beta)) \in S$$

which this type follows in a play of the PD where $\boldsymbol{a}_h(\alpha, \beta)$ is the prescribed choice a_h for each information set h for type (α, β).

4.2 The three-mode strategy

We examine a Bayesian strategy named *three-mode strategy* that assigns a strategy from a particular set of three behavior modes. They are called C-mode, CDD-mode, and DDD-mode and are defined as the following strategies in S where C-mode is described as a reduced strategy with a_C, a_D, and $a_\emptyset$ for unreached information sets left unspecified.

$$\begin{aligned} C &= (C, a_C, a_D, a_\emptyset) \\ CDD &= (\emptyset, C, D, D) \\ DDD &= (\emptyset, D, D, D) \end{aligned}$$

These strategies in S are called *behavior modes* in a three-mode strategy when we stress the feature that each of them are followed by a mass of types in T according to the Bayesian strategy.

Formally, for a Bayesian strategy $\boldsymbol{s} : T \to S$, let

$$\begin{aligned} T_C(\boldsymbol{s}) &= \{(\alpha, \beta) \in T | \boldsymbol{s}(\alpha, \beta) = C\} \\ T_{CDD}(\boldsymbol{s}) &= \{(\alpha, \beta) \in T | \boldsymbol{s}(\alpha, \beta) = CDD\} \\ T_{DDD}(\boldsymbol{s}) &= \{(\alpha, \beta) \in T | \boldsymbol{s}(\alpha, \beta) = DDD\} \end{aligned}$$

denote the sets of types who follow C, CDD, and DDD, respectively. Let ϕ denote the probability measure over T induced by the prior density f. For a Borel subset B of T, it assigns the probability of a player being of a type in B by

$$\phi(B) = \int_B f(\alpha, \beta) d(\alpha, \beta). \tag{3}$$

Then, the three-mode strategy is defined as follows.

Definition 1. *A Bayesian strategy $\boldsymbol{s} : T \to S$ is a three-mode strategy if $(T_C(\boldsymbol{s}), T_{CDD}(\boldsymbol{s}), T_{DDD}(\boldsymbol{s}))$ is a partition of T and $\phi(T_C(\boldsymbol{s})) > 0$, $\phi(T_{CDD}(\boldsymbol{s})) > 0$, $\phi(T_{DDD}(\boldsymbol{s})) > 0$.*

When players play a three-mode strategy $\boldsymbol{s}$, four kinds of play paths will be realized with positive probabilities. These are shown in Table 5. For example, (C, CDD) cell shows a play path in which player i chooses C at timing 1 according to $s_i = C$ and player j waits at timing 1 and chooses C at timing 2 according to $s_j = CDD$. This is the leadership process.

s_i / s_j	C	CDD	DDD
C	$C \quad , C$	$C \quad , \emptyset \to C$	$C \quad , \emptyset \to D$
CDD	$\emptyset \to C, C$	$\emptyset \to D, \emptyset \to D$	$\emptyset \to D, \emptyset \to D$
DDD	$\emptyset \to D, C$	$\emptyset \to D, \emptyset \to D$	$\emptyset \to D, \emptyset \to D$

Table 5: Play paths of three-mode strategies

4.3 The three-mode equilibrium

We consider a symmetric sequential equilibrium in three-mode strategies. We call this equilibrium a *three-mode equilibrium*.

We prepare a characterization of the three-mode equilibrium by an associated distribution of the three behavior modes. Let $\mu = (\mu_C, \mu_{CDD}, \mu_{DDD})$ denote a probability distribution according to which a player follows strategies C, CDD, and DDD. Let

$$\Delta \equiv \{\mu = (\mu_C, \mu_{CDD}, \mu_{DDD}) | 0 \leq \mu_C, \mu_{CDD}, \mu_{DDD} \leq 1 \text{ and } \mu_C + \mu_{CDD} + \mu_{DDD} = 1\}$$

be the set of distributions over the three strategies. Then,

Definition 2. *A distribution $\mu \in \Delta$ is called a three-mode distribution if $\mu_C > 0$, $\mu_{CDD} > 0$, and $\mu_{DDD} > 0$.*

A three-mode strategy $\boldsymbol{s}$ is associated with a three-mode distribution μ by the relation $\mu_C = \phi(T_C(\boldsymbol{s}))$, $\mu_{CDD} = \phi(T_{CDD}(\boldsymbol{s}))$, and $\mu_{DDD} = \phi(T_{DDD}(\boldsymbol{s}))$. A three-mode distribution is called a *three-mode equilibrium distribution* if a three-mode strategy that generates the distribution is a three-mode equilibrium strategy.

Imagine that a player holds a belief $\mu \in \Delta$ according to which his opponent follows strategies C, CDD, and DDD. Then, the expected utility for a type (α, β) to follow C given μ is given by

$$U_{(\alpha,\beta)}(C, \mu) = \mu_C a + \mu_{CDD} a + \mu_{DDD}(c - \alpha(b - c))$$

because, as is shown in Table 5, the outcome will be (C, C), (C, C), and (C, D) when the opponent follows C, CDD, and DDD, respectively. Similarly, the expected utilities from CDD and DDD are:

$$\begin{aligned} U_{(\alpha,\beta)}(CDD, \mu) &= \mu_C a + \mu_{CDD} d + \mu_{DDD} d \\ U_{(\alpha,\beta)}(DDD, \mu) &= \mu_C(b - \beta(b - c)) + \mu_{CDD} d + \mu_{DDD} d. \end{aligned}$$

We define the following sets of types that are supposed to describe those sets of types for whom C, CDD, and DDD are sequentially rational responses to μ.

$$\begin{aligned} T^*_C(\mu) &= \{(\alpha, \beta) \in T | U_{(\alpha,\beta)}(C, \mu) \geq U_{(\alpha,\beta)}(CDD, \mu), U_{(\alpha,\beta)}(DDD, \mu)\} \\ T^*_{CDD}(\mu) &= \{(\alpha, \beta) \in T | U_{(\alpha,\beta)}(CDD, \mu) \geq U_{(\alpha,\beta)}(C, \mu), U_{(\alpha,\beta)}(DDD, \mu) \text{ and } \beta \geq \beta^*\} \\ T^*_{DDD}(\mu) &= \{(\alpha, \beta) \in T | U_{(\alpha,\beta)}(DDD, \mu) \geq U_{(\alpha,\beta)}(C, \mu), U_{(\alpha,\beta)}(CDD, \mu) \text{ and } \beta \leq \beta^*\}, \end{aligned}$$

where we define $\beta^* \equiv \frac{b-a}{b-c}$. The sets $T^*_C(\mu)$, $T^*_{CDD}(\mu)$ and $T^*_{DDD}(\mu)$ are called the *best-response type sets*. The set $T^*_C(\mu)$ is a set of types for whom C is the best among the three strategies when he evaluates them at the information set at timing 1 given the belief μ. The set $T^*_{CDD}(\mu)$ is a set of types for whom CDD is the best among the three strategies at two information sets. It is the best at the information set at timing 1 given the belief μ, which is parallel to $T^*_C(\mu)$. It is also the best at the information set at timing 2 where he observes that his opponent chooses C at timing 1. He chooses between C and D at this information set. He prefers C to D if $a > b - \beta(b-c)$, he prefers the opposite if $a < b - \beta(b-c)$, and he is indifferent if $a = b - \beta(b-c)$. Therefore, the sequential rationality for CDD requires $\beta \geq \beta^*$. The set $T^*_{DDD}(\mu)$ is parallel to $T^*_{CDD}(\mu)$.

Now, we show that the three-mode equilibrium is characterized by a three-mode distribution, as follows.

Lemma 1. *Consider a function ψ that assigns to each $\mu \in \Delta$ a vector $\psi(\mu) = (\psi_C(\mu), \psi_{CDD}(\mu), \psi_{DDD}(\mu))$ by $\psi_C(\mu) = \phi(T^*_C(\mu))$, $\psi_{CDD}(\mu) = \phi(T^*_{CDD}(\mu))$, and $\psi_{DDD}(\mu) = \phi(T^*_{DDD}(\mu))$. Then, $\psi(\mu) \in \Delta$ for every $\mu \in \Delta \setminus (1,0,0)$ and ψ is continuous on $\Delta \setminus (1,0,0)$. Furthermore,*

(1) *If $\boldsymbol{s} : T \to S$ is a three-mode equilibrium strategy, then $\mu = (\phi(T_C(\boldsymbol{s})), \phi(T_{CDD}(\boldsymbol{s})), \phi(T_{DDD}(\boldsymbol{s})))$ is a three-mode distribution such that $\mu = \psi(\mu)$.*

(2) *If $\mu \in \Delta$ is a three-mode distribution such that $\mu = \psi(\mu)$, then any three-mode strategy $\boldsymbol{s} : T \to S$ that satisfies $T_C(\boldsymbol{s}) \subseteq T^*_C(\mu)$, $T_{CDD}(\boldsymbol{s}) \subseteq T^*_{CDD}(\mu)$, and $T_{DDD}(\boldsymbol{s}) \subseteq T^*_{DDD}(\mu)$ is a three-mode equilibrium strategy. Furthermore, $\phi(T_C(\boldsymbol{s})) = \phi(T^*_C(\mu)) = \mu_C$, $\phi(T_{CDD}(\boldsymbol{s})) = \phi(T^*_{CDD}(\mu)) = \mu_{CDD}$, and $\phi(T_{DDD}(\boldsymbol{s})) = \phi(T^*_{DDD}(\mu)) = \mu_{DDD}$.*

Lemma 1 means that the function ψ generates a distribution over strategies C, CDD, and DDD by generating probability measures of best-response type sets against a belief $\mu \in \Delta \setminus (1,0,0)$ and that a three-mode equilibrium strategy is identified as a three-mode strategy that generates a three-mode distribution as a fixed point of the function ψ.

4.4 The best-response type sets

To characterize a fixed point of ψ, let us study the best-response type sets $T^*_C(\mu)$, $T^*_{CDD}(\mu)$, and $T^*_{DDD}(\mu)$. We explore them by examining preferences over C, CDD, and DDD because the best-response type sets are defined by the preferences. Take the case of μ with $0 < \mu_C < 1$. First, consider a preference over C and CDD. A player of (α, β) type prefers C to CDD if and only if $\mu_C a + \mu_{CDD} a + \mu_{DDD}(c - \alpha(b-c)) > \mu_C a + \mu_{CDD} d + \mu_{DDD} d$. Let

$$\alpha^*(\mu) \equiv \frac{a-d}{b-c}\frac{\mu_{CDD}}{\mu_{DDD}} - \frac{d-c}{b-c} \tag{4}$$

where we set $\alpha^*(\mu) = \infty$ when $\mu_{DDD} = 0$. Then, (α, β) type prefers C to CDD if and only if $\alpha < \alpha^*(\mu)$. He prefers the opposite when $\alpha > \alpha^*(\mu)$, and he is indifferent when $\alpha = \alpha^*(\mu)$. The threshold $\alpha^*(\mu)$ lies in the interval $(0, \bar{\alpha})$ and partitions the type space T if and only if

$$\frac{d-c}{a-d} < \frac{\mu_{CDD}}{\mu_{DDD}} < \frac{d-c}{a-d} + \frac{b-c}{a-d}\bar{\alpha}. \tag{5}$$

Second, consider a preference over CDD and DDD. A player of (α, β) type prefers CDD to DDD if and only if $\mu_C a + \mu_{CDD} d + \mu_{DDD} d > \mu_C(b - \beta(b-c)) + \mu_{CDD} d + \mu_{DDD} d$; that is, $\beta > \beta^*$ where the threshold β^* is defined previously when we introduced the best-response type sets. He prefers the opposite when $\beta < \beta^*$, and he is indifferent when $\beta = \beta^*$. Note that the threshold β^* is $0 < \beta^* = \frac{b-a}{b-c} < 1$ so that it partitions the type space.

Finally, consider a preference over C and DDD. A player of (α, β) type prefers C to DDD if and only if $\mu_C a + \mu_{CDD} a + \mu_{DDD}(c - \alpha(b-c)) > \mu_C(b - \beta(b-c)) + \mu_{CDD} d + \mu_{DDD} d$. Let

$$\begin{aligned} H(\alpha|\mu) &\equiv \frac{\mu_{DDD}}{\mu_C}\alpha + \left[\frac{b-a}{b-c} + \frac{\mu_{DDD}(d-c) - \mu_{CDD}(a-d)}{\mu_C(b-c)}\right] \\ &= \frac{\mu_{DDD}}{\mu_C}(\alpha - \alpha^*(\mu)) + \beta^*. \end{aligned} \tag{6}$$

Then, (α, β) type prefers C to DDD if and only if $\beta > H(\alpha|\mu)$. He prefers the opposite when $\beta < H(\alpha|\mu)$, and he is indifferent when $\beta = H(\alpha|\mu)$.

Note that, as a consequence of transitivity of preferences, the two thresholds $\alpha^*(\mu)$, β^* introduced so far satisfy

$$U_{(\alpha^*(\mu),\beta^*)}(C, \mu) = U_{(\alpha^*(\mu),\beta^*)}(CDD, \mu) = U_{(\alpha^*(\mu),\beta^*)}(DDD, \mu).$$

This means that $\beta^* = H(\alpha^*(\mu)|\mu)$. Therefore, the best-response type sets to μ with $0 < \mu_C < 1$ are characterized immediately, as follows.

Lemma 2. *For μ with $0 < \mu_C < 1$, the best response type sets are*

$$T^*_C(\mu) = T \cap \{(\alpha, \beta) | \alpha \leq \alpha^*(\mu), \beta \geq H(\alpha|\mu)\} \tag{7}$$

$$T^*_{CDD}(\mu) = T \cap \{(\alpha, \beta) | \alpha \geq \alpha^*(\mu), \beta \geq \beta^*\} \tag{8}$$

$$T^*_{DDD}(\mu) = T \cap \{(\alpha, \beta) | \beta \leq H(\alpha|\mu), \beta \leq \beta^*\}. \tag{9}$$

An example of the best-response type sets are illustrated in Figure 1. This example is obtained when μ satisfies $\beta^* < \alpha^*(\mu) < \bar{\alpha}$ and $H(0|\mu) > 0$. All of the best-response type sets are nondegenerate in this example.

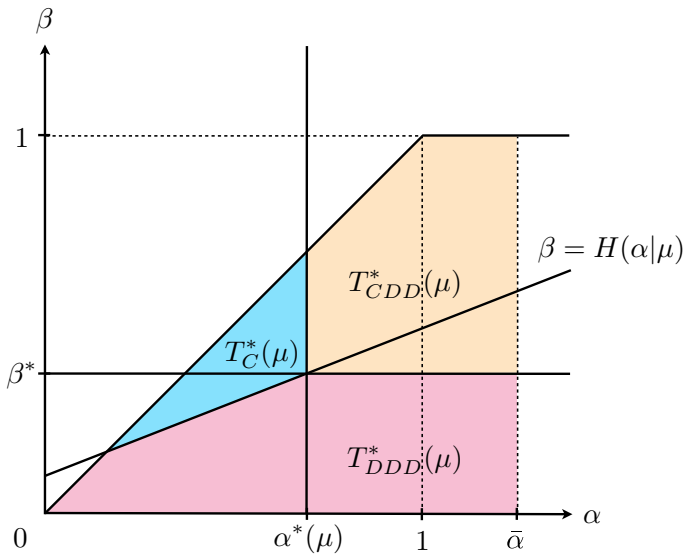


Figure 1: Inequity concerned leader

Consider the case of μ with $\mu_C = 0$. Then, $U_{(\alpha,\beta)}(CDD, \mu) = U_{(\alpha,\beta)}(DDD, \mu) = d$ for any $(\alpha, \beta) \in T$. Therefore, it is straightforward to see that the best-response type sets are as follows.

Lemma 3. *For μ with $\mu_C = 0$, the best-response type sets are*

$$T^*_C(\mu) = T \cap \{(\alpha, \beta) | \alpha \leq \alpha^*(\mu)\} \tag{10}$$

$$T^*_{CDD}(\mu) = T \cap \{(\alpha, \beta) | \alpha \geq \alpha^*(\mu), \beta \geq \beta^*\} \tag{11}$$

$$T^*_{DDD}(\mu) = T \cap \{(\alpha, \beta) | \alpha \geq \alpha^*(\mu), \beta \leq \beta^*\}. \tag{12}$$

5 Existence of the three-mode equilibrium

5.1 μ_C-fixed point and fixed point of ψ

Our first result is a characterization of the existence of a three-mode equilibrium. In light of Lemma 1, we study a fixed point of ψ. Since a three-mode equilibrium distribution must be fully mixed, we must exclude a fixed point on the boundaries of Δ.

We study a fixed point of ψ in two steps.[28] First, we identify the set of μ_C-fixed points of ψ. A μ_C-fixed point $\hat{\mu}$ is a belief in Δ with a property that the probability of strategy C remains unchanged by ψ; that is, $\psi_C(\hat{\mu}) = \hat{\mu}_C$. The second step is to identify from the set of μ_C-fixed points a belief $\hat{\mu}$ with an additional property that the probability of strategy DDD also remains unchanged by ψ; that is, $\psi_{DDD}(\hat{\mu}) = \hat{\mu}_{DDD}$.

Now, we begin the first step to find μ_C-fixed points of ψ. In particular, we are interested in those μ_C-fixed points that are three-mode distributions. For $\hat{\mu}$ to satisfy $\psi_C(\hat{\mu}) = \hat{\mu}_C$ and $\hat{\mu}_C > 0$, the best-response type set $T^*_C(\hat{\mu})$ must have a positive probability measure. By Lemma 2, this implies that $\alpha^*(\hat{\mu}) > 0$ for the threshold (4). By the condition (5), this means that $\hat{\mu}_{CDD} > \frac{d-c}{a-d}\hat{\mu}_{DDD}$. Therefore, we study those $\hat{\mu}$ with $\psi_C(\hat{\mu}) = \hat{\mu}_C$ in a subset

$$\Delta' \equiv \{\tilde{\mu} \in \Delta \setminus (1,0,0) | \tilde{\mu}_{CDD} \geq \frac{d-c}{a-d}\tilde{\mu}_{DDD}\}$$

of $\Delta \setminus (1,0,0)$. The following establishes that those points are an arc in Δ'.

Lemma 4. *There exists a unique $\bar{\mu}_C$ in $(0,1)$ such that for each $\mu_C \in [0, \bar{\mu}_C]$, there exists a unique $\hat{\mu} \in \Delta'$ that satisfies $\psi_C(\hat{\mu}) = \hat{\mu}_C = \mu_C$. We denote by $\hat{\mu} = \hat{\mu}(\mu_C)$ a function that assigns to each $\mu_C \in [0, \bar{\mu}_C]$ the corresponding μ_C-fixed point $\hat{\mu}$. Then,*

(1) *for $\mu_C = 0$, $\hat{\mu}(0) = (0, \frac{d-c}{a-c}, \frac{a-d}{a-c})$ and $\psi(\hat{\mu}(0)) = (0, \phi(\boldsymbol{\beta} > \beta^*), \phi(\boldsymbol{\beta} < \beta^*))$,*

(2) *for $\mu_C = \bar{\mu}_C$, $\hat{\mu}(\bar{\mu}_C) = (\bar{\mu}_C, 1 - \bar{\mu}_C, 0)$ and $\psi(\hat{\mu}(\bar{\mu}_C)) = (\bar{\mu}_C, 0, 1 - \bar{\mu}_C)$,*

(3) *for $\mu_C \in (0, \bar{\mu}_C)$, $\hat{\mu}(\mu_C)$ is in the interior of Δ'.*

Furthermore, the function $\hat{\mu}(\mu_C)$ is continuous.

Lemma 4 is illustrated in Figure 2. The distribution $\hat{\mu} = (0, \frac{d-c}{a-c}, \frac{a-d}{a-c})$ is defined by a boundary $\hat{\mu}_C = 0$ of Δ and a boundary $\hat{\mu}_{CDD} = \frac{d-c}{a-d}\hat{\mu}_{DDD}$ of Δ'. It is a μ_C-fixed point for $\mu_C = 0$ because the latter condition (that is $\alpha^*(\hat{\mu}) = 0$) means that $\hat{\mu}$

[28] Alternatively, we can study a fixed point of ψ by a standard method that relies on the Kakutani fixed-point theorem by extending a function ψ to a correspondence defined over Δ. We use the two-step analysis developed below because this analysis is also useful for the comparative static analysis of the three-mode equilibrium, which we develop later.

is mapped by ψ to $\psi(\hat{\mu})$ such that $\psi_C(\hat{\mu}) = 0$. There exists another μ_C-fixed point $\hat{\mu} = (\bar{\mu}_C, 1 - \bar{\mu}_C, 0)$ for some $\bar{\mu}_C \in (0, 1)$ on the boundary $\mu_{DDD} = 0$ of Δ. To see it, note that any distribution μ on this boundary is mapped by ψ to $\psi(\mu)$ such that $\psi_{CDD}(\mu) = 0$. The reason is that, as we noted in the illustrative example of Section 2, when the opponent is expected to never follow DDD (that is, $\mu_{DDD} = 0$), there is no risk of being betrayed by an opponent. This makes C dominate CDD. Therefore, a best response must be either C or DDD. The distribution $\hat{\mu} = (\bar{\mu}_C, 1 - \bar{\mu}_C, 0)$ is a distribution for which the best response generates the same probability $\psi_C(\hat{\mu}) = \bar{\mu}_C$ of C and a probability $\psi_{DDD}(\hat{\mu}) = 1 - \bar{\mu}_C$ of DDD, which is equal in size to the probability $\hat{\mu}_{CDD}$ of CDD in $\hat{\mu}$. Lemma 4 means that the set of μ_C-fixed point is an arc that connects these two extreme points because a μ_C-fixed point exists continuously depending on μ_C from $\mu_C = 0$ to $\mu_C = \bar{\mu}_C$.

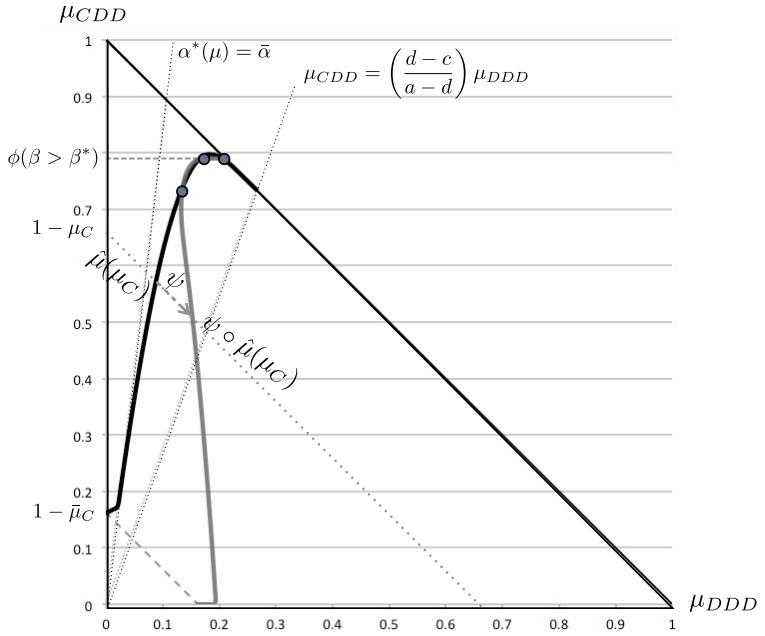


Figure 2: μ_C-fixed points and their ψ images under $f(\alpha, \beta) = 60(1 - \alpha)\beta^2$

The second step to find a fixed point of ψ is to find a belief $\hat{\mu}$ along the arc $\hat{\mu} = \hat{\mu}(\mu_C)$ of μ_C-fixed points for which the probability of strategy DDD also remains unchanged by ψ; that is, $\psi_{DDD}(\hat{\mu}) = \hat{\mu}_{DDD}$. Figure 2 also illustrates this exercise. We map each μ_C-fixed point $\hat{\mu}$ on the arc $\hat{\mu} = \hat{\mu}(\mu_C)$ to a belief in Δ by ψ. The image of the arc by ψ is also an arc in Δ because ψ is continuous. As is reported in Lemma 4, it starts at $\psi(\hat{\mu}(0)) = (0, \phi(\boldsymbol{\beta} > \beta^*), \phi(\boldsymbol{\beta} < \beta^*))$ for $\mu_C = 0$ and ends at $\psi(\hat{\mu}(\bar{\mu})) = (\bar{\mu}_C, 0, 1 - \bar{\mu}_C)$ for $\mu_C = \bar{\mu}_C$. A crossing point of this image arc with the arc $\hat{\mu} = \hat{\mu}(\mu_C)$ is a fixed point of ψ.

For the purpose of this analysis, we define a function λ that assigns to each $\mu_C \in [0, \bar{\mu}_C]$

$$\lambda(\mu_C) = \psi_{DDD}(\hat{\mu}(\mu_C)) - \hat{\mu}_{DDD}(\mu_C) \tag{13}$$

where we write the components of a μ_C-fixed point $\hat{\mu}(\mu_C)$ as $\hat{\mu}(\mu_C) = (\hat{\mu}_C(\mu_C), \hat{\mu}_{CDD}(\mu_C), \hat{\mu}_{DDD}(\mu_C))$. Figure 3 shows a graph of this function (13) that corresponds to Figure 2. When a μ_C-fixed point $\hat{\mu}$ is mapped by ψ to an image $\psi(\hat{\mu})$ located north-west of $\hat{\mu}$ on a line with the same μ_C in Figure 2, the corresponding value of λ is negative in Figure 3. When it is mapped in the opposite direction in Figure 2, λ is positive in Figure 3. The

crossing points of the arc and its image in Figure 2 correspond to the solutions to the equation $\lambda(\mu_C) = 0$ in Figure 3.

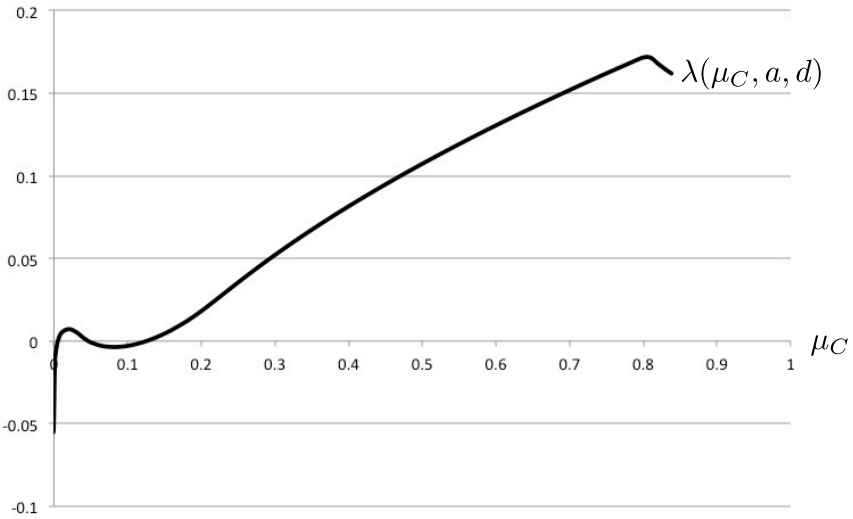


Figure 3: λ under $f(\alpha, \beta) = 60(1-\alpha)\beta^2$

Now, a three-mode equilibrium distribution is characterized as follows.

Lemma 5. *A belief $\mu \in \Delta$ is a three-mode equilibrium distribution if and only if*

(1) $\mu = \hat{\mu}(\mu_C)$

(2) $\lambda(\mu_C) = 0$

(3) $\mu_C > 0$.

Condition (1) is a requirement that $\mu_C = \psi_C(\mu)$. Condition (2) is a requirement that $\mu_{DDD} = \psi_{DDD}(\mu)$. These conditions guarantee that μ is a fixed point of ψ in $\Delta \setminus (1,0,0)$. Condition (3) excludes a fixed point at an extreme point $\hat{\mu}(0) = (0, \frac{d-c}{a-c}, \frac{a-d}{a-c})$ of the arc $\hat{\mu} = \hat{\mu}(\mu_C)$.

In Lemma 5, we do not need a condition to exclude a fixed point at another extreme point $\hat{\mu}(\bar{\mu}_C) = (\bar{\mu}_C, 1-\bar{\mu}_C, 0)$ of the arc $\hat{\mu} = \hat{\mu}(\mu_C)$. From Lemma 4, we already know that $\hat{\mu}(\bar{\mu}_C) = (\bar{\mu}_C, 1-\bar{\mu}_C, 0)$ is mapped to $\psi(\hat{\mu}(\bar{\mu}_C)) = (\bar{\mu}_C, 0, 1-\bar{\mu}_C)$. Therefore, $\hat{\mu}_{DDD}(\bar{\mu}_C) = 0 < 1-\bar{\mu}_C = \psi_{DDD}(\hat{\mu}(\bar{\mu}_C))$. Hence, $\hat{\mu}(\bar{\mu}_C)$ is never a fixed point of ψ.

5.2 A sufficient condition for the existence of the three-mode equilibrium

Now, we develop a sufficient condition for the existence of a three-mode equilibrium. The sufficient condition also conveys a simple and straightforward logic by which leadership should be expected to emerge if a prisoner's dilemma satisfies the condition.

Lemma 5 shows that a three-mode equilibrium exists if there exists a solution $\mu_C > 0$ to the equation $\lambda(\mu_C) = 0$. As we noted after the lemma, it is always the case that $\lambda(\bar{\mu}) > 0$. Therefore, if $\lambda(0) < 0$, the continuity of $\lambda(\mu_C)$ guarantees a solution $\mu_C > 0$ to the equation $\lambda(\mu_C) = 0$. This leads us to the following sufficient condition for the existence of the three-mode equilibrium.

Theorem 1. *There exists a three-mode equilibrium in* $PD((a,b,c,d),f)$ *if*

$$\phi(\boldsymbol{\beta} > \beta^*) > \frac{d-c}{a-c}\ . \tag{14}$$

Recall that β^* is a threshold such that a player of type (α,β) with $\beta > \beta^*$ prefers CDD to DDD; that is, he is a potential conditional cooperator. Theorem 1 says that a three-mode equilibrium exists if the probability $\phi(\boldsymbol{\beta} > \beta^*)$ of those potential conditional cooperators is large enough. The condition (14) provides a specific value $\frac{d-c}{a-c}$ for this probability.

The condition (14) has the following simple meaning. Suppose that no type takes the leadership behavior. In particular, consider a Bayesian strategy $\boldsymbol{s}: T \to S$ in which all the types (α,β) with $\beta > \beta^*$ follow CDD and all the types (α,β) with $\beta \leq \beta^*$ follow DDD. Then, the belief μ consistent with this Bayesian strategy is $\mu_C = \phi(T_C(\boldsymbol{s})) = 0$, $\mu_{CDD} = \phi(T_{CDD}(\boldsymbol{s})) = \phi(\boldsymbol{\beta} > \beta^*)$, and $\mu_{DDD} = \phi(T_{DDD}(\boldsymbol{s})) = \phi(\boldsymbol{\beta} \leq \beta^*)$. When the condition (14) is satisfied, a player of type $(0,0)$ prefers C to CDD and DDD given this belief μ because

$$\begin{aligned}
U_{(0,0)}(C,\mu) - U_{(0,0)}(CDD,\mu) &= [\phi(\boldsymbol{\beta} > \beta^*)a + \phi(\boldsymbol{\beta} \leq \beta^*)c] - [\phi(\boldsymbol{\beta} > \beta^*)d + \phi(\boldsymbol{\beta} \leq \beta^*)d] \\
&= (a-c)\Big[\phi(\boldsymbol{\beta} > \beta^*) - \frac{d-c}{a-c}\Big] > 0 \\
U_{(0,0)}(C,\mu) - U_{(0,0)}(DDD,\mu) &= (a-c)\Big[\phi(\boldsymbol{\beta} > \beta^*) - \frac{d-c}{a-c}\Big] > 0.
\end{aligned}$$

Thus, the condition (14) means the situation in which the pure materialist ($\alpha = \beta = 0$) has an incentive to take the leadership behavior if no one takes the leadership, all the types with $\beta > \beta^*$ follow CDD, and all the others follow DDD.

When the condition (14) holds, $\lambda(0) = \psi_{DDD}(\hat{\mu}(0)) - \hat{\mu}_{DDD}(0) < 0$ is guaranteed for the following reason. Lemma 4 states that the μ_C-fixed point $\hat{\mu}(0)$ for $\mu_C = 0$ is mapped by ψ to $\psi(\hat{\mu}(0)) = (0, \phi(\boldsymbol{\beta} > \beta^*), \phi(\boldsymbol{\beta} < \beta^*))$. Then, the above meaning of the condition (14) states that, in response to this belief $\psi(\hat{\mu}(0))$, strategy C is followed by a set of types with a positive probability measure, that is, $\psi_C(\psi(\hat{\mu}(0))) > 0$. Therefore, when we compare $\hat{\mu}(0)$ and $\psi(\hat{\mu}(0))$, the property of the belief $\hat{\mu}(0)$ that $\psi_C(\hat{\mu}(0)) = 0$ (that is, no one takes the leadership) requires that the belief $\hat{\mu}(0)$ places more weight on DDD than the belief $\psi(\hat{\mu}(0))$; that is, $\hat{\mu}_{DDD}(0) > \psi_{DDD}(\hat{\mu}(0))$.

Given its meaning, the reason why the condition (14) guarantees the existence of a three-mode equilibrium is now transparent, as follows. We consider the arc $\hat{\mu} = \hat{\mu}(\mu_C)$ in Δ, along which it holds that $\psi_C(\hat{\mu}) = \hat{\mu}_C$. The extreme points of the arc are $\hat{\mu}(0) = (0, \frac{d-c}{a-c}, \frac{a-d}{a-c})$ and $\hat{\mu}(\bar{\mu}_C) = (\bar{\mu}_C, 1-\bar{\mu}_C, 0)$. The former corresponds to the belief with the lowest μ_C on the arc and the latter corresponds to the belief with the highest μ_C on the arc. As we explained above, when the condition (14) holds, it holds that $\psi_{DDD}(\hat{\mu}(0)) < \hat{\mu}_{DDD}(0)$ at the belief $\hat{\mu}(0)$ with the lowest μ_C on the arc; that is, a best response to the belief lowers the probability of DDD-mode. (Figure 2 exhibits this case.) On the other hand, $\psi_{DDD}(\hat{\mu}(\bar{\mu}_C)) > \hat{\mu}_{DDD}(\bar{\mu}_C)$ at the belief $\hat{\mu}(\bar{\mu}_C)$ with the highest μ_C on the arc; that is, a best response to the belief raises the probability of DDD-mode. Hence, there must exist a belief between the two extreme points along the arc that maintains the probability of DDD-mode under a best response.

5.3 A characterization of the existence of the three-mode equilibrium

Theorem 1 establishes a sufficient condition (14) for the existence of the three-mode equilibrium. Then, we proceed to develop a characterization of the existence of the three-mode equilibrium.

For this purpose, let us restate Theorem 1 as a sufficient condition for the existence of the three-mode equilibrium in a space of prisoner's dilemma parameters (a, d), as follows.

Corollary 1. *For each $a \in (c, b)$, define*

$$\bar{d}(a) \equiv \phi(\boldsymbol{\beta} > \beta^*)a + (1 - \phi(\boldsymbol{\beta} > \beta^*))c. \tag{15}$$

Then,

(1) *it is $c < \bar{d}(a) < a$, $\dfrac{d}{da}\bar{d}(a) > 0$, $\lim_{a \to c} \bar{d}(a) = c$, $\lim_{a \to b} \bar{d}(a) = b$, and*

(2) *there exists a three-mode equilibrium in $PD((a, b, c, d), f)$ if*

$$c < d < \bar{d}(a) \tag{16}$$

The upper bound $\bar{d}(a)$ defined by (15) is a value of d for which the condition (14) in Theorem 1 holds in equality, that is,

$$\phi(\boldsymbol{\beta} > \beta^*) = \frac{\bar{d}(a) - c}{a - c} \quad . \tag{17}$$

This gives the area of (a, d) in the space of a normalized prisoner's dilemma for which a three-mode equilibrium is guaranteed to exist. It is illustrated by a dashed line in Figure 4.[29]

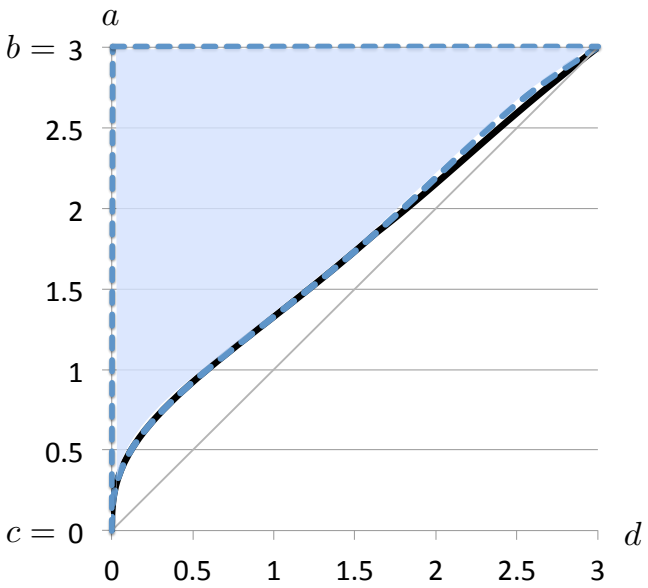


Figure 4: Sufficient condition for the existence of the three-mode equilibrium

Now, we fix f and determine for each (a, d) with $\bar{d}(a) \leq d < a$ whether there exists a three-mode equilibrium in $PD((a, b, c, d), f)$. Note that, in the definition of

[29] Figure 4 shows the area of the sufficiency condition (16) for the case of $b = 3$, $c = 0$, and $f(\alpha, \beta) = 27\{(\alpha - \frac{2}{3})^4 + \beta^4\}$.

the function $\lambda(\mu_C)$ in (13), the function $\hat{\mu}(\mu_C)$ depends on a, d and the function $\psi(\mu)$ depends on a, d so that the function $\lambda(\mu_C)$ depends on a, d. To be explicit about this fact, let us rewrite the definition as follows.

$$\lambda(\mu_C, a, d) = \psi_{DDD}(\hat{\mu}(\mu_C, a, d), a, d) - \hat{\mu}_{DDD}(\mu_C, a, d) \tag{18}$$

Similarly, let us denote $\bar{\mu}$ as $\bar{\mu}(a, d)$ to express the fact that the upper bound for μ_C that admits μ_C-fixed points depends on a, d. The function $\lambda(\mu_C, a, d)$ in (18) and its domain depend on the prisoner's dilemma parameters (a, d), as follows.

Lemma 6.

(1) *$\bar{\mu}(a, d)$ is strictly increasing in a and strictly decreasing in d.*

(2) *$\lambda(\mu_C, a, d)$ is strictly decreasing in a and strictly increasing in d.*

The intuition for why $\lambda(\mu_C, a, d)$ is strictly increasing in d is as follows.[30] Compare $PD((a, b, c, d'), f)$ and $PD((a, b, c, d''), f)$ such that $d' > d''$. Then, strategies CDD and DDD, from which a player may receive a payoff d, are more attractive in $PD((a, b, c, d'), f)$ than in $PD((a, b, c, d''), f)$. This makes the probability of DDD by best response at least as high in $PD((a, b, c, d'), f)$ as in $PD((a, b, c, d''), f)$; that is, $\psi_{DDD}(\hat{\mu}(\mu_C, a, d'), a, d') \geq \psi_{DDD}(\hat{\mu}(\mu_C, a, d''), a, d'')$. The increase in attractiveness of CDD and DDD also reduces the incentive for a player to follow strategy C in $PD((a, b, c, d'), f)$. Therefore, for a given μ_C, a μ_C-fixed point $\hat{\mu}$ that achieves the same μ_C by best response must place less weight on DDD to create a stronger incentive to follow C in $PD((a, b, c, d'), f)$ than in $PD((a, b, c, d''), f)$; that is, $\hat{\mu}_{DDD}(\mu_C, a, d') < \hat{\mu}_{DDD}(\mu_C, a, d'')$. Hence, $\lambda(\mu_C, a, d') > \lambda(\mu_C, a, d'')$.

By this monotonicity of a function λ in (a, d), we establish the following characterization that the existence of the three-mode equilibrium is monotone in (a, d).

Theorem 2. *For each $a \in (c, b)$, there exists $\hat{d}(a) \in [\bar{d}(a), a)$ such that*

(1) *there exists a three-mode equilibrium in $PD((a, b, c, d), f)$ if*

$$c < d < \hat{d}(a), \tag{19}$$

(2) *there exists no three-mode equilibrium in $PD((a, b, c, d), f)$ if*

$$\hat{d}(a) < d < a. \tag{20}$$

Furthermore, the bound $\hat{d}(a)$ is continuous, strictly increasing, $\lim_{a \to c} \hat{d}(a) = c$, and $\lim_{a \to b} \hat{d}(a) = b$.

Theorem 2 is illustrated in Figure 4. A set of prisoner's dilemma games (a, d) is partitioned by a function $\hat{d} = \hat{d}(a)$ (denoted as a bold line) into a north-west part that supports a three-mode equilibrium and a south-east part that fails to support a three-mode equilibrium.

Theorem 2 also shows that the upper bound $\hat{d}(a)$ for prisoner's dilemma games (a, d) with a three-mode equilibrium for a given a is bounded away from a. In words, the nature of the prisoner's dilemma that there is an opportunity of Pareto improvement from (d, d) to (a, a) does not guarantee successful leadership. The leadership will not emerge if the gain from the Pareto improvement is limited.

[30] The intuitions for the other claims in Lemma 6 are similar.

The threshold $\hat{d}(a)$ for the existence of the three-mode equilibrium in Theorem 2 may or may not coincide with the sufficiency bound $\bar{d}(a)$ in Corollary 1. In Appendix A, we provide a sufficient condition under which they coincide. In general, however, the threshold $\hat{d}(a)$ in Theorem 2 may be strictly higher than the sufficiency bound $\bar{d}(a)$ in Corollary 1, for the following reason. The bound $\bar{d}(a)$ corresponds to the prisoner's dilemma in which a pure materialist ($\alpha = \beta = 0$) is indifferent between taking the leadership behavior and following DDD if no type takes the leadership at all. Then, all the types, other than a pure materialist, are not willing to take the leadership behavior either given the same belief. However, although no type has an incentive to take the leadership if the other types are not expected to take the leadership, some types may be willing to take the leadership if some other types are expected to take the leadership. Then, $\bar{d}(a) < \hat{d}(a)$. Figure 5 is an example of this case $\bar{d}(a) < \hat{d}(a)$.[31]

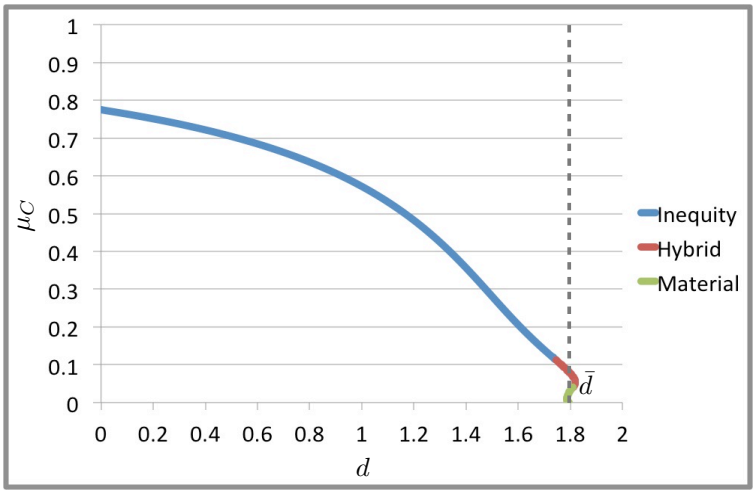


Figure 5: Equilibrium locus under $f(\alpha, \beta) = 27\{(\alpha - 2/3)^4 + \beta^4\}$

6 Comparative statics of the three-mode equilibrium

Our second result is comparative statics of the three-mode equilibrium. Parallel to our first result on the existence of the three-mode equilibrium, we fix f and conduct a comparative statics analysis with respect to the prisoner's dilemma parameters (a, d).

6.1 Comparative statics of the three-mode equilibrium distribution

We conduct a comparative statics analysis of the three-mode equilibrium distribution μ. As we noted after Theorem 2, there may exist multiple three-mode equilibria in a prisoner's dilemma and this is the case shown in Figure 3. Therefore, we pursue a monotone comparative statics analysis of the three-mode equilibrium distributions with the maximum of μ_C and with the minimum of μ_C. Recall from Lemma 5 that the value of μ_C in the three-mode equilibrium is determined by $\lambda(\mu_C, a, d) = 0$. We

[31] In addition to Theorem 2, Figure 5 demonstrates that there may exist a three-mode equilibrium in $PD((a, b, c, \hat{d}(a)), f)$ at the threshold $\hat{d}(a)$. In light of Lemma 6, it is easily verified that this occurs if and only if there exists $\mu > 0$ such that $\lambda(\mu, a, \bar{d}(a)) = 0$ at the sufficiency bound $\bar{d}(a)$ in Corollary 1.

consider

$$\mu_C^{\max}(a,d) \equiv \max\{\mu_C \in (0,1) | \lambda(\mu_C, a, d) = 0\}$$
$$\mu_C^{\min}(a,d) \equiv \min\{\mu_C \in (0,1) | \lambda(\mu_C, a, d) = 0\}.$$

The values $\mu_C^{\max}(a,d)$ and $\mu_C^{\min}(a,d)$ are well-defined because $\lambda(\mu_C, a, d)$ is continuous in μ_C so that the set of μ_C in the three-mode equilibrium is compact. Then, we define the maximum leadership distribution and the minimum leadership distribution by

$$\mu^{\max}(a,d) = (\mu_C^{\max}(a,d), \mu_{CDD}^{\max}(a,d), \mu_{DDD}^{\max}(a,d)) \equiv \hat{\mu}(\mu_C^{\max}(a,d))$$
$$\mu^{\min}(a,d) = (\mu_C^{\min}(a,d), \mu_{CDD}^{\min}(a,d), \mu_{DDD}^{\min}(a,d)) \equiv \hat{\mu}(\mu_C^{\min}(a,d)).$$

First, we conduct a comparative statics analysis of μ_C. Lemma 6 shows the monotonicity of λ in that $\lambda(\mu_C, a, d)$ is strictly decreasing in a and strictly increasing in d. This immediately leads us to the following comparative statics result.[32]

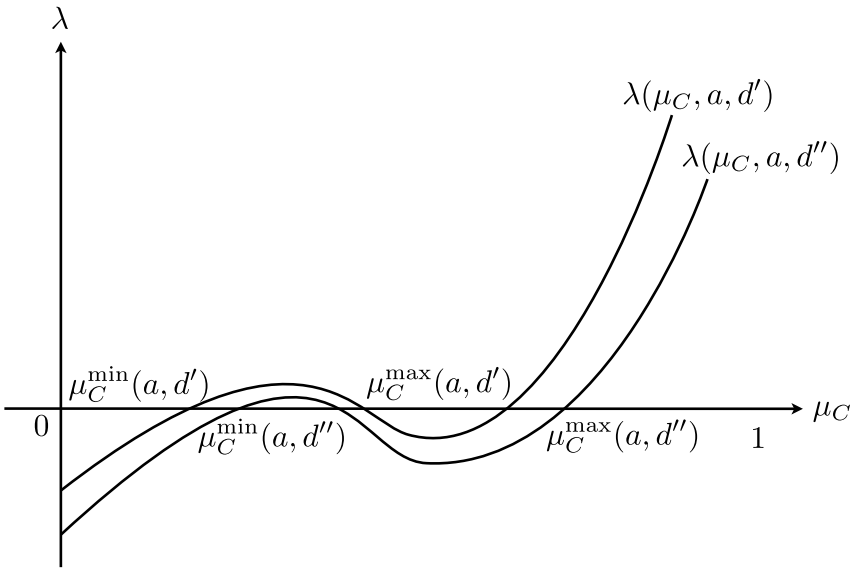
Theorem 3.

(1) *$\mu_C^{\max}(a,d)$ is strictly increasing in a and strictly decreasing in d.*

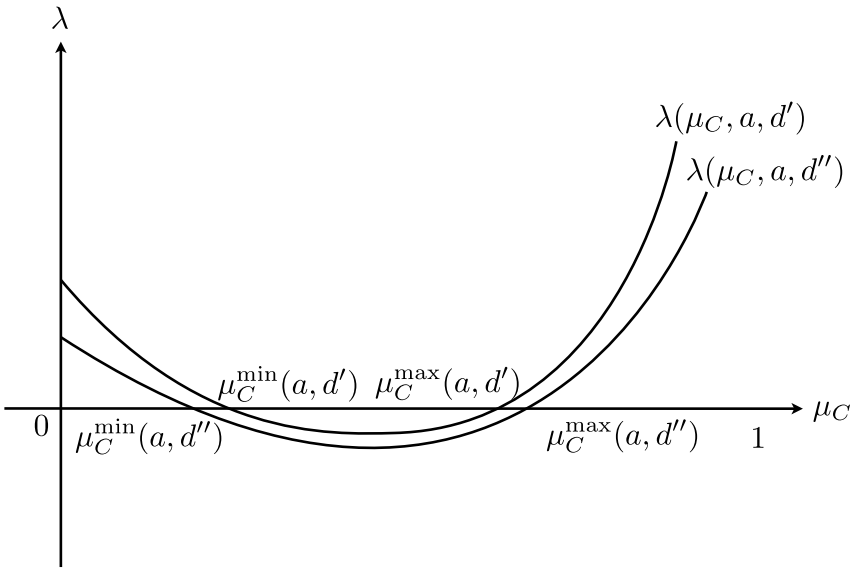
(2) *$\mu_C^{\min}(a,d)$ is strictly increasing in a and strictly decreasing in d for a range of (a,d) with $c < d < \bar{d}(a)$, while it is strictly decreasing in a and strictly increasing in d for a range of (a,d) with $\bar{d}(a) < d < \hat{d}(a)$.*

This is illustrated in Figure 6. Panel A illustrates the monotone comparative statics result for a range of (a,d) with $c < d < \bar{d}(a)$, while Panel B illustrates the monotone comparative statics result for a range of (a,d) with $\bar{d}(a) < d < \hat{d}(a)$.

[32]Theorem 3 states that $\mu_C^{\min}(a,d)$ behaves in opposite directions to $\mu_C^{\max}(a,d)$ for a range of (a,d) with $\bar{d}(a) < d < \hat{d}(a)$. This occurs because we are interested in the three-mode equilibrium. If we consider the sequential equilibrium that assigns one of C-mode, CDD-mode, and DDD-mode without restricting its distribution to the interior of Δ, the $\mu_C^{\min}(a,d)$ behaves in a parallel way to $\mu_C^{\max}(a,d)$. As we will discuss in Section 9, a no-leadership equilibrium ($\mu_C = 0$) exists for a range of (a,d) with $\bar{d}(a) < d < \hat{d}(a)$ (and $\hat{d}(a) \leq d < a$). Therefore, Theorem 3-(2) is rewritten for $\mu_C^{\min}(a,d)$ over the entire Δ as stating that $\mu_C^{\min}(a,d)$ is (strictly) increasing in a and (strictly) decreasing in d, where the strictness applies for a range of (a,d) with $c < d < \bar{d}(a)$. The same remark applies to all the results on the minimum leadership distribution hereafter. This parallel monotonicity of $\mu_C^{\min}(a,d)$ must be noticed when we test our theory in experiments using the comparative statics results. We will return to this issue in Section 9 where we discuss the existence of a no-leadership equilibrium.



Panel A



Panel B

Figure 6: Graphs of λ when $d' > d''$

Next, we consider the comparative statics of μ_{CDD}. Note in Lemma 1 that $\mu_{CDD} = \phi(T^*_{CDD}(\mu))$ for a three-mode equilibrium distribution and recall from Lemma 2 that $T^*_{CDD}(\mu)$ is determined by $\alpha^*(\mu)$ and β^* as an area of $\alpha \geq \alpha^*(\mu)$ and $\beta \geq \beta^*$ in T. We can show a comparative statics result of $\alpha^*(\mu)$ and β^*.

Lemma 7.

(1) (a) $\alpha^*(\mu^{\max}(a,d))$ *is strictly increasing in* a *and strictly decreasing in* d.

(b) $\alpha^*(\mu^{\min}(a,d))$ *is strictly increasing in* a *and strictly decreasing in* d *for a range of* (a,d) *with* $c < d < \bar{d}(a)$, *while it is strictly increasing in* d *for a*

range with $\bar{d}(a) < d < \hat{d}(a)$.[33]

(2) β^* *is strictly decreasing in* a *and constant in* d.

This comparative statics result immediately leads us to the following comparative statics of μ_{CDD}.

Theorem 4.

(1) *For a range of* (a, d) *with* $\alpha^*(\mu^{\max}(a, d)) \leq \beta^*$, $\mu^{\max}_{CDD}(a, d) = \phi(\boldsymbol{\beta} > \beta^*)$ *and it is constant in* d *and strictly increasing in* a. $\mu^{\min}_{CDD}(a, d)$ *is parallel.*

(2) (a) *For a range of* (a, d) *with* $\alpha^*(\mu^{\max}(a, d)) > \beta^*$, $\mu^{\max}_{CDD}(a, d)$ *is strictly increasing in* d.

(b) *For a range of* (a, d) *with* $\alpha^*(\mu^{\min}(a, d)) > \beta^*$, $\mu^{\min}_{CDD}(a, d)$ *is strictly increasing in* d *for a range with* $c < d < \bar{d}(a)$, *while it is strictly decreasing in* d *for a range with* $\bar{d}(a) < d < \hat{d}(a)$.

Note that $\mu^{\max}_{CDD}(a, d)$ (and $\mu^{\min}_{CDD}(a, d)$) may not be monotone with respect to a for a range of (a, d) with $\alpha^*(\mu^{\max}(a, d)) > \beta^*$ (and $\alpha^*(\mu^{\min}(a, d)) > \beta^*$) because $\alpha^*(\mu^{\max}(a, d))$ (and $\alpha^*(\mu^{\min}(a, d))$ for a range with $c < d < \bar{d}(a)$) is strictly increasing in a while β^* is strictly decreasing in a.

Finally, we consider the comparative statics of μ_{DDD}. As is stated in Theorem 4, $\mu^{\max}_{CDD}(a, d)$ and $\mu^{\min}_{CDD}(a, d)$ have a constant value $\phi(\boldsymbol{\beta} > \beta^*)$ for a range of (a, d) with $\alpha^*(\mu^{\max}(a, d)) \leq \beta^*$ and $\alpha^*(\mu^{\min}(a, d)) \leq \beta^*$, respectively. Therefore, the comparative statics of μ_{DDD} for this range is obtained from the comparative statics of μ_C in Theorem 3 and the comparative statics of β^* in Lemma 7, as follows.

Theorem 5.

(1) *For a range of* (a, d) *with* $\alpha^*(\mu^{\max}(a, d)) \leq \beta^*$, $\mu^{\max}_{DDD}(a, d)$ *is strictly decreasing in* a *and strictly increasing in* d.

(2) *For a range of* (a, d) *with* $\alpha^*(\mu^{\min}(a, d)) \leq \beta^*$, $\mu^{\min}_{DDD}(a, d)$ *is strictly decreasing in* a *and strictly increasing in* d *for a range of* (a, d) *with* $c < d < \bar{d}(a)$, *while it is strictly decreasing in* d *for a range of* (a, d) *with* $\bar{d}(a) < d < \hat{d}(a)$.

Note that $\mu^{\max}_{DDD}(a, d)$ (and $\mu^{\min}_{DDD}(a, d)$) may not be monotone with respect to a and d for a range of (a, d) with $\alpha^*(\mu^{\max}(a, d)) > \beta^*$ (and $\alpha^*(\mu^{\min}(a, d)) > \beta^*$). For example, for a range of (a, d) with $\alpha^*(\mu^{\max}(a, d)) > \beta^*$ and $c < d < \bar{d}(a)$, $\mu^{\max}_{C}(a, d)$ is strictly decreasing in d and $\mu^{\max}_{CDD}(a, d)$ is strictly increasing in d so that $\mu^{\max}_{DDD}(a, d) = 1 - \mu^{\max}_{C}(a, d) - \mu^{\max}_{CDD}(a, d)$ may or may not be increasing in d. Figure 7 below demonstrates that $\mu^{\max}_{DDD}(a, d)$ is not monotone in d.[34]

[33] $\alpha^*(\mu^{\min}(a, d))$ may not be monotone with respect to a for a range of (a, d) with $\bar{d}(a) < d < \hat{d}(a)$.

[34] The relation $\alpha^*(\mu^{\max}(a, d)) > \beta^*$ applies to the parts "Inequity" and "Hybrid" of the graph in Figure 7. The three categories of three-mode equilibrium in Figure 7 will be introduced formally in Section 7.1.

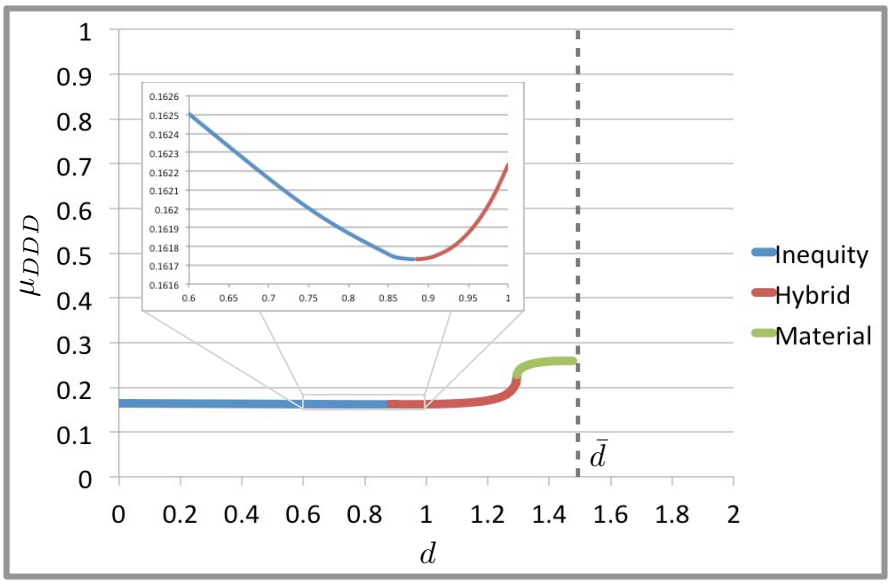


Figure 7: Non-monotonic move of μ_{DDD} under $f(\alpha, \beta) = 6\beta$

6.2 Comparative statics of outcome distributions in the three-mode equilibrium

The comparative statics of the three-mode equilibrium distribution enables us to conduct the comparative statics of the outcome distributions in the three-mode equilibrium. Recall from Table 5 that four kinds of outcomes are realized with positive probabilities in a three-mode equilibrium: (C, C) at timing 1, $C \to C$, $C \to D$, and (D, D) at timing 2. When a three-mode equilibrium with a three-mode equilibrium distribution μ prevails, probabilities for outcomes (C, C) at timing 1, $C \to C$, $C \to D$, and (D, D) at timing 2 are given by $(\mu_C)^2$, $2\mu_C\mu_{CDD}$, $2\mu_C\mu_{DDD}$, and $(\mu_{CDD} + \mu_{DDD})^2 = (1 - \mu_C)^2$. Therefore, the following is concluded from Theorems 3, 4, and 5.

Theorem 6.

(1) *The probability of the outcome of* (C, C) *at timing 1 under* $\mu^{\max}$ *is strictly increasing in* a *and strictly decreasing in* d. *The probability of the outcome of* (C, C) *at timing 1 under* $\mu^{\min}$ *is strictly increasing in* a *and strictly decreasing in* d *for a range of* (a, d) *with* $c < d < \bar{d}(a)$ *and it is strictly decreasing in* a *and strictly increasing in* d *for a range of* (a, d) *with* $\bar{d}(a) < d < \hat{d}(a)$.

(2) *The probability of the outcome of* $C \to C$ *under* $\mu^{\max}$ *is strictly increasing in* a *and strictly decreasing in* d *for a range of* (a, d) *with* $\alpha^*(\mu^{\max}(a, d)) \leq \beta^*$. *The probability of the outcome of* $C \to C$ *under* $\mu^{\min}$ *is parallel for a range of* (a, d) *with* $\alpha^*(\mu^{\max}(a, d)) \leq \beta^*$ *and* $c < d < \bar{d}(a)$, *while it is strictly increasing in* d *for a range of* (a, d) *with* $\alpha^*(\mu^{\max}(a, d)) \leq \beta^*$ *and* $\bar{d}(a) < d < \hat{d}(a)$.

(3) *The probability of the outcome of* (D, D) *at timing 2 under* $\mu^{\max}$ *is strictly decreasing in* a *and strictly increasing in* d. *The probability of the outcome of* (D, D) *at timing 2 under* $\mu^{\min}$ *is strictly decreasing in* a *and strictly increasing in* d *for a range of* (a, d) *with* $c < d < \bar{d}(a)$, *while it is strictly increasing in* a *and strictly decreasing in* d *for a range of* (a, d) *with* $\bar{d}(a) < d < \hat{d}(a)$.

The behaviors of the probabilities of the outcomes that are not stated in Theorem 6 are more complex. For example, the probability of the outcome $C \to C$ does not necessarily move monotonically with respect to a and d for the region that Theorem 6-(2) does not cover. In this region, an increase in the probability of C-mode is partly realized by having some conditional cooperator become a leader type. Because of this shift from CDD-mode to C-mode, the matching probability of a leader type and conditional cooperator becomes complex. Figure 8 shows an example of a nonmonotonic move of the probability of the outcome $C \to C$.[35] A similar complexity applies to the outcome $C \to D$.

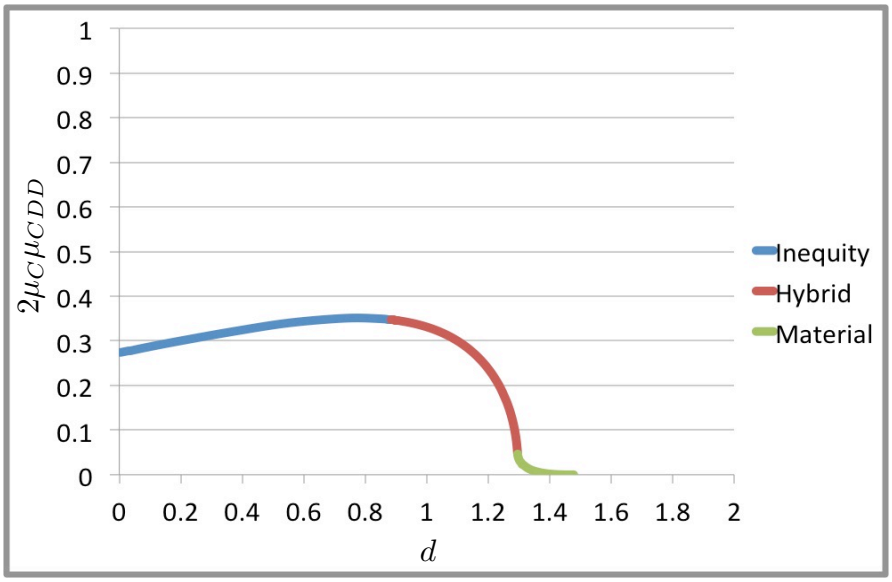


Figure 8: Non-monotonic probability of $C \to C$ under $f(\alpha, \beta) = 6\beta$

7 Who takes the leadership? - Fairness concerns, incentive to lead, and leadership patterns -

Our third result is a characterization of leadership patterns. A leadership pattern is a description of who takes the leadership and it is given by the set of leader types. When a three-mode equilibrium with a distribution μ prevails in a prisoner's dilemma game $PD((a, b, c, d), f)$, the leadership pattern of this equilibrium is the set $T_C^*(\mu)$ and this set depends on the parameters (a, b, c, d) and f in the prisoner's dilemma. We characterize the leadership patterns $T_C^*(\mu)$ in the three-mode equilibrium. In particular, we explore how each pattern is generated from the underlying incentive for each type to lead and how the incentive to lead is connected to the fairness concerns of the type.

7.1 The incentive to lead and a leadership pattern under a given belief

We begin by fixing a prisoner's dilemma $PD((a, b, c, d)f)$ and charactering the leadership pattern under a given belief; that is, a best-response type set $T_C^*(\mu)$ to an

[35] The relation $\alpha^*(\mu^{\max}(a, d)) > \beta^*$ applies to the parts "Inequity" and "Hybrid" of the graph in Figure 8.

arbitrarily given three-mode distribution μ.

To understand the best-response type set in light of the incentive to lead, let us introduce an ordering of types by their degrees of incentive to lead.

Definition 3. *Fix a positive number $\gamma > 0$. We say that a type (α, β) has a stronger incentive to lead under the belief ratio γ than a type (α', β') when*

(1) *under any three-mode distribution μ with $\frac{\mu_{DDD}}{\mu_C} = \gamma$, if the type (α', β') prefers C to CDD and DDD, then the type (α, β) also prefers C to CDD and DDD, and*

(2) *there exists a three-mode distribution μ with $\frac{\mu_{DDD}}{\mu_C} = \gamma$ under which the type (α', β') does not prefer C to CDD and DDD, while the type (α, β) prefers C to CDD and DDD.*

We say that a type (α, β) has the strongest incentive to lead under the belief ratio γ when no other type in T has a stronger incentive to lead under the belief ratio γ than the type (α, β). This type is called the strongest incentive to lead type.

The notion of the strongest incentive to lead is important for our understanding of leadership patterns. First, the best-response type set $T^*_C(\mu)$ is not empty if and only if the strongest incentive to lead type prefers C to CDD and DDD. In words, whether the leadership behavior occurs with positive probabilities is identified by whether the strongest incentive to lead type is willing to take the leadership. Second, when the best-response type set $T^*_C(\mu)$ is not empty, the strongest incentive to lead type serves as a representative element of the set $T^*_C(\mu)$. All the types in the set $T^*_C(\mu)$ are ordered according to the degree of incentive to lead from the top by the strongest incentive to lead type to the lowest on the boundaries of $T^*_C(\mu)$.

For each $\gamma > 0$, the set of the strongest incentive to lead types can be identified in the following way. A type (α, β) prefers C to CDD and DDD under a belief μ if and only if

$$U_{(\alpha,\beta)}(C, \mu) > \max[U_{(\alpha,\beta)}(CDD, \mu), U_{(\alpha,\beta)}(DDD, \mu)].$$

This condition is rewritten as

$$\mu_{CDD}(a - d) > \mu_{DDD}(d - (c - \alpha(b - c))) + \mu_C(\max[a, b - \beta(b - c)] - a). \tag{21}$$

The left-hand side is a benefit of leading; that is, a utility increase from inducing C from the opponent who follows CDD. The right-hand side is a cost of leading, which is a sum of a utility decrease from being betrayed by the opponent who follows DDD and a utility decrease from losing an opportunity to betray the opponent who follows C. The benefit of leading is independent of the type because it compares symmetric outcomes (C, C) and (D, D). Therefore, when beliefs μ with a given belief ratio $\frac{\mu_{DDD}}{\mu_C} = \gamma$ are considered, a type has a stronger incentive to lead than another type if and only if the cost of leading is lower for the former type than for the latter type. The type who has the strongest incentive to lead can be identified by minimizing the cost of leading, as follows.

Lemma 8. *Let $\gamma > 0$ be given. A set of types who has the strongest incentive to lead under the belief ratio γ is $\{(0, 0)\}$, $\{(\alpha, \beta) = t(0, 0) + (1 - t)(\beta^*, \beta^*) | 0 \leq t \leq 1\}$, and $\{(\beta^*, \beta^*)\}$ for $\gamma > 1$, $\gamma = 1$, and $\gamma < 1$, respectively.*

The intuition of Lemma 8 is straightforward. An isocost curve of leading in the type space is given by

$$\mu_{DDD}(d-(c-\alpha(b-c)))+\mu_C(\max[a,b-\beta(b-c)]-a)=\text{constant}. \tag{22}$$

For a type (α,β) with $\beta\geq\beta^*$, it holds that $\max[a,b-\beta(b-c)]-a=0$; that is, there is no cost of losing an opportunity to betray the opponent who follows C because it is the best for this type not to betray the opponent if he is to move at timing 2. Hence, the isocost curve of leading (22) is vertical for a range of types (α,β) with $\beta\geq\beta^*$. On the other hand, for a type (α,β) with $\beta<\beta^*$, both kinds of costs of leading matter, and the isocost curve of leading (22) becomes

$$\mu_{DDD}(d-(c-\alpha(b-c)))+\mu_C((b-\beta(b-c))-a)=\text{constant}.$$

The envy parameter α is the agent's evaluation of the disadvantageous inequality $b-c$ in the cost of being betrayed and the guilt parameter β is that of the advantageous inequality $b-c$ in the cost of losing the opportunity to betray. The same size of inequity $b-c$ is evaluated in the opposite direction. Therefore, the slope of the isocost curve of leading is $\frac{d\beta}{d\alpha}=\frac{\mu_{DDD}(b-c)}{\mu_C(b-c)}=\frac{\mu_{DDD}}{\mu_C}=\gamma$. The isocost curve of leading (22) corresponds to a lower cost of leading when it is located toward a lower α and a higher β. Hence, the strongest incentive to lead type must lie in a subset of the boundary of T with $\alpha=\beta$ and $0\leq\beta\leq\beta^*$. An extreme point $(\alpha,\beta)=(0,0)$ of this subset becomes a unique type of the strongest incentive to lead when $\gamma>1$. The other extreme point $(\alpha,\beta)=(\beta^*,\beta^*)$ of this subset becomes a unique type of the strongest incentive to lead when $\gamma<1$.

The two candidates for the strongest incentive to lead type, Materialist and the type (β^*,β^*), represent two extreme cases of fairness concerns. The type (β^*,β^*) maximizes β to minimize the effect of advantageous inequity on the cost of leading. Maximizing β is accompanied by raising α to β^* along the boundary of T with $\alpha=\beta$. This means that the type (β^*,β^*) is the type who places the minimum weight on the cost of losing an opportunity to betray the opponent who follows C and the maximum weight to the cost of being betrayed. In contrast, Materialist minimizes α to minimize the effect of advantageous inequity on the cost of leading. Minimizing α is accompanied by lowering β to 0. This means that Materialist is the type who places the maximum weight on the cost of losing an opportunity to betray the opponent who follows C and the minimum weight on the cost of being betrayed.

The two canonical forms of fairness concern held by Materialist and the type (β^*,β^*) leads us to classify leadership patterns according to which form of fairness concern leads to the leadership behavior. We may call a best-response type set $T_C^*(\mu)$ an *inequity concerned leader pattern* (*icl-pattern*) when $(\beta^*,\beta^*)\in T_C^*(\mu)$ and $(0,0)\notin T_C^*(\mu)$. This occurs when the type (β^*,β^*) is the unique strongest incentive to lead type and Materialist does not take the leadership. An opposite pattern in which $(0,0)\in T_C^*(\mu)$ and $(\beta^*,\beta^*)\notin T_C^*(\mu)$ may be called a *materialist leader pattern* (*ml-pattern*). This occurs when Materialist is the unique strongest incentive to lead type and the type (β^*,β^*) dose not take the leadership. When it happens to be that $(0,0)\in T_C^*(\mu)$ and $(\beta^*,\beta^*)\in T_C^*(\mu)$, we may call it a *hybrid leader pattern* (*hl-pattern*). This occurs in a particular case of $\gamma=1$ in which both Materialist and the type (β^*,β^*) have the strongest incentive to lead. However, this also occurs generically when Materialist is the unique strongest incentive to lead type and when the type (β^*,β^*) is the unique strongest incentive to lead type.

Figure 1, which we used to explain best-response type sets in Section 4, illustrates an inequity concerned leader pattern. The type (β^*,β^*) has the strongest incentive to

lead and receives a surplus $\mu_{CDD}(a-d)-\mu_{DDD}((d-c)+(b-a))$ by leading. As we move in the type space T by raising α and keeping $\beta=\beta^*$, this surplus is decreased at a rate of $\mu_{DDD}(b-c)$. It is decreased to zero when α is increased to $\alpha^*(\mu)$. The boundary of the set $T^*_C(\mu)$ is stretched out from the type $(\alpha^*(\mu),\beta^*)$ vertically for a range with $\beta>\beta^*$ and with a positive slope γ for a range with $\beta<\beta^*$. The set $T^*_C(\mu)$ is triangle shaped because $(0,0)\notin T^*_C(\mu)$.

Figure 9 illustrates a materialist leader pattern. Materialist has the strongest incentive to lead and receives a surplus $\mu_{CDD}(a-d)-\mu_{DDD}(d-c)-\mu_C(b-a)$ by leading. As we move in the type space T by raising α and keeping $\beta=0$, this surplus is decreased at a rate of $\mu_{DDD}(b-c)$. It is decreased to zero when α is increased to $\alpha^*(\mu)-\frac{\beta^*}{\gamma}$. The boundary of the set $T^*_C(\mu)$ is stretched out from the type $(\alpha^*(\mu)-\frac{\beta^*}{\gamma},0)$ with a positive slope γ. The set $T^*_C(\mu)$ is also triangle shaped because $(\beta^*,\beta^*)\notin T^*_C(\mu)$.

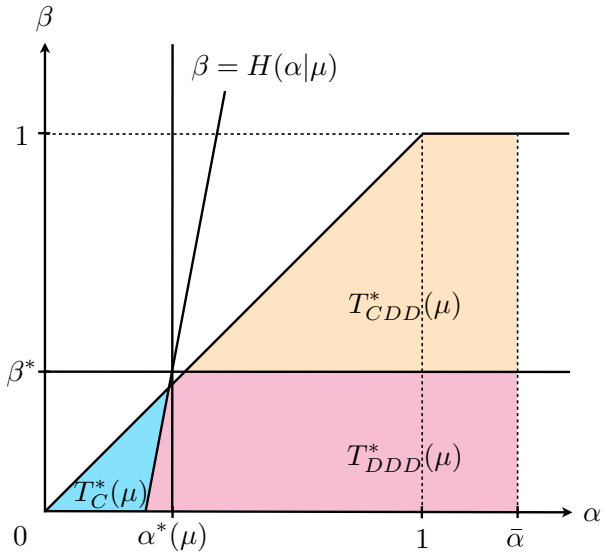
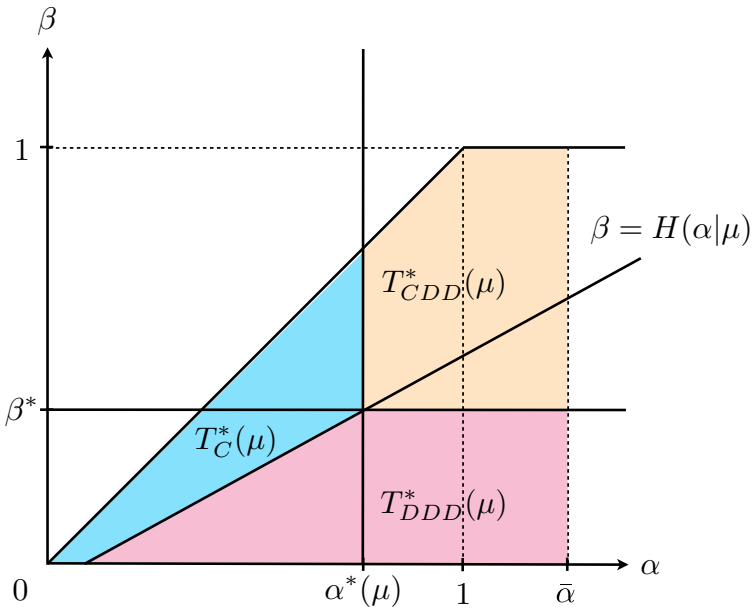
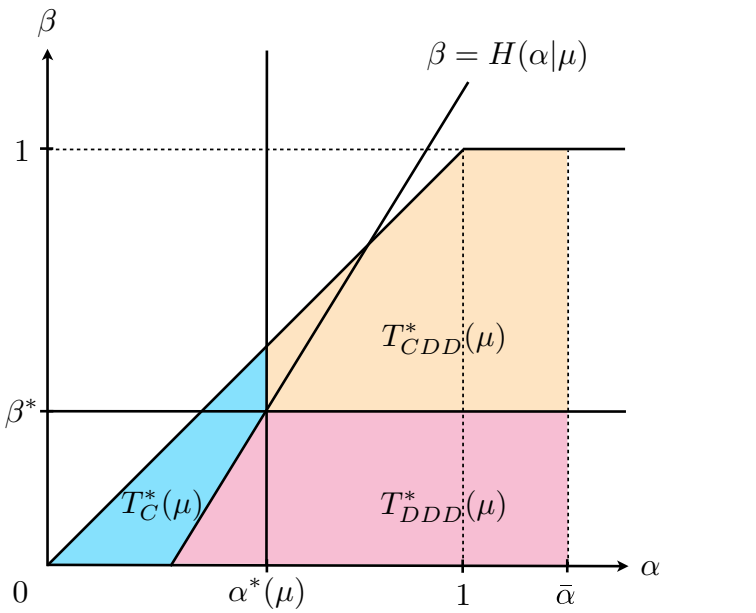


Figure 9: Materialist leader

Figure 10 illustrates hybrid leader patterns. The set $T^*_C(\mu)$ with $\gamma<1$ is a case in which the type (β^*,β^*) is the unique strongest incentive to lead type and the boundary of the set $T^*_C(\mu)$ is stretched out from the type $(\alpha^*(\mu),\beta^*)$. It is quadrilateral shaped because it also includes Materialist. The set $T^*_C(\mu)$ with $\gamma>1$ is a case in which Materialist is the unique strongest incentive to lead type and the boundary of the set $T^*_C(\mu)$ is stretched out from the type $(\alpha^*(\mu)-\frac{\beta^*}{\gamma},0)$. It is also quadrilateral shaped because it also includes the type (β^*,β^*).



Panel A: The case of $\gamma < 1$



Panel B: The case of $\gamma > 1$

Figure 10: Hybrid leader patterns

7.2 The incentive to lead and a leadership pattern in the three-mode equilibrium

We characterize the leadership pattern that prevails in the three-mode equilibrium. In the previous section, we studied the leadership pattern under a given belief by considering a three-mode distribution arbitrarily. In the three-mode equilibrium, however, the leadership pattern is determined jointly with a three-mode distribution in the equilibrium. Therefore, we explore how a three-mode equilibrium distribution is determined depending on the underlying incentive for each type to lead and how the incentive to

lead in the equilibrium is connected with the fairness concerns of the type. We explore it by studying how the leadership pattern in the three-mode equilibrium differs in each prisoner's dilemma. For this purpose, we combine the comparative statics results in Section 6 and the results on leadership patterns under a given belief in the previous Subsection 7.1, and we answer two particular questions: (1) how the leadership pattern in the three-mode equilibrium differs according to the prisoner's dilemmas, and (2) how the type of the strongest incentive to lead in the three-mode equilibrium differs according to the prisoner's dilemma.

7.2.1 Comparison of the leadership patterns

First, we compare the leadership patterns in the three-mode equilibrium for different prisoner's dilemmas. Formally, we consider how the leadership patterns $T_C^*(\mu^{max}(a,d))$ and $T_C^*(\mu^{min}(a,d))$ differ according to prisoner's dilemma parameters (a,d). To compare $T_C^*(\mu^{max}(a,d))$ and $T_C^*(\mu^{min}(a,d))$ for different values of (a,d), we introduce the following ordering of sets of types who take the leadership.

Definition 4. *Let T_C and T_C' be sets of types who take the leadership. We say that T_C is the leadership pattern of a more inequity concerned leader than T_C' if the following conditions hold.*

(1) $\exists(\hat{\alpha},\hat{\beta}) \in T_C$ s.t.$\forall(\alpha',\beta') \in T_C'; (\alpha',\beta') \leq (\hat{\alpha},\hat{\beta})$[36], *and*

(2) $\forall(\alpha',\beta') \in T_C' \setminus T_C, \forall(\alpha,\beta) \in T_C \setminus T_C'; (\alpha',\beta') \leq (\alpha,\beta)$.

Conditions (1) and (2) define a set of natural requirements for a set T_C of types that consists of more inequity concerned types than another set T_C'. Condition (1) requires that there is a type $(\hat{\alpha},\hat{\beta})$ in T_C that is dominant over T_C' in the sense that the type $(\hat{\alpha},\hat{\beta})$ is more inequity conscious in both envy and guilt for any type in T_C'. This dominance also implies that a set difference $T_C \setminus T_C'$ exists and the dominant type $(\hat{\alpha},\hat{\beta})$ lies in this difference. Condition (2) requires that the set differences $T_C \setminus T_C'$ and $T_C' \setminus T_C$ are in a specific relation such that the set difference $T_C \setminus T_C'$ is dominant over the set difference $T_C' \setminus T_C$ in the sense that any type in the set difference $T_C \setminus T_C'$ is more inequity concerned in both envy and guilt than any type in the set difference $T_C' \setminus T_C$.[37] In words, the set T_C is obtained by adding types to and extracting types from the set T_C' in such a way that any added type is more inequity conscious than any extracted type and some added type is also more inequity conscious than all the remaining types.

Recall from the analysis in Subsection 7.1 that a leadership pattern $T_C^*(\mu)$ is determined if we have three values $\alpha^*(\mu)$, β^*, and $\gamma^*(\mu)$ that define the boundary of $T_C^*(\mu)$ where we denote $\gamma^*(\mu) \equiv \frac{\mu_{DDD}}{\mu_C}$. These numbers for the maximum leadership distribution and the minimum leadership distribution depend on prisoner's dilemma parameters (a,d), as follows. (Lemma 9-(1) and (2) are restatements of Lemma 7.)

Lemma 9.

[36] We denote $(\alpha',\beta') \leq (\hat{\alpha},\hat{\beta})$ if and only if $\alpha' \leq \hat{\alpha}, \beta' \leq \hat{\beta}$, and at least one inequality holds strictly.

[37] Economists have applied various set orderings for the purpose of monotone comparative statics. The set order advocated by Topkis (1998) and Milgrom and Shannon (1994) is the usual strong order. Depending on the monotone comparative statics, however, economists sometimes need orderings that fit the objectives of their analysis. The set order by dominance in set differences is one of the naturally conceivable set orderings. Indeed, Kukushkin (2013) considers the set ordering by condition (2).

(1) $\alpha^*(\mu^{\max}(a,d),a,d)$ *is strictly increasing in* a *and strictly decreasing in* d. $\alpha^*(\mu^{\min}(a,d),a,d)$ *is strictly increasing in* a *and strictly decreasing in* d *for a range of* (a,d) *with* $c<d<\bar{d}(a)$.

(2) β^* *is strictly decreasing in* a *and independent of* d.

(3) $\gamma^*(\mu^{\max}(a,d))$ *is strictly decreasing in* a *and strictly increasing in* d. $\gamma^*(\mu^{\min}(a,d))$ *is strictly decreasing in* a *and strictly increasing in* d *for a range of* (a,d) *with* $c<d<\bar{d}(a)$.

This characterization of boundaries of the set of types who take the leadership in the three-mode equilibrium establishes the following comparison of leadership patterns.

Theorem 7.

(1) *If* $a>a'$ *and* $d<d'$, *then the leadership pattern* $T_C^*(\mu^{\max}(a,d),a,d)$ *in the maximum leadership distribution in prisoner's dilemma* $PD((a,b,c,d),f)$ *is a leadership pattern of more inequity concerned leader than the leadership pattern* $T_C^*(\mu^{\max}(a',d'),a',d')$ *in the maximum leadership distribution in prisoner's dilemma* $PD((a',b,c,d'),f)$.

(2) *If* $a>a'$, $d<d'$, *and* $d'<\bar{d}(a')$, *then the leadership pattern* $T_C^*(\mu^{\min}(a,d),a,d)$ *in the minimum leadership distribution in prisoner's dilemma* $PD((a,b,c,d),f)$ *is a leadership pattern of more inequity concerned leader than the leadership pattern* $T_C^*(\mu^{\min}(a',d'),a',d')$ *in the minimum leadership distribution in prisoner's dilemma* $PD((a',b,c,d'),f)$.

Roughly speaking, Theorem 7 states that as a is increased and d is decreased, the leadership pattern in the equilibrium becomes a more inequity concerned leader pattern.

The intuition for this comparison is as follows. For any type $(\alpha,\beta)\in T$, if a is increased and d is decreased in the condition (21) for the type (α,β) to lead under a given belief μ, the benefit of leading is increased and the cost of leading is decreased. Hence, when we compare prisoner's dilemmas $PD((a,b,c,d),f)$ and $PD((a',b,c,d'),f)$ with $a>a'$ and $d<d'$ and we imagine that the maximum leadership distribution $\mu^{\max}(a',d')$ of prisoner's dilemma $PD((a',b,c,d'),f)$ prevails as a belief in both prisoner's dilemmas, more types in prisoner's dilemma $PD((a,b,c,d),f)$ take the leadership behavior as best-responses to $\mu^{\max}(a',d')$ than in prisoner's dilemma $PD((a',b,c,d'),f)$, that is, $T_C^*(\mu^{\max}(a',d'),a',d')\subset T_C^*(\mu^{\max}(a',d'),a,d)$. Hence, the leadership pattern in $PD((a,b,c,d),f)$ tends to be of a more inequity concerned leader than the leadership pattern in $PD((a',b,c,d'),f)$ in the sense of an expanding set of types who take the leadership. However, the expansion of the set of types who take the leadership causes the probability of C-mode to be increased in $PD((a,b,c,d),f)$. This is the driving force for the result that $\mu_C^{\max}(a',d')<\mu_C^{\max}(a,d)$, which we established in Theorem 3. As we discussed in a comparison of incentives to lead between Materialist and the type (β^*,β^*), an increase in the probability of C-mode reduces the incentive to lead for those types with low guilt parameters including Materialist. The types with low guilt parameters are more inclined to follow DDD. This effect reduces the belief ratio γ from $\gamma^*(\mu^{\max}(a',d'))$ to $\gamma^*(\mu^{\max}(a,d))$. Thus, the change in probabilities of C-mode further shifts the leadership pattern in $PD((a,b,c,d),f)$ toward a more inequity concerned leader. The same explanation applies to a comparison of $T_C^*(\mu^{\min}(a,d),a,d)$ and $T_C^*(\mu^{\min}(a',d'),a',d')$ for a range of $d<d'<\bar{d}(a)$.

7.2.2 Comparison of who has the strongest incentive to lead

Second, we compare the leadership patterns in the three-mode equilibrium for different prisoner's dilemmas more specifically in terms of who has the strongest incentive to lead. Lemma 8 provides the criterion for which of Materialist or the type (β^*, β^*) is the type of the strongest incentive to lead. Together with Lemma 9-(3), we know that when we compare prisoner's dilemmas $PD((a, b, c, d), f)$ and $PD((a', b, c, d'), f)$ with $a > a'$ and $d < d'$, the type who has the strongest incentive to lead is more likely to be the type (β^*, β^*) rather than Materialist in $PD((a, b, c, d), f)$ than in $PD((a', b, c, d'), f)$. We can state this fact more precisely for the maximum leadership distribution and for the minimum leadership distribution, as follows.

Theorem 8.

(1) *For each $d \in (c, b)$, there exists $a_L^{\max}(d)$ with $\hat{d}^{-1}(d) \leq a_L^{\max}(d) < b$ such that in a prisoner's dilemma $PD((a, b, c, d), f)$ for a range $c < d < \hat{d}(a)$, the strongest incentive to lead type in the three-mode equilibrium distribution $\mu^{\max}(a, d)$ is Materialist if $a < a_L^{\max}(d)$ and the type (β^*, β^*) if $a > a_L^{\max}(d)$.*

(2) *For each $d \in (c, b)$, there exists $a_L^{\min}(d)$ with $\bar{d}^{-1}(d) \leq a_L^{\min}(d) < b$ such that in a prisoner's dilemma $PD((a, b, c, d), f)$ for a range $c < d < \bar{d}(a)$, the strongest incentive to lead type in the three-mode equilibrium distribution $\mu^{\min}(a, d)$ is Materialist if $a < a_L^{\min}(d)$ and the type (β^*, β^*) if $a > a_L^{\min}(d)$.*

(3) *The two functions $a = a_L^{\max}(d)$ and $a = a_L^{\min}(d)$ are increasing and related as $a_L^{\max}(d) \leq a_L^{\min}(d)$ for any $d \in (c, b)$.*

Theorem 8 states that a class of prisoner's dilemma that admits a three-mode equilibrium is divided into a Materialist class (a subclass in which Materialist is the unique strongest incentive to lead type) and a β^*-class (a subclass in which the type (β^*, β^*) is the unique strongest incentive to lead type). Furthermore, roughly speaking, the β^*-class is located north-west (a higher a and lower d) of the Materialist class.

Theorem 8 together with Theorem 7 suggests the following relationship between a prisoner's dilemma and a leadership pattern in the three-mode equilibrium in the dilemma. Suppose that some prisoner's dilemma $PD((a, b, c, d), f)$ has a three-mode equilibrium with the materialist leader pattern.[38] Then, as we move from $PD((a, b, c, d), f)$ by lowering d and raising a, the corresponding leadership patterns in the three-mode equilibrium turn from a materialist leader pattern into a hybrid pattern and then an inequity concerned pattern.[39] Figure 11 demonstrates that this leadership pattern transition occurs for a particular prior $f(\alpha, \beta) = 6\beta$.

[38] Note that although Theorem 8 guarantees that the β^*-class is nonempty, it does not guarantee that the Materialist class is. If it were the case that $\hat{d}^{-1}(d) = a_L^{\max}(d)$ and $\bar{d}^{-1}(d) = a_L^{\min}(d)$ for all $d \in (c, b)$, then every prisoner's dilemma with the three-mode equilibrium is in the β^*-class. Then, the materialist leader pattern would never prevail in the equilibrium.

[39] Theorem 8 guarantees that a hybrid pattern prevails in some prisoner's dilemma $PD((a, b, c, d), f)$ because, as long as a is increased close enough to b, $PD((a, b, c, d), f)$ belongs to the β^*-class. However, as we will explore below in Appendix B, it may be the case for some prior f that an inequity concerned pattern never prevails in the three-mode equilibrium.

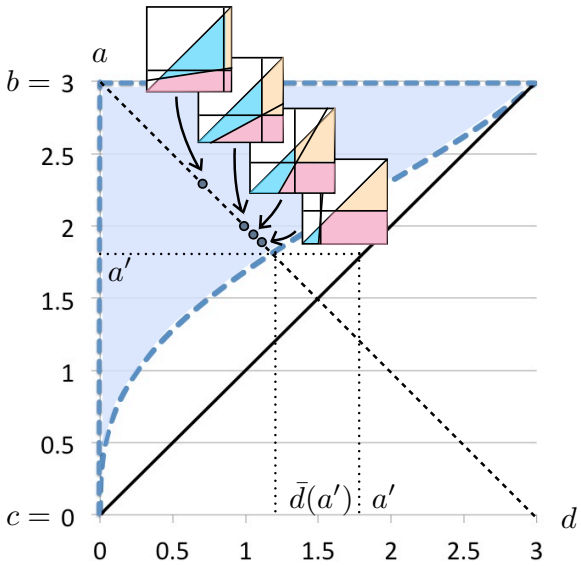


Figure 11: Leadership pattern transitions

More precisely, we can provide a sufficient condition on the prisoner's dilemma under which the leadership pattern in the three-mode equilibrium is a materialist leader pattern.

Theorem 9. *Let* $\bar{d}_{ml}(a) \equiv a - (b-c)\min_{\alpha^* \in [0,\bar{\alpha}]} \phi(\boldsymbol{\beta} < \beta^* | \alpha^* \leq \alpha)$. *Suppose that* $\min_{\alpha^* \in [0,\bar{\alpha}]} \phi(\boldsymbol{\beta} < \beta^* | \alpha^* \leq \alpha) > 0$. *Then,*

(1) $\bar{d}_{ml}(a)$ *is strictly increasing,* $\bar{d}_{ml}(a) < a$, $\lim_{a \to c} \bar{d}_{ml}(a) = c - (b-c)$, $\lim_{a \to b} \bar{d}_{ml}(a) = b$, *and there exists* $\bar{a}_{ml} \in (c, b)$ *such that* $\bar{d}_{ml}(\bar{a}_{ml}) = c$.

(2) *for each* $a \in (c, \bar{a}'_{ml})$, *if* $\max(c, \bar{d}_{ml}(a)) < d < \hat{d}(a)$, *then the leadership pattern in the three-mode equilibrium in* $PD((a,b,c,d), f)$ *is a materialist leader pattern.*

Figure 12 illustrates Theorem 9.[40] The curve $d = \hat{d}(a)$, which is identified in Theorem 2, divides the space of a normalized prisoner's dilemma and a three-mode equilibrium exists in $PD((a,b,c,d), f)$ if and only if it is located north-west of the curve $d = \hat{d}(a)$. The curve $d = \bar{d}_{ml}(a)$ crosses the curve $d = \hat{d}(a)$. The area between the curve $d = \bar{d}_{ml}(a)$ and the curve $d = \hat{d}(a)$ is a subset of a prisoner's dilemma in which the leadership pattern in the three-mode equilibrium in $PD((a,b,c,d), f)$ is guaranteed to be a materialist leader pattern.

[40] Figure 12 is drawn for the uniform distribution over $T = \{(\alpha, \beta) | \alpha \geq \beta, 0 \leq \alpha, \beta \leq 1\}$, for which $\min_{\alpha^* \in [0,\bar{\alpha}]} \phi(\boldsymbol{\beta} < \beta^* | \alpha^* \leq \alpha) = \beta^* > 0$.

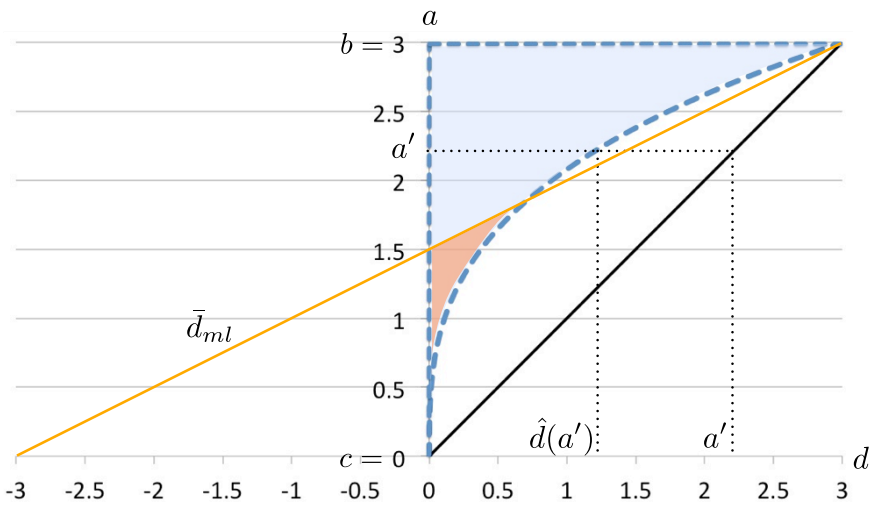


Figure 12: Sufficient condition for materialist leader pattern under $f(\alpha, \beta) = 2$

The bound $\bar{d}_{ml}(a)$ identified in Theorem 9 has the following simple meaning. Take any $\alpha^* \geq \beta^*$ and suppose that all the types (α, β) with $\alpha < \alpha^*$ follow C, while a type (α, β) with $\alpha > \alpha^*$ follows CDD if $\beta > \beta^*$ and DDD if $\beta < \beta^*$. Then, the belief μ that is consistent with this Bayesian strategy $\boldsymbol{s} : T \to S$ is $\mu_C = \phi(T_C(\boldsymbol{s})) = \phi(\alpha^* > \boldsymbol{\alpha})$, $\mu_{CDD} = \phi(T_{CDD}(\boldsymbol{s})) = \phi(\alpha^* \leq \boldsymbol{\alpha}, \boldsymbol{\beta} \geq \beta^*)$, and $\mu_{DDD} = \phi(T_{DDD}(\boldsymbol{s})) = \phi(\alpha^* \leq \boldsymbol{\alpha}, \boldsymbol{\beta} < \beta^*)$. Consider a player of type (β^*, β^*). When $d > \bar{d}_{ml}(a)$, he prefers CDD to C given this belief μ because

$$U_{(\beta^*,\beta^*)}(CDD, \mu) - U_{(\beta^*,\beta^*)}(C, \mu) = (b - c)\phi(\alpha^* \leq \boldsymbol{\alpha})\left[\phi(\boldsymbol{\beta} < \beta^* | \alpha^* \leq \boldsymbol{\alpha}) - \frac{a - d}{b - c}\right] > 0.$$

Thus, the condition $d > \bar{d}_{ml}(a)$ means a situation in which the type (β^*, β^*) has an incentive not to take the leadership behavior if all the types with $\alpha < \alpha^*$ take the leadership, all the types with $\alpha \geq \alpha^*$ and $\beta \geq \beta^*$ follow CDD, and all the others follow DDD. In such a situation, the type (β^*, β^*) never takes the leadership in any equilibrium, which would mean $\alpha^* = \alpha^*(\mu) > \beta^*$ for the corresponding belief μ. Therefore, only the three-mode equilibrium with a materialist leader pattern exists.

Although a materialist leader pattern prevails in a three-mode equilibrium in some prisoner's dilemmas for any prior f, there may exist no three-mode equilibrium with an inequity concerned leader pattern in any prisoner's dilemma, depending on a prior f. The uniform distribution over a type space $T = \{(\alpha, \beta) | \alpha \geq \beta, 0 \leq \alpha, \beta \leq 1\}$ is an example (see Appendix B). However, we can provide a sufficient condition for the prisoner's dilemma and a prior under which the leadership pattern in the three-mode equilibrium is an inequity concerned leader pattern.

Theorem 10. *For a, b, c with $c < a < b$ and f given, if*

$$\phi\Big(\boldsymbol{\alpha} < \beta^*(1 - \beta^*)\frac{\phi(\boldsymbol{\beta} > \beta^*)}{1 - \phi(\boldsymbol{\beta} > \beta^*)}\Big|\boldsymbol{\beta} > \beta^*\Big) > 1 - \beta^*, \tag{23}$$

then there exists $\bar{d}_{icl}(a) \in (c, a)$ such that

(1) $c < \bar{d}_{icl}(a) < \bar{d}(a)$

(2) *for any $d \in (c, \bar{d}_{icl}(a))$, there exists a three-mode equilibrium in $PD((a, b, c, d), f)$ and any three-mode equilibrium is an inequity concerned leader pattern.*

The intuition of Theorem 10 is as follows. A type (α, β) with $\beta > \beta^*$ is the type who is a potential conditional cooperator. The condition (23) requires that the distribution f over the set of potential conditional cooperators is more concentrated toward lower envy types and provides an explicit threshold $\beta^*(1-\beta^*)\frac{\phi(\boldsymbol{\beta}>\beta^*)}{1-\phi(\boldsymbol{\beta}>\beta^*)}$ for the envy parameter and an explicit bound $1-\beta^*$ for the probability of an envy parameter being lower than the threshold. Note from Lemma 2 that among the potential conditional cooperators it is those with lower envy parameters that in fact take the leadership and choose C at timing 1 rather than behave as conditional cooperators. Therefore, the condition (23) makes the distribution of C and CDD chosen by the potential conditional cooperators concentrated toward a choice of C. This induces Materialist to follow DDD and betray the choice of C made by the potential conditional cooperators with low envy parameters rather than to follow C and take the leadership himself in expectation of inducing the conditional cooperator with high envy parameters to respond with C. Hence, only an inequity concerned leader pattern is possible.

8 An extension to many timings

So far we have considered a prisoner's dilemma in which a player chooses C or D at one of two timings he prefers. The theory, however, applies to a prisoner's dilemma with many timings in general. The mechanism of leadership elucidated by our analysis is also valid and our conclusion on the resolution of social dilemmas remains unchanged even if we allow a player to choose his timing of move from not only two but many timings.[41]

Suppose that there are K possible timings from timing 1 to timing K with $2 \leq K < \infty$. Consider a prisoner's dilemma in which a player is allowed to choose C or D at a timing from K possible timings. We can extend the notion of a three-mode strategy naturally to this prisoner's dilemma with K timings. C_k-mode is a strategy that commands a player to take the leadership behavior at timing k; that is, a strategy that prescribes $\emptyset$ for timings 1 to $k-1$ and C for timing k when his opponent chooses $\emptyset$ at all the timings from 1 through $k-1$.[42] CDD_k-mode is a strategy that prescribes $\emptyset$ for timings 1 to $k-1$, and prescribes C for timing k if his opponent chooses C before timing k, D if his opponent chooses D before timing k, and D if his opponent chooses $\emptyset$ at all the timings 1 through $k-1$. DDD_k-mode is parallel to CDD_k-mode. A Bayesian strategy is a three-mode strategy for a prisoner's dilemma with K timings if there exists k^* with $1 \leq k^* < K$ such that a player follows either a C_k-mode with $k \leq k^*$, a CDD_k-mode with $k > k^*$, or a DDD_k-mode with $k > k^*$, and a probability of some C_k-mode being followed, a probability of some CDD_k-mode being followed, and a probability of some DDD_k-mode being followed are positive and summed to one; that is, $0 < \phi(\cup_{1\leq k\leq k^*} T_{C_k}), \phi(\cup_{k^*<k\leq K} T_{CDD_k}), \phi(\cup_{k^*<k\leq K} T_{DDD_k}) < 1$, and $\phi(\cup_{1\leq k\leq k^*} T_{C_k}) + \phi(\cup_{k^*<k\leq K} T_{CDD_k}) + \phi(\cup_{k^*<k\leq K} T_{DDD_k}) = 1$ where T_{C_k}, T_{CDD_k}, T_{DDD_k} denote the set of types who follow C_k-mode, CDD_k-mode, and DDD_k-mode

[41] We thank Klaus Schmidt for suggesting this point to us.

[42] Precisely speaking, there are two distinguished C_k-mode, depending on the prescriptions when the opponent moves before the prescribed leadership timing k. One C_k-mode prescribes the response C at timing k to his opponent's choice of C before timing k and the response D to his opponent's choice of D. The other C_k-mode prescribes the response with D at timing k to any choice by the opponent before timing k. When a player follows C_k-mode in a three-mode strategy, a player with type $\beta \geq \beta^*$ follows the former C_k-mode, while a player with type $\beta < \beta^*$ follows the latter C_k-mode. This distinction of two distinguished modes of leadership did not arise in a prisoner's dilemma with two timings, because there is no pretiming before timing 1, for which C-mode prescribes the leadership behavior of choosing C.

respectively. A three-mode strategy with a threshold timing k^* divides K timings into a span of earlier timings $k = 1, \cdots, k^*$ in which the leadership behavior occurs and a span of later timings $k = k^* + 1, \cdots, K$ in which the follower behaviors occur.

A three-mode strategy in a prisoner's dilemma with K timings is called a two-timing three-mode strategy if it is a particular three-mode strategy with k^* such that a player follows one of C_{k^*}-mode, CDD_{k^*+1}-mode, or DDD_{k^*+1}-mode. A two-timing three-mode strategy is special in that a player is supposed to move with positive probabilities at only two adjacent timings among K available timings.

There is a natural correspondence of a two-timing three-mode strategy with a three-mode strategy in a prisoner's dilemma with two timings. Consider a three-mode strategy $\boldsymbol{s}$ in a prisoner's dilemma with two timings. Then, for each k^* with $1 \leq k^* < K$, the naturally corresponding two-timing three-mode strategy with a threshold k^* in a prisoner's dilemma with K timings assigns C_{k^*}-mode, CDD_{k^*+1}-mode, and DDD_{k^*+1}-mode to those types $(\alpha, \beta) \in T_C(\boldsymbol{s})$, $(\alpha, \beta) \in T_{CDD}(\boldsymbol{s})$, and $(\alpha, \beta) \in T_{DDD}(\boldsymbol{s})$, respectively. This natural correspondence is shown to preserve the equilibrium condition, as follows.

Theorem 11. *Consider a three-mode strategy $\boldsymbol{s}$ in a prisoner's dilemma with two timings. Then, the corresponding two-timing three-mode strategy in a prisoner's dilemma with K timings is a sequential equilibrium if and only if $\boldsymbol{s}$ is a sequential equilibrium.*

When a two-timing three-mode strategy with a threshold k^* is played in a prisoner's dilemma with K timings, there are opportunities for players to take the leadership behavior at timings $k = 1, \cdots, k^* - 1$ ahead of any planned moves of his opponent. However, these opportunities are not taken. The reason is that the benefit of leadership is to induce the choice of C from the opponent who follows CDD_{k^*+1}. This benefit is fully utilized if the leadership behavior is taken before $k^* + 1$. On the other hand, taking the leadership behavior of choosing C before k^* opens an opportunity for the opponent to observe this choice of C irrespective of the opponent's strategy. This is harmful if the opponent who follows C_{k^*} is a type (α, β) with $\beta < \beta^*$. This opponent responds with D to a choice of C made at any timing k before k^*, while the opponent chooses C given that no player moves before k^*. Thus, taking the leadership behavior before timing k^* simply increases the risk of being betrayed. This force makes a player concentrate on timing k^* for the leadership behavior.

The force of betrayal risk of early leadership exists when a player considers the leadership in any three-mode strategy. We can show that any three-mode equilibrium with a threshold k^* in a prisoner's dilemma with K timings is essentially a two-timing three-mode strategy in the following sense.

Theorem 12. *Consider a three-mode strategy with a threshold k^* in a prisoner's dilemma with K timings. If it is a sequential equilibrium, then*

(1) $\phi(T_{C_1}) = \cdots = \phi(T_{C_{k^*-1}}) = 0$ *and* $\phi(T_{C_{k^*}}) > 0$, *and*

(2) $\phi(T_{CDD_{k^*+1}}) > 0$.

Property (1) means that the leadership behavior must take place at exactly one timing k^* in a three-mode equilibrium with a threshold k^*. Property (2) means that the follower behavior based on the conditional cooperation must emerge with a positive probability at the adjacent timing $k^* + 1$.

An inessential delay of CDD_k-mode or DDD_k-mode with $k > k^* + 1$ may be possible in a three-mode equilibrium with a threshold k^*. For example, if $\boldsymbol{s}$ is a three-mode equilibrium in a prisoner's dilemma with two timings, then a combination of

C_{k^*}-mode for $T_C(\boldsymbol{s})$, CDD_{k^*+1}-mode for $T_{CDD}(\boldsymbol{s})$, and DDD_k-mode for $T_{DDD}(\boldsymbol{s})$ for any k with $k^* + 1 \leq k \leq K$ is supported by a sequential equilibrium in a prisoner's dilemma with K timings.

In spite of this kind of possible delay, however, a distribution $(\phi(T_{C_{k^*}}), \phi(\cup_{k^*<k\leq K} T_{CDD_k}), \phi(\cup_{k^*<k\leq K} T_{DDD_k}))$ over C-mode, CDD-mode, and DDD-mode in any three-mode equilibrium in a prisoner's dilemma with K timings must correspond to some three-mode equilibrium distribution μ in a prisoner's dilemma with two timings in that $\mu_C = \phi(T_{C_{k^*}})$, $\mu_{CDD} = \phi(\cup_{k^*<k\leq K} T_{CDD_k})$, and $\mu_{DDD} = \phi(\cup_{k^*<k\leq K} T_{DDD_k})$, because it is also a three-mode equilibrium for any type who follows a CDD_k-mode or DDD_k-mode with $k > k^* + 1$ to follow CDD_{k^*+1}-mode or DDD_{k^*+1}-mode, respectively. Hence, as far as an outcome distribution is concerned, our conclusion in a prisoner's dilemma with two timings remains unchanged when we allow many timings in a prisoner's dilemma.

9 Discussion

We studied the leadership in a prisoner's dilemma by considering a three-mode equilibrium in a Bayesian model of a prisoner's dilemma with endogenous moves. We briefly discuss some issues that we excluded from our analysis.

9.1 Other equilibria

Although we focused on the three-mode equilibrium, there may exist other equilibria in PD.

9.1.1 Other leadership equilibrium

We studied the leadership in a prisoner's dilemma with the notion of the three-mode equilibrium in which players follow one of three behavior modes C, CDD, or DDD. Discussion on other possibilities of the emergence of leadership is in order.

One can show that the leadership is never realized by an equilibrium in which players follow one particular behavior mode or one of two behavior modes. However, when there exists a three-mode equilibrium in a prisoner's dilemma, leadership can be realized by an alternative equilibrium in which players follow a behavior mode from a different set of three behavior modes: C, CDD, and D that prescribes a choice of D at timing 1. The type who follows DDD in our three-mode equilibrium follows D in this alternative equilibrium. The incentive for players to follow D in the alternative equilibrium is similar to the incentive for them to follow DDD in our three-mode equilibrium. Therefore, the alternative equilibrium can be regarded as substantially the same as our three-mode equilibrium.

There might exist an equilibrium with a more complex set of behavior modes in some prisoner's dilemma. From the above discussion, an obvious one is a four-mode equilibrium in which players follow one of four behavior modes: C, CDD, DDD, or D. This is a combination of our three-mode equilibrium and the alternative three-mode equilibrium.

Furthermore, one can consider a five-mode equilibrium in which players follow one of five behavior modes: C, CDD, DDD, D, or CDC, which prescribes $a_C = C$, $a_D = D$, and $a_\emptyset = C$. The additional behavior mode CDC differs from CDD in that a player chooses C at timing 2 when his opponent postpones his choice to timing 2.

We can verify that the abovementioned equilibria exhaust the list of equilibrium in which leadership is realized. We focused on our three-mode equilibrium in our analysis of leadership in a prisoner's dilemma because it is the simplest manner in which leadership is realized in a prisoner's dilemma.

9.1.2 No-leadership equilibrium

We studied the emergence of leadership in equilibrium in a prisoner's dilemma. However, there may exist in the same prisoner's dilemma another sequential equilibrium in which both players choose D.

In particular, consider a wait-and-defect equilibrium in which all the types postpone their choices to timing 2 and then choose D. From the analysis leading to Lemma 2 and Lemma 3, it is easy to see that this sequential equilibrium must assign CDD to those types with $\beta > \beta^*$ and DDD to those types with $\beta < \beta^*$. By the discussion after Theorem 1, such a strategy is a sequential equilibrium strategy if and only if the $(0,0)$ type has no incentive to deviate to choosing C at timing 1; that is, if and only if $\bar{d}(a) \leq d$ where $\bar{d}(a)$ is the sufficiency bound for the existence of a three-mode equilibrium in Corollary 1.

Recall from Theorem 2 that there exists a threshold $\hat{d}(a)$ such that there exists a three-mode equilibrium if and only if $d < \hat{d}(a)$, where $\bar{d}(a) \leq \hat{d}(a)$ and the threshold $\hat{d}(a)$ may or may not coincide with the sufficiency bound $\bar{d}(a)$. Then, we have conditions for combinations of the (C,C) outcome by the three-mode equilibrium and the (D,D) outcome by the wait-and-defect equilibrium in terms of payoff parameters a, b, c, d and a type density f. For $d < \bar{d}(a)$, a three-mode equilibrium exists and the wait-and-defect equilibrium does not exist. Then, it is not possible for all the players to wait and choose D at timing 2. For $\bar{d}(a) \leq d < \hat{d}(a)$, both a three-mode equilibrium and the wait-and-defect equilibrium exist.[43] Then, the players need to coordinate on a three-mode equilibrium to achieve cooperation. For $\hat{d}(a) < d$, a three-mode equilibrium does not exist and the wait-and-defect equilibrium exists. Then, it is not possible to achieve cooperation through the leadership by a three-mode equilibrium.

For a prisoner's dilemma with $d < \bar{d}(a)$ in which a three-mode equilibrium exists and the wait-and-defect equilibrium does not exist, there still remains the possibility of the (D,D) outcome by a wider class of equilibria $\boldsymbol{s} : T \to S$ that assigns to each type (α, β) either D, CDD, or DDD. This strategy prescribes an early defection at timing 1 for those types who are assigned D. The analysis leading to Lemma 2 and Lemma 3 states that CDD must be assigned only to those types with $\beta \geq \beta^*$ and DDD must be assigned only to those types with $\beta \leq \beta^*$. This means that $0 \leq \phi(T_{CDD}(\boldsymbol{s})) \leq \phi(\boldsymbol{\beta} > \beta^*)$. By the same argument as the discussion after Theorem 1, such a $\boldsymbol{s}$ is a sequential equilibrium strategy if and only if the $(0,0)$ type has no incentive to deviate from $\boldsymbol{s}$ by choosing C at timing 1; that is, if and only if $\phi(T_{CDD}(\boldsymbol{s}))a + (1 - \phi(T_{CDD}(\boldsymbol{s})))c \leq d$. This condition is rewritten as $0 \leq \phi(T_{CDD}(\boldsymbol{s})) \leq \bar{\mu}_{CDD}(d)$ by defining $\bar{\mu}_{CDD}(d) = \min(\frac{d-c}{a-c}, \phi(\boldsymbol{\beta} > \beta^*))$. When $d < \bar{d}(a)$, one can find such a

[43] As we remarked in footnote 32, this fact must be noted when we test our theory by experiments. The theory of the three-mode equilibrium implies that the leadership behavior should be observed at a nonnegligible frequency in a prisoner's dilemma with $\bar{d}(a) < d < \hat{d}(a)$. However, the existence of a no-leadership equilibrium may cause subjects to rarely take the leadership behavior in such a prisoner's dilemma. Furthermore, when subjects play various prisoner's dilemmas with $c < d < \hat{d}(a)$, the observed frequency of the leadership behavior may exhibit a monotone decrease in d over the whole range $(c, \hat{d}(a))$, although the comparative statics results of the three-mode equilibrium (Theorem 3) states that $\mu_C^{\min}(a,d)$ for the three-mode equilibrium is strictly decreasing in d over the range $(c, \bar{d}(a))$ and strictly increasing in d over the range $(\bar{d}(a), \hat{d}(a))$. Hence, we should not rush to overthrow our theory of leadership even if the data show the overall monotonicity rather than the asymmetry at $\bar{d}(a)$.

$\boldsymbol{s}$ that $0 \leq \phi(T_{CDD}(\boldsymbol{s})) \leq \bar{\mu}_{CDD}(d) < \phi(\boldsymbol{\beta} > \beta^*)$. This means that if many types who would become conditional cooperators (that is, those types with $\beta > \beta^*$ with a probability more than $\phi(\boldsymbol{\beta} > \beta^*) - \bar{\mu}_{CDD}(d) > 0$) do not wait until timing 2 and choose D early at timing 1, the uncooperative outcome (D, D) prevails in equilibrium. Then, to achieve cooperation through leadership, the players need to coordinate on a three-mode equilibrium.

9.1.3 Cooperation without leadership

We studied the issue of time and social dilemmas by focusing on a mechanism in which the leadership resolves social dilemmas through dynamic decision making. However, even when agents have the freedom to choose the timing of moves, cooperation may be realized without leadership.

We say that a Bayesian strategy involves the leadership if it generates a positive probability for an outcome in which an agent chooses C at timing 1 and the other agent waits and chooses C at timing 2. We say that a Bayesian strategy supports cooperation without leadership if it does not involve the leadership and it generates a positive probability for an outcome in which both agents choose C. Consider the following Bayesian strategy $\boldsymbol{s} : T \to S$ which assigns either DDD, CDD, CDC, or D under a belief $\mu \in \Delta(\{DDD, CDD, CDC, D\})$.

$$\boldsymbol{s}(\alpha, \beta) = \begin{cases} DDD & \text{if } \beta \leq \beta^* \\ CDD & \text{if } \beta^* < \beta \leq \min[\frac{b-d}{b-c}, \beta^* + \frac{\mu_{CDD}+\mu_{DDD}}{\mu_{CDC}}(\alpha + \frac{d-c}{b-c})] \\ CDC & \text{if } \alpha \leq \frac{\mu_{CDC}}{\mu_{CDD}+\mu_{DDD}}\frac{a-d}{d-c} - \frac{d-c}{b-c} \text{ and } \beta > \beta^* + \frac{\mu_{CDD}+\mu_{DDD}}{\mu_{CDC}}(\alpha + \frac{d-c}{b-c}) \\ D & \text{if } \alpha > \frac{\mu_{CDC}}{\mu_{CDD}+\mu_{DDD}}\frac{a-d}{d-c} - \frac{d-c}{b-c} \text{ and } \beta > \frac{b-d}{b-c}\ . \end{cases} \tag{24}$$

This Bayesian strategy does not involve the leadership because it does not assign C. However, when both agents are assigned CDC, they wait and choose C at timing 2. Hence, this Bayesian strategy supports cooperation without leadership.

One can show that the Bayesian strategy (24) may be a sequential equilibrium, depending on a prior f[44], and that there is no other sequential equilibrium that supports cooperation without leadership.

The cooperation occurs by simultaneous choices of C at timing 2 in the equilibrium in the Bayesian strategy (24). However, note that the freedom to choose the timing of moves plays a crucial role in realizing this cooperation. Although the types who follow DDD, CDD, and CDC choose to move at timing 2, the types who follow D choose to move at timing 1. The types who are assigned D are those types who are endowed with the highest α and the highest β. They most dislike both being betrayed and betraying. These types prefer choosing D at timing 1 because, as long as they wait until timing 2, they cannot avoid either being betrayed by the types who follow DDD and CDD or betraying the types who follow CDC, so that choosing D at timing 1 is the only way to guarantee (D, D), which involves neither being betrayed nor betraying. This behavior by the most fairness concerned types makes the other types refrain from taking the leadership. In particular, the types who follow CDC are willing to choose C at timing 2 in spite of the risk of being betrayed by the types who follow DDD and

[44] To see this, consider a degenerate prior that assigns a probability 0.2 to the type $(0, 0)$, a probability 0.7 to the type $(\frac{2}{3}, \frac{2}{3})$, and a probability 0.1 to the type $(1, 1)$. Then, consider a special case of the Bayesian strategy (24) that assigns DDD to the type $(0, 0)$, CDC to the type $(\frac{2}{3}, \frac{2}{3})$, and D to the type $(1, 1)$. Then, this strategy is a sequential equilibrium in a prisoner's dilemma with $a = 2, b = 3, c = 0, d = 1$. Smooth out this degenerate prior over a type space $T = \{(\alpha, \beta) | 0 \leq \alpha \leq 1, \beta \leq \alpha\}$. Then, a Bayesian strategy (24) under the resulting prior continues to be a sequential equilibrium.

CDD. If they choose C at timing 1 instead of timing 2, they would be better off if there were no types who follow D, because those types who follow CDD will respond with C. In fact, however, choosing C at timing 1 also entails an additional risk of being betrayed by the type who follows D, which can be avoided if they stick to the assigned CDC-mode. Thus, it is essential for the cooperation at timing 2 that the most fairness concerned types choose to move at timing 1 and the other types choose to move at timing 2 under the freedom to choose the timing of moves.

9.2 Other models of a prisoner's dilemma in the presence of fairness concerns

To study the issue of time and social dilemmas in the presence of fairness concerns, we consider a Bayesian model of a prisoner's dilemma with endogenous moves. In particular situations, a prisoner's dilemma may be modeled in rather restricted ways.

9.2.1 Complete information

We studied the leadership in a prisoner's dilemma under incomplete information about player's preferences. However, there are also some cases that are appropriate to be modeled as complete information games. We briefly discuss leadership in a prisoner's dilemma under complete information.

One can show that, under complete information of preferences, (C, C) is realized by leadership if at least one of the players is a high-guilt type. Namely, if a player is a high-guilt type and the other player is a low-guilt type, then the low-guilt type takes the leadership and the high-guilt type becomes a follower in equilibrium. If both players are high-guilt types, then there are two leadership equilibria and each player becomes a leader in one of the equilibria.

The mechanism of leadership under complete information of preferences is much simpler than the mechanism of this paper. A player becomes a leader because he knows that his opponent is a high-guilt type who becomes a conditional cooperator. Neither the envy parameter nor the guilt parameter of the leader plays a role in making him a leader. A follower becomes a follower because he knows that his opponent takes the leadership behavior knowing that the follower becomes a conditional cooperator. The envy parameter of the follower plays no role in making him a follower.

In contrast, in the leadership mechanism of this paper, both the envy parameter and the guilt parameter play essential roles in making a player a leader and making a player a follower. This enables us to explain not only how the leadership emerges endogenously, but also why a particular player becomes a leader when a player's preferences are his private information.

9.2.2 Exogenous sequence of moves

We studied the possibility of cooperation in a prisoner's dilemma under the freedom to choose the timing of moves. We argue that when the players have no freedom to choose the timing of moves and a sequence of moves is exogenously given, cooperation is also possible depending on the parameters of the game, but how social preferences make the cooperation realized differs.

Similar to the PD games, we consider a SPD game defined as follows. There are two roles: leader and follower. Before playing the game, the type of each player is realized by f independently. Then, under incomplete information about their utility functions, the players play the prisoner's dilemma in Table 3 in the following sequence.

At timing 1, the leader must choose either C or D. The follower has no move at timing 1 and observes the choice of the leader. Then, the follower chooses C or D at timing 2. This is the end of play. The players receive the payoffs in Table 3 corresponding to the pair of their choices.

The leader has one information set at timing 1. He must choose C or D. The follower has two information sets corresponding to the leader's choice at timing 1 being either C or D. His (pure) strategy is a complete plan that assigns either C or D to each of these information sets. Call a strategy CD when it prescribes C if the leader chooses C and D if the leader chooses D. Call a strategy DD when it prescribes D always. The Bayesian strategy is defined in the same manner as for the PD game.

The sequential equilibrium in a SPD game is characterized as follows. The sequential equilibrium strategy of the follower is to follow CD if his type (α, β) is $\beta > \beta^*$ and DD if it is $\beta < \beta^*$. The follower's strategy generates a distribution of follower's behaviors. If the leader chooses C, then the follower chooses C with a probability $\phi(\boldsymbol{\beta} > \beta^*)$ and D with a probability $1 - \phi(\boldsymbol{\beta} > \beta^*)$. Therefore, the expected utility for the leader from choosing C is $\phi(\boldsymbol{\beta} > \beta^*)a + (1 - \phi(\boldsymbol{\beta} > \beta^*))[c - \alpha(b - c)]$. On the other hand, if the leader chooses D, then the follower responds with D for sure. The expected utility is d. Therefore, the leader chooses C if $\phi(\boldsymbol{\beta} > \beta^*)a + (1 - \phi(\boldsymbol{\beta} > \beta^*))[c - \alpha(b - c)] > d$. The leader chooses D if the opposite inequality holds. The leader is indifferent between C and D if and only if $\phi(\boldsymbol{\beta} > \beta^*)a + (1 - \phi(\boldsymbol{\beta} > \beta^*))[c - \alpha(b - c)] = d$. Solve this equation for α and we obtain

$$\alpha^{**} \equiv \frac{1}{(1 - \phi(\boldsymbol{\beta} > \beta^*))(b - c)} \left[\bar{d}(a) - d\right]$$

where $\bar{d}(a) \equiv \phi(\boldsymbol{\beta} > \beta^*)a + (1 - \phi(\boldsymbol{\beta} > \beta^*))c$ is a bound defined by (15) in Corollary 1. The leader chooses C if $\alpha < \alpha^{**}$ and D if $\alpha > \alpha^{**}$.

The SPD game differs from the PD game in how social preferences make the leadership realized. Two differences are worth mentioning. First, whether a leader chooses C or not is solely determined by his envy parameter α. In the SPD game, the opponent of the leader is necessarily a follower. This means that the leader has no opportunity to betray the opponent. Therefore, in terms of the incentive to lead, which we explored in Section 7, the only tradeoff that the leader faces from choosing C is between the benefit of inducing C from the follower and the risk of being betrayed by the opponent. Hence, the leadership pattern is independent of the guilt parameter β in the SPD game.

Second, the leadership is realized with a positive probability in a SPD game if and only if $\alpha^{**} > 0$; that is, $c < d < \bar{d}(a)$. In the SPD game, the leader is the only player who possibly chooses C at timing 1. This means that the necessary and sufficient condition for the leadership to be realized with a positive probability is that the leader of the pure materialist ($\alpha = \beta = 0$) is willing to take the leadership when the follower follows CD if his type (α, β) is $\beta > \beta^*$ and DD if it is $\beta < \beta^*$. This condition corresponds to the condition for the PD game that the pure materialist has an incentive to take a leadership behavior if no one else takes the leadership (so that the opponent is a follower for sure), all the types with $\beta > \beta^*$ follow CDD, and all the others follow DDD. Hence, the condition $c < d < \bar{d}(a)$ is the necessary and sufficient condition for leadership in the SPD game, while it is only a sufficient condition in the PD game.

These differences raise the issue of social dilemmas and social roles in dynamic decision making in the presence of fairness concerns. The leader role and the follower role are predetermined and assigned to agents by the rule in the SPD game. In

contrast, agents are only provided an alternative of taking the role of leader or not in the PD game. The leader role and the follower role are realized by the agents' voluntary will if and only if one of the agents takes the leader role and the other agent chooses to become a follower. The above differences in the way of leadership in the SPD game and the PD game show that how social roles are distributed matters for a possibility of resolving social dilemmas. In particular, one may tend to think that it is more difficult to support leadership in equilibrium in the PD game, in which an individual has the freedom not to take the responsibility of leadership, than in the SPD game, in which the leader agent cannot escape the role of leader. Contrary to this first thought, the second comparison suggests that an agent takes the leadership behavior with a positive probability in a wider class of prisoner's dilemmas in the PD game than in the SPD game. Note, however, that when we consider a prisoner's dilemma that supports a leadership equilibrium both in the SPD game and the PD game, it is not clear whether the probability of leadership is higher in the SPD game or in the PD game, because the probability of leadership in the PD game is determined endogenously in equilibrium. This raises an interesting issue for future research.

9.2.3 Simultaneous moves

We studied the possibility of cooperation in a prisoner's dilemma by the leadership in a model in which the players choose C or D over a span of time. We argue that when the players have no freedom to choose the timing of moves and they are forced by the rules of the game to move simultaneously, cooperation is also possible depending on the parameters of the game.

Consider a simultaneous-move prisoner's dilemma that is the same game as PD except that there is a single timing for moves at which each player must choose C or D simultaneously and independently. A (pure) Bayesian strategy $\boldsymbol{s}$ assigns to each type $(\alpha, \beta) \in T$ a choice $\boldsymbol{s}(\alpha, \beta) \in \{C, D\}$. We describe the (symmetric pure) Bayesian Nash equilibrium condition in terms of payoff parameters a, b, c, d and a type density f.[45]

Let $\mu_C = \phi(\boldsymbol{s}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = C)$ denote the consistent belief that a player chooses C. Then, a player of type (α, β) prefers C to D if and only if $\mu_C a + (1-\mu_C)\{c - \alpha(b-c)\} > \mu_C\{b - \beta(b-c)\} + (1-\mu_C)d$, or equivalently:

$$\beta > \frac{1-\mu_C}{\mu_C}\alpha + \frac{1-\mu_C}{\mu_C}\frac{d-c}{b-c} + \frac{b-a}{b-c} =: \widetilde{H}(\alpha|\mu_C). \tag{25}$$

From this, we can derive a fixed-point characterization of symmetric equilibria. The belief μ_C supports an equilibrium if and only if:

$$\mu_C = \phi\left(\boldsymbol{\beta} > \widetilde{H}(\boldsymbol{\alpha}|\mu_C)\right). \tag{26}$$

From the linear relation between α and β in (25), it is straightforward to verify that the right-hand side of (26) equals zero for all $\mu_C \leq \frac{b-c+d-c}{a-c+b-c+d-c} =: \underline{\mu}_C$ and monotonically reaches $\phi(\boldsymbol{\beta} > \beta^*)$ as μ_C goes to 1, where $\beta^* = \frac{b-a}{b-c}$ was defined when we introduced the best-response type sets. This implies that (1) $\mu_C = 0$ always supports an equilibrium in which all the players choose D irrespective of their types, (2) if a positive μ_C supports an equilibrium, then $\mu_C > \underline{\mu}_C$, and (3) if there exists a positive μ_C that supports an equilibrium, then there generically exists at least one

[45] See also Duffy and Muñoz-García (2011).

other positive belief μ'_C that supports another equilibrium. Figures 13 and 14 below illustrate these results.

Figure 13 displays the best-response type set given by equation (25). The blue area describes the set of types who optimally chooses C under the belief μ_C. Combined with ϕ, a belief μ_C induces a probability that a player chooses C given μ_C; that is, $\phi\big(\boldsymbol{\beta} > \widetilde{H}(\boldsymbol{\alpha}|\mu_C)\big)$. Figure 14 shows two kinds of graph of $\phi\big(\boldsymbol{\beta} > \widetilde{H}(\boldsymbol{\alpha}|\mu_C)\big)$ and hence illustrates the fixed-point characterizations of (26). The left graph displays the case in which no positive μ_C supports an equilibrium, while the right graph illustrates that we have two positive beliefs as fixed points. Therefore, cooperation is possible in a simultaneous-move prisoner's dilemma depending on the parameters (a, b, c, d) and f.

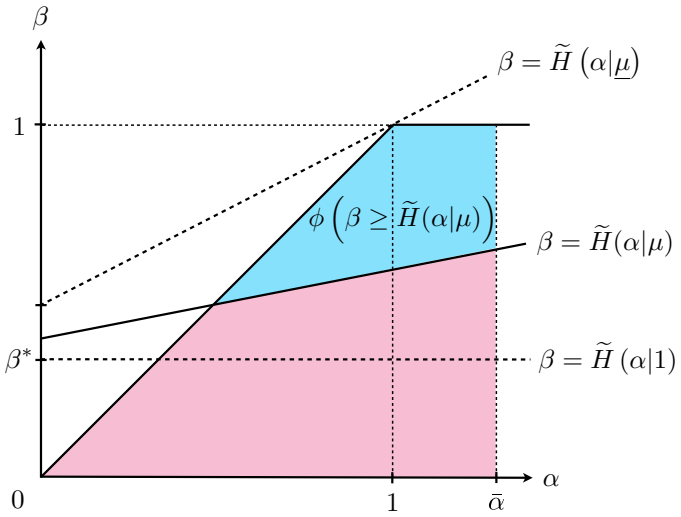


Figure 13: The area of $\phi\big(\boldsymbol{\beta} > \widetilde{H}(\boldsymbol{\alpha}|\mu_C)\big)$

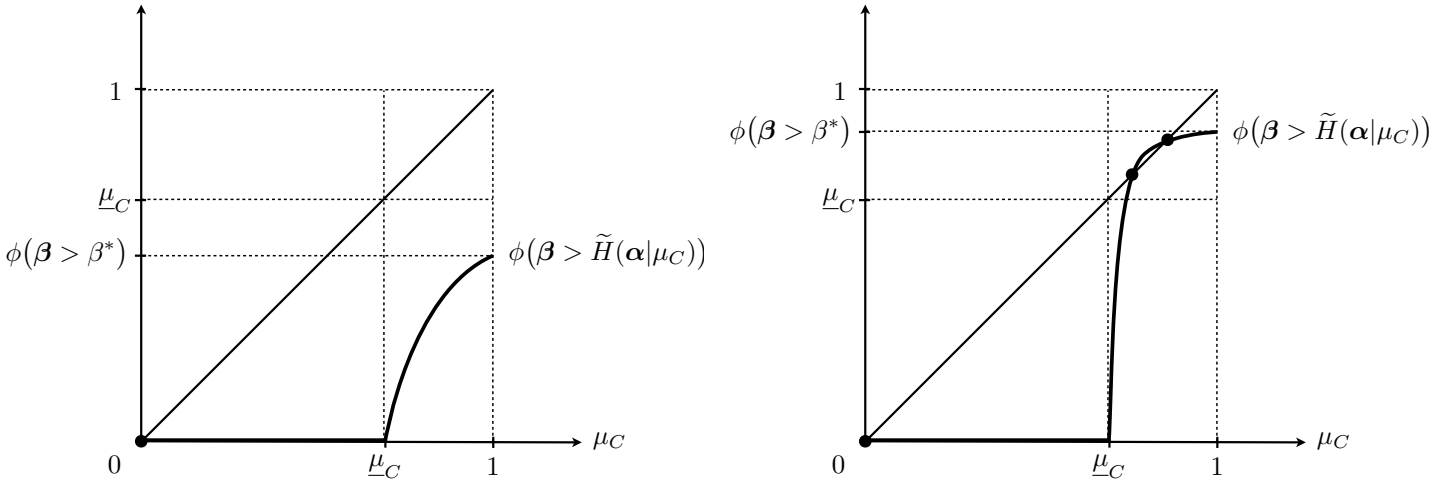


Figure 14: The graphs of $\phi\big(\boldsymbol{\beta} > \widetilde{H}(\boldsymbol{\alpha}|\mu_C)\big)$

Furthermore, we can discuss how cooperation by the leadership differs from cooperation in simultaneous-move games. In a simultaneous-move prisoner's dilemma, if $\phi(\boldsymbol{\beta} > \beta^*) \leq \underline{\mu}_C$, we never have an equilibrium in which a player cooperates, which is

illustrated by the left graph of Figure 14. On the other hand, Theorem 1 states that if $\phi(\boldsymbol{\beta} > \beta^*) > \frac{d-c}{a-c}$, there exists a three-mode equilibrium in a (endogenous timing) PD game. These results imply that if the game parameters satisfy $\frac{d-c}{a-c} < \phi(\boldsymbol{\beta} > \beta^*) \leq \underline{\mu}_C$, then the cooperation outcome (C, C) is realized as an equilibrium outcome with a positive probability in the (endogenous timing) PD game, whereas we never observe cooperation in the simultaneous-move prisoner's dilemma with the same parameters.[46] This provides a case in which a prisoner's dilemma with the chance of voluntary moves is superior in achieving cooperation to a prisoner's dilemma without the chance of voluntary moves. This case may suggest that working face-to-face is different from working in isolation in light of advancing cooperation among the players. We leave a comprehensive understanding of this issue for future research.

Appendix A: The case of an exact bound for the existence of the three-mode equilibrium

We can elaborate Theorem 2 and show that under certain conditions on f, the sufficiency bound $\bar{d}(a)$ for the existence of a three-mode equilibrium in Corollary 1 is in fact the exact bound. We say that a prior f of types admits the exact bound for the existence of a three-mode equilibrium when for any pair (a, d) with $c < d < a < b$ there exists a three-mode equilibrium in $PD((a, b, c, d), f)$ if and only if $c < d < \bar{d}(a)$. When f admits the exact bound, the threshold $\hat{d}(a)$ for the existence of a three-mode equilibrium stated in Theorem 2 is identical to the sufficiency bound of Corollary 1 itself; that is, $\hat{d}(a) = \bar{d}(a)$ for any $a \in (c, b)$.[47]

We characterize a prior f of types that admits the exact bound for the existence of a three-mode equilibrium. Let $\phi(\boldsymbol{\beta} < \beta^* | \alpha^* \leq \boldsymbol{\alpha})$ denote the conditional probability of $\boldsymbol{\beta} < \beta^*$ given an event of $\alpha^* \leq \boldsymbol{\alpha}$. We set

$$\phi(\boldsymbol{\beta} < \beta^* | \bar{\alpha} \leq \boldsymbol{\alpha}) = \lim_{\alpha^* \to \bar{\alpha}} \phi(\boldsymbol{\beta} < \beta^* | \alpha^* \leq \boldsymbol{\alpha})$$

for the case of $\alpha^* = \bar{\alpha}$. Then, we can show that a prior f admits the exact bound for the existence of a three-mode equilibrium if it satisfies the following condition.

Theorem 13. *Fix f. Suppose that*

$$\frac{1 - \tilde{\beta}}{1 - \tilde{\beta} + \tilde{\alpha}} < \frac{\phi(\boldsymbol{\beta} < \tilde{\beta} | \tilde{\alpha} \leq \boldsymbol{\alpha})}{\phi(\boldsymbol{\beta} < \tilde{\beta})} \tag{27}$$

holds for any $\tilde{\beta} \in (0, 1)$ and $\tilde{\alpha} \in (0, \bar{\alpha})$. Then, there exists a three-mode equilibrium in $PD((a, b, c, d), f)$ if and only if $c < d < \bar{d}(a)$.

Theorem 13 says that if the conditional probability of $\boldsymbol{\beta} < \tilde{\beta}$ given an event of $\tilde{\alpha} \leq \boldsymbol{\alpha}$ for $\tilde{\alpha} \in (0, \bar{\alpha})$ is high enough in comparison with the corresponding probability in the extreme case of $\tilde{\alpha} = 0$, that is the prior probability $\phi(\boldsymbol{\beta} < \tilde{\beta})$, there is no three-mode equilibrium in $PD((a, b, c, d), f)$ with $d \geq \bar{d}(a)$. The condition (27) provides a particular bound $\frac{1-\tilde{\beta}}{1-\tilde{\beta}+\tilde{\alpha}}$ for a ratio of $\phi(\boldsymbol{\beta} < \tilde{\beta} | \tilde{\alpha} \leq \boldsymbol{\alpha})$ against $\phi(\boldsymbol{\beta} < \tilde{\beta})$. An example

[46] To state it by treating the payoff parameters and type distribution separately, if $\frac{d-c}{a-c} < \underline{\mu}_C$, or equivalently $(b-c)(a-d) > (d-c)^2$, then there exists a type density f such that cooperation outcome (C, C) is realized as an equilibrium outcome with a positive probability in the (endogenous timing) PD game with those parameters and is never realized in the simultaneous-move prisoner's dilemma.

[47] To be precise, when f admits the exact bound, there exists no three-mode equilibrium in $PD((a, b, c, d), f)$ for an interior case of $\hat{d}(a) < d < a$ but also for a boundary case of $d = \hat{d}(a)$.

of a prior that admits the exact bound for the existence of a three-mode equilibrium is the uniform distribution over $T = \{(\alpha, \beta) | 0 \leq \alpha, \beta \leq 1 \text{ and } \beta \leq \alpha\}$.

The bound in the condition (27) has the following simple meaning. Set $\tilde{\beta} = \beta^* = \frac{b-a}{b-c}$ for $PD((a,b,c,d), f)$. Then, for each $\tilde{\alpha}$ given, consider a "C if $\alpha < \tilde{\alpha}$" belief $\tilde{\mu}$ such that a type (α, β) with $\alpha < \tilde{\alpha}$ follows C, a type (α, β) with $\alpha \geq \tilde{\alpha}$ and $\beta \geq \beta^*$ follows CDD, and a type (α, β) with $\alpha \geq \tilde{\alpha}$ and $\beta < \beta^*$ follows DDD. Then, a threshold type $(\tilde{\alpha}, \beta)$'s preference for C over CDD is written as

$$
\begin{aligned}
&U_{(\tilde{\alpha},\beta)}(C, \tilde{\mu}) - U_{(\tilde{\alpha},\beta)}(CDD, \tilde{\mu}) \\
&= \Big(\phi(\tilde{\alpha} > \boldsymbol{\alpha})a + \phi(\tilde{\alpha} \leq \boldsymbol{\alpha})[\phi(\boldsymbol{\beta} \geq \beta^* | \tilde{\alpha} \leq \boldsymbol{\alpha})a + \phi(\boldsymbol{\beta} < \beta^* | \tilde{\alpha} \leq \boldsymbol{\alpha})(c - \tilde{\alpha}(b-c))]\Big) \\
&\quad - \Big(\phi(\tilde{\alpha} > \boldsymbol{\alpha})a + \phi(\tilde{\alpha} \leq \boldsymbol{\alpha})[\phi(\boldsymbol{\beta} \geq \beta^* | \tilde{\alpha} \leq \boldsymbol{\alpha})d + \phi(\boldsymbol{\beta} < \beta^* | \tilde{\alpha} \leq \boldsymbol{\alpha})d]\Big) \\
&= \phi(\tilde{\alpha} \leq \boldsymbol{\alpha})\Big[\Big(a - \phi(\boldsymbol{\beta} < \beta^* | \tilde{\alpha} \leq \boldsymbol{\alpha})(1 - \beta^* + \tilde{\alpha})(b-c)\Big) - d\Big].
\end{aligned}
$$

The bound in condition (27) means that this preference (the value in the brackets) is maximized at the extreme case of $\tilde{\alpha} = 0$. In words, the pure materialist ($\alpha = \beta = 0$) is the type who is the most inclined to adopt strategy C over strategy CDD given a "C if $\alpha < \tilde{\alpha}$" belief across all the candidate thresholds $\tilde{\alpha} \in [0, \bar{\alpha})$.

From this meaning of the condition (27), it is now straightforward to see why there is no three-mode equilibrium in $PD((a,b,c,d), f)$ with $d \geq \bar{d}(a)$. Suppose that there is a three-mode equilibrium distribution μ. As we noted after Theorem 1, the pure materialist is not willing to follow C (or at most is indifferent between C and DDD in the case of $d = \bar{d}(a)$) given a "C if $\alpha < \tilde{\alpha}$" belief that corresponds to the extreme case of $\tilde{\alpha} = 0$. Then, the condition (27) means that a threshold type $(\tilde{\alpha}, \beta)$ with $\tilde{\alpha} = \alpha^*(\mu) > 0$ in the presumed equilibrium also strictly prefers CDD to C given a "C if $\alpha < \tilde{\alpha}$" belief $\tilde{\mu}$ that corresponds to $\tilde{\alpha} = \alpha^*(\mu)$. Note that this belief $\tilde{\mu}$ is a belief that gives greater advantage to strategy C than the presumed distribution μ because Lemma 2 indicates that when μ is a three-mode equilibrium distribution, a player is subject to a risk of being betrayed not only by the opponent of type (α, β) with $\alpha \geq \alpha^*(\mu)$ and $\beta \leq \beta^*$, but also by the opponent of some type (α, β) with $\alpha < \alpha^*(\mu)$. Hence, the threshold type $(\tilde{\alpha}, \beta)$ with $\tilde{\alpha} = \alpha^*(\mu)$ must strictly prefer CDD to C given the presumed distribution μ. Hence, μ cannot be a three-mode equilibrium distribution.

When a prior f admits the exact bound for the existence of a three-mode equilibrium, our results developed in this paper can be stated in a much sharper way. As for the results on the existence of a three-mode equilibrium in Theorems 1 and 2, we can use the explicit formula (15) in Corollary 1 to test whether there exists a three-mode equilibrium in a prisoner's dilemma $PD((a,b,c,d), f)$.

The comparative statics results in Theorems 3 through 10 can be stated as the comparative statics with respect to the three-mode equilibrium in general by referring to the case of $c < d < \bar{d}(a)$ without the necessity of referring to the case of $\bar{d}(a) < d < \hat{d}(a)$. Then, the comparative statics results with respect to the minimum leadership equilibrium $\mu^{\min}$ are much simpler and parallel to the comparative statics results with respect to the maximum leadership equilibrium $\mu^{\max}$. Furthermore, the characterization of the materialist leader pattern in Theorem 9 can be strengthened. First, we can say that the leadership probability μ_C shrinks to 0 as d approaches the bound $\bar{d}(a)$.

Lemma 10. *Suppose that a prior f admits the exact bound for the existence of a three-mode equilibrium. Then, for any given $m > 0$, there exists $d_m(a)$ with $c \leq$*

$d_m(a) < \bar{d}(a)$ *such that* $0 < \mu_C < m$ *for any three-mode equilibrium distribution* $\mu = (\mu_C, \mu_{CDD}, \mu_{DDD})$ *in any* $PD((a,b,c,d),f)$ *with* $d_m(a) < d < \bar{d}(a)$.

Then, we can say that any three-mode equilibrium is a materialist leader pattern in any prisoner's dilemma $PD((a,b,c,d),f)$ close enough to the bound $\bar{d}(a)$.

Theorem 14. *Suppose that a prior* f *admits the exact bound for the existence of a three-mode equilibrium. Then, there exists* $\hat{d}_{ml}(a)$ *with* $c \leq \hat{d}_{ml}(a) < \bar{d}(a)$ *such that any three-mode equilibrium in any prisoner's dilemma* $PD((a,b,c,d),f)$ *with* $\hat{d}_{ml}(a) < d < \bar{d}(a)$ *is of a materialist leader pattern while there exists a three-mode equilibrium of either a hybrid leader pattern or a inequity concerned leader pattern in any prisoner's dilemma* $PD((a,b,c,d),f)$ *with* $c < d < \hat{d}_{ml}(a)$. *Furthermore,* $\hat{d}_{ml}(a)$ *is increasing,* $\hat{d}_{ml}(a) \leq \bar{d}_{ml}(a)$, *and* $\lim_{a \to b} \hat{d}_{ml}(a) = b$ *where* $\bar{d}_{ml}(a)$ *is the bound stated in Theorem 9.*

As we noted after Theorem 8, Theorems 7 and 8 suggest that if a prisoner's dilemma $PD((a,b,c,d),f)$ supports a three-mode equilibrium with a materialist leader pattern, then the leadership pattern turns from the materialist leader pattern to a hybrid pattern (and then possibly to an inequity concerned leader pattern) as we move from the prisoner's dilemma $PD((a,b,c,d),f)$ by lowering d and raising a. Figure 11 demonstrates this leadership pattern transition for a particular prior $f(\alpha,\beta) = 6\beta$. Theorem 14 means that when a prior f admits the exact bound for the existence of a three-mode equilibrium, there exists in fact a nondegenerate area along the boundary $d = \bar{d}(a)$ in which any three-mode equilibrium is a materialist leader pattern so that the leadership pattern transition from a materialist leader pattern to a hybrid pattern (and then possibly to an inequity concerned leader pattern) occurs as we start from any prisoner's dilemma close enough to the boundary $d = \bar{d}(a)$ and move by lowering d and raising a.

Appendix B: An example of no inequity concerned leader pattern

Consider the uniform distribution over a type space $T = \{(\alpha,\beta) | \alpha \geq \beta, 0 \leq \alpha, \beta \leq 1\}$. Then, for this prior, there exists no three-mode equilibrium with an inequity concerned leader pattern in any prisoner's dilemma. Suppose to the contrary that there exists a three-mode equilibrium distribution μ with an inequity concerned leader pattern in some prisoner's dilemma. Let γ denote the slope of the boundary of the set $T^*_C(\mu)$ for an area of $\beta < \beta^*$ in the type space. Then, the three-mode equilibrium requires that $\gamma = \frac{\mu_{DDD}}{\mu_C}$, and the inequity concerned leader pattern requires that $\gamma \leq \frac{\beta^*}{\alpha^*(\mu)}$. Now, consider a triangle T_C that connects $(0,0)$, $(\alpha^*(\mu),\beta^*)$, and $(\alpha^*(\mu),\alpha^*(\mu))$ in the type space T. Then, a probability that a type is realized in T_C is $(\alpha^*(\mu) - \beta^*)\alpha^*(\mu)$ under the uniform distribution. This probability is no less than μ_C because the set $T^*_C(\mu)$ is a subset of T_C in the inequity concerned leader pattern. Therefore, $(\alpha^*(\mu) - \beta^*)\alpha^*(\mu) \geq \mu_C$. Similarly, consider a trapezoid T_{DDD} that connects $(0,0)$, $(1,0)$, $(1,\beta^*)$, and $(\alpha^*(\mu),\beta^*)$. Then, a probability that a type is realized in T_{DDD} is $((1-\alpha^*(\mu))+1)\beta^*$ under the uniform distribution. This probability is no more than μ_{DDD} because T_{DDD} is a subset of the set $T^*_{DDD}(\mu)$ in the inequity concerned leader pattern. Therefore, $((1-\alpha^*(\mu))+1)\beta^* \leq \mu_{DDD}$. Hence, we must have

$$\frac{\beta^*}{\alpha^*(\mu)} \geq \gamma = \frac{\mu_{DDD}}{\mu_C} \geq \frac{((1-\alpha^*(\mu))+1)\beta^*}{(\alpha^*(\mu)-\beta^*)\alpha^*(\mu)}.$$

This holds only when $\alpha^*(\mu) = 1$ and $\beta^* = 0$. This is a contradiction.

Appendix C: Proofs

Proof of Lemma 1

[**Step 1**] Verify that $\psi(\mu) \in \Delta$ for every $\mu \in \Delta \setminus (1,0,0)$. This follows from Lemma 2 and Lemma 3. These lemmas show that $T^*_C(\mu)$, $T^*_{CDD}(\mu)$, $T^*_{DDD}(\mu)$ cover T and have degenerate intersections to each other so that $\psi_C(\mu) + \psi_{CDD}(\mu) + \psi_{DDD}(\mu) = \phi(T^*_C(\mu)) + \phi(T^*_{CDD}(\mu)) + \phi(T^*_{DDD}(\mu)) = \phi(T^*_C(\mu) \cup T^*_{CDD}(\mu) \cup T^*_{DDD}(\mu)) = \phi(T) = 1$.

[**Step 2**] We confirm that $\psi(\mu)$ is continuous.[48] First, the sets $T^*_C(\mu)$, $T^*_{CDD}(\mu)$, and $T^*_{CDD}(\mu)$ are continuous in μ in the Housdorff metric. Second, the sets $T^*_C(\mu)$, $T^*_{CDD}(\mu)$, and $T^*_{CDD}(\mu)$ are closed subsets in Δ and the function ϕ is continuous in Borel sets in Δ.

[**Step 3**] We show Lemma 1-(1). Suppose that $\boldsymbol{s} : T \to S$ is a three-mode equilibrium strategy. Then, $\mu = (\phi(T_C(\boldsymbol{s})), \phi(T_{CDD}(\boldsymbol{s})), \phi(T_{DDD}(\boldsymbol{s})))$ is a three-mode distribution. Furthermore, μ is a consistent belief at timing 1. Take $(\alpha, \beta) \in T_C(\boldsymbol{s})$. Then, it is necessary for the sequential rationality of this type at timing 1 that $U_{(\alpha,\beta)}(C, \mu) \geq U_{(\alpha,\beta)}(CDD, \mu), U_{(\alpha,\beta)}(DDD, \mu)$. This means that $T_C(\boldsymbol{s}) \subseteq T^*_C(\mu)$. Take $(\alpha, \beta) \in T_{CDD}(\boldsymbol{s})$. Then, it is necessary for the sequential rationality of this type at timing 1 that $U_{(\alpha,\beta)}(CDD, \mu) \geq U_{(\alpha,\beta)}(C, \mu), U_{(\alpha,\beta)}(DDD, \mu)$. As is explained in the text, $\beta \geq \beta^*$ is also necessary for the sequential rationality of this type at timing 2 after the opponent chooses C. This means that $T_{CDD}(\boldsymbol{s}) \subseteq T^*_{CDD}(\mu)$. Similarly, $T_{DDD}(\boldsymbol{s}) \subseteq T^*_{DDD}(\mu)$. Then, $\phi(T_C(\boldsymbol{s})) = \phi(T^*_C(\mu))$, $\phi(T_{CDD}(\boldsymbol{s})) = \phi(T^*_C(\mu))$, and $\phi(T_{DDD}(\boldsymbol{s})) = \phi(T^*_{DDD}(\mu))$ must hold because $T^*_C(\mu)$, $T^*_{CDD}(\mu)$, $T^*_{DDD}(\mu)$ have degenerate intersections to each other. This means $\mu = \psi(\mu)$.

[**Step 4**] We show Lemma 1-(2). Suppose a three-mode distribution $\mu \in \Delta$ such that $\mu = \psi(\mu)$. Take any three-mode strategy $\boldsymbol{s} : T \to S$ that satisfies $T_C(\boldsymbol{s}) \subseteq T^*_C(\mu)$, $T_{CDD}(\boldsymbol{s}) \subseteq T^*_{CDD}(\mu)$, and $T_{DDD}(\boldsymbol{s}) \subseteq T^*_{DDD}(\mu)$. Then, $\phi(T_C(\boldsymbol{s})) = \phi(T^*_C(\mu))$, $\phi(T_{CDD}(\boldsymbol{s})) = \phi(T^*_C(\mu))$, and $\phi(T_{DDD}(\boldsymbol{s})) = \phi(T^*_{DDD}(\mu))$ because $T^*_C(\mu)$, $T^*_{CDD}(\mu)$, $T^*_{DDD}(\mu)$ have degenerate intersections to each other. This means that μ is a belief at timing 1 that is consistent with $\boldsymbol{s}$ because $\mu = \psi(\mu)$.

Now, we verify that $\boldsymbol{s}$ satisfies sequential rationality under μ. Consider the information set at timing 2 after the opponent chooses C. It is sequentially rational for a type $(\alpha, \beta) \in T_{CDD}(\boldsymbol{s})$ to choose C because $T_{CDD}(\boldsymbol{s}) \subseteq T^*_{CDD}(\mu)$, and as is explained in the text, it is sequentially rational for a type $(\alpha, \beta) \in T^*_{CDD}(\mu)$ to choose C. Similarly, it is sequentially rational for a type $(\alpha, \beta) \in T_{DDD}(\boldsymbol{s})$ to choose D.[49]

Consider the information set at timing 2 after the opponent chooses D. Then, it is sequentially rational for both a type $(\alpha, \beta) \in T_{CDD}(\boldsymbol{s})$ and a type $(\alpha, \beta) \in T_{DDD}(\boldsymbol{s})$ to choose D because the best response to D is D irrespective of a player's inequity aversion.

Consider the information set at timing 2 after the opponent chooses $\emptyset$. Then, the consistent belief is that the opponent is either a type $(\alpha, \beta) \in T_{CDD}(\boldsymbol{s})$ or a type $(\alpha, \beta) \in T_{DDD}(\boldsymbol{s})$. Both types choose $a_\emptyset = D$ at timing 2 after the player in question also chooses $\emptyset$ at timing 1. Hence, it is sequentially rational for the player to choose D against the expected choice of D irrespective of whether he is a type $(\alpha, \beta) \in T_{CDD}(\boldsymbol{s})$ or a type $(\alpha, \beta) \in T_{DDD}(\boldsymbol{s})$.

Finally, consider the information set at timing 1. First, note that the optimality of

[48] Note that $\psi(\mu)$ also depends on (a, d). The function ψ ia also continuous in (a, d) by the same argument.

[49] C-mode is left unspecified in timing 2 actions $a_C, a_D, a_\emptyset$. To be explicit, a type $(\alpha, \beta) \in T_C(\boldsymbol{s})$ follows (C, C, D, D) if $\beta > \beta^*$ and (C, D, D, D) if $\beta < \beta^*$. Then, it is sequentially rational for a type $(\alpha, \beta) \in T_C(\boldsymbol{s})$ to take these actions.

a strategy among those with $a_1 = \emptyset$ is equivalent to the sequential rationality at timing 2. As is established above, either CDD or DDD is optimal against the three-mode strategy $\boldsymbol{s} : T \to S$. Second, the strategies with $a_1 \neq \emptyset$ are C and $D = (D, a_C, a_D, a_\emptyset)$. Observe that DDD and D generate the same outcomes given the three-mode strategy $\boldsymbol{s} : T \to S$. If the opponent follows C-mode, then both the outcome from following DDD and the outcome from following D are (D, C). If the opponent follows either CDD-mode or DDD-mode, then both the outcome from following DDD and the outcome from following D are (D, D). Therefore, DDD is as good as D given the three-mode strategy $\boldsymbol{s} : T \to S$. Hence, if a strategy is optimal among C, CDD, and DDD, then it is optimal in S. This means that it is sequentially rational for a type $(\alpha, \beta) \in T_C(\boldsymbol{s})$ to choose C at timing 1 because $T_C(\boldsymbol{s}) \subseteq T^*_C(\mu)$, and it is sequentially rational for a type $(\alpha, \beta) \in T_{CDD}(\boldsymbol{s})$ and a type $(\alpha, \beta) \in T_{DDD}(\boldsymbol{s})$ to choose $\emptyset$ because $T_{CDD}(\boldsymbol{s}) \subseteq T^*_{CDD}(\mu)$ and $T_{DDD}(\boldsymbol{s}) \subseteq T^*_{DDD}(\mu)$. (Q.E.D.)

Proof of Lemma 4

[**Step 1**] Consider $\psi(\mu)$ for $\mu = (\mu_C, 1 - \mu_C, 0)$ with $\mu_C \in [0, 1)$. Then, there exists $\bar{\mu}_C \in (\frac{a-d}{b-d}, 1)$ such that $\psi_C(\mu_C, 1 - \mu_C, 0) > \mu_C$ for $\mu_C \in [0, \bar{\mu}_C)$, $\psi_C(\bar{\mu}_C, 1 - \bar{\mu}_C, 0) = \bar{\mu}_C$, and $\psi_C(\mu_C, 1 - \mu_C, 0) < \mu_C$ for $\mu_C \in (\bar{\mu}_C, 1)$.

(Proof)

Consider $\mu = (\mu_C, 1 - \mu_C, 0)$ with $\mu_C \in [0, 1)$. Then, $T^*_{CDD}(\mu) = \emptyset$ because

$$U_{(\alpha,\beta)}(C, \mu) - U_{(\alpha,\beta)}(CDD, \mu) = (1 - \mu_C)(a - d) > 0$$

for any $(\alpha, \beta) \in T$. Therefore, $(\alpha, \beta) \in T^*_C(\mu)$ if and only if

$$U_{(\alpha,\beta)}(C, \mu) - U_{(\alpha,\beta)}(DDD, \mu) = \mu_C\{a - [b - \beta(b - c)]\} + (1 - \mu_C)(a - d) \geq 0.$$

Hence, we have

$$\psi_C(\mu) = \phi\Big(\mu_C(b - c)\boldsymbol{\beta} \geq \mu_C(b - d) - (a - d)\Big).$$

This means that $\psi_C(\mu) = 1$ for $\mu_C \in [0, \frac{a-d}{b-d}]$ and $\psi_C(\mu)$ is strictly decreasing in μ_C over $\mu_C \in (\frac{a-d}{b-d}, 1)$. Furthermore, $\psi : \Delta \setminus (1, 0, 0) \to \Delta$ is continuous in μ by Lemma 1. Hence, we have the desired $\bar{\mu}_C$. ||

[**Step 2**] Consider $\psi(\mu)$ for $\mu = (\mu_C, \mu_{CDD}, \mu_{DDD})$ with $\mu_C \in [0, 1)$ and $\mu_{CDD} = \frac{d-c}{a-d}\mu_{DDD}$. Then, the threshold $\alpha^*(\mu)$ defined by (4) is $\alpha^*(\mu) = 0$. Hence, it follows from Lemma 2 and Lemma 3 that $\psi(\mu) = (0, \phi(\boldsymbol{\beta} > \beta^*), \phi(\boldsymbol{\beta} < \beta^*))$.

[**Step 3**] Fix $\mu_C \in [0, 1)$ and consider $\psi(\mu'), \psi(\mu'')$ for $\mu' = (\mu'_C, \mu'_{CDD}, \mu'_{DDD}), \mu'' = (\mu''_C, \mu''_{CDD}, \mu''_{DDD})$ with $\mu'_C = \mu''_C = \mu_C$, $\mu'_{CDD} \geq \frac{d-c}{a-d}\mu'_{DDD}$, and $\mu''_{CDD} \geq \frac{d-c}{a-d}\mu''_{DDD}$. Then, if $\mu'_{CDD} > \mu''_{CDD}$ (and so $\mu'_{DDD} < \mu''_{DDD}$), then

(1) $\psi_C(\mu') \geq \psi_C(\mu'')$, and

(2) it is never the case that $0 < \psi_C(\mu') = \psi_C(\mu'') < 1$.

(Proof)

Consider $\mu' = (\mu'_C, \mu'_{CDD}, \mu'_{DDD}), \mu'' = (\mu''_C, \mu''_{CDD}, \mu''_{DDD})$ with $\mu'_C = \mu''_C = \mu_C \in (0, 1)$ and $\mu'_{CDD} > \mu''_{CDD}$ (and so $\mu'_{DDD} < \mu''_{DDD}$). We show that $T^*_C(\mu') \supseteq T^*_C(\mu'')$. Take $(\alpha, \beta) \in T^*_C(\mu'')$. Then, $U_{(\alpha,\beta)}(C, \mu'') \geq U_{(\alpha,\beta)}(CDD, \mu'')$ and $U_{(\alpha,\beta)}(C, \mu'') \geq$

$U_{(\alpha,\beta)}(DDD, \mu'')$. Then, together with $\mu'_{CDD} > \mu''_{CDD}$ and $\mu'_{DDD} < \mu''_{DDD}$, we have

$$\begin{aligned}
U_{(\alpha,\beta)}(C, \mu') - U_{(\alpha,\beta)}(CDD, \mu') &= \mu_C(a-a) + \mu'_{CDD}(a-d) + \mu'_{DDD}\{[c-\alpha(b-c)]-d\} \\
&> \mu_C(a-a) + \mu''_{CDD}(a-d) + \mu''_{DDD}\{[c-\alpha(b-c)]-d\} \\
&= U_{(\alpha,\beta)}(C, \mu'') - U_{(\alpha,\beta)}(CDD, \mu'') \\
&\geq 0 \\
U_{(\alpha,\beta)}(C, \mu') - U_{(\alpha,\beta)}(DDD, \mu') &= \mu_C\{a-[b-\beta(b-c)]\} + \mu'_{CDD}(a-d) + \mu'_{DDD}\{[c-\alpha(b-c)]-d\} \\
&> \mu_C\{a-[b-\beta(b-c)]\} + \mu''_{CDD}(a-d) + \mu''_{DDD}\{[c-\alpha(b-c)]-d\} \\
&= U_{(\alpha,\beta)}(C, \mu'') - U_{(\alpha,\beta)}(DDD, \mu'') \\
&\geq 0.
\end{aligned}$$

Therefore, $(\alpha, \beta) \in T^*_C(\mu')$. Hence, $T^*_C(\mu') \supseteq T^*_C(\mu'')$. This establishes $\psi_C(\mu') \geq \psi_C(\mu'')$.

Suppose that $0 < \psi_C(\mu'') < 1$. Then, it is easily verified from Lemmas 2 and 3 that there exists $(\alpha, \beta) \in T^*_C(\mu'') \cap T^*_{DDD}(\mu'')$ such that any open neighborhood N of (α, β) contains an open set in $(N \cap T^*_{DDD}(\mu'')) \setminus T^*_C(\mu'')$. On the other hand, the argument in the proof of (1) immediately guarantees the existence of an open neighborhood N' of the (α, β) such that C is the best among C, CDD, DDD under μ' for every $(\tilde{\alpha}, \tilde{\beta}) \in N'$; that is, $N' \subseteq T^*_C(\mu')$. N' contains an open set in $(N' \cap T^*_{DDD}(\mu'')) \setminus T^*_C(\mu'')$. This open set is contained in $T^*_C(\mu') \setminus T^*_C(\mu'')$, and hence $\psi_C(\mu') > \psi_C(\mu'')$. ||

[**Step 4**] For each $\mu_C \in [0, \bar{\mu}_C]$, there uniquely exists $\hat{\mu} \in \Delta$ with $\hat{\mu}_{CDD} \geq \frac{d-c}{a-d}\hat{\mu}_{DDD}$ such that $\psi_C(\hat{\mu}) = \hat{\mu}_C = \mu_C$. Denote it as $\hat{\mu}(\mu_C)$. Then,

(1) $\hat{\mu}(0) = (0, \hat{\mu}_{CDD}(0), \hat{\mu}_{DDD}(0))$ with $\hat{\mu}_{CDD}(0) = \frac{d-c}{a-d}\hat{\mu}_{DDD}(0)$ and $\psi(\hat{\mu}(0)) = (0, \phi(\boldsymbol{\beta} > \beta^*), \phi(\boldsymbol{\beta} < \beta^*))$

(2) $\hat{\mu}(\bar{\mu}_C) = (\bar{\mu}_C, 1-\bar{\mu}_C, 0)$ and $\psi(\hat{\mu}(\bar{\mu}_C)) = (\bar{\mu}_C, 0, 1-\bar{\mu}_C)$

(3) $\hat{\mu}_{CDD}(\mu_C) > \frac{d-c}{a-d}\hat{\mu}_{DDD}(\mu_C)$, $\hat{\mu}_{DDD}(\mu_C) > 0$ for $\mu_C \in (0, \bar{\mu}_C)$

(Proof)

Consider $\mu_C = 0$. Consider $\mu = (0, \mu_{CDD}, \mu_{DDD})$ with $\mu_{CDD} = \frac{d-c}{a-d}\mu_{DDD}$. This μ satisfies the required property for $\hat{\mu}$ that $\psi_C(\hat{\mu}) = \hat{\mu}_C = \mu_C = 0$ because $\psi(\mu) = (0, \phi(\boldsymbol{\beta} > \beta^*), \phi(\boldsymbol{\beta} < \beta^*))$ holds by Step 2. Consider any $\mu = (0, \mu_{CDD}, \mu_{DDD})$ with $\mu_{CDD} > \frac{d-c}{a-d}\mu_{DDD}$. Then,

$$U_{(0,0)}(C, \mu) - U_{(0,0)}(CDD, \mu) = U_{(0,0)}(C, \mu) - U_{(0,0)}(DDD, \mu) = \mu_{CDD}(a-d) + \mu_{DDD}(c-d) > 0.$$

Hence, $T^*_C(\mu)$ contains some neighborhood of $(0, 0)$. Therefore, $\psi_C(\mu) > 0$. This μ fails to satisfy the required property for $\hat{\mu}$ that $\psi_C(\hat{\mu}) = \hat{\mu}_C = \mu_C = 0$. Thus, (1) is established.

Consider $\mu_C = \bar{\mu}_C$. Consider $\mu = (\bar{\mu}_C, 1-\bar{\mu}_C, 0)$. This μ satisfies the required property for $\hat{\mu}$ that $\psi_C(\hat{\mu}) = \hat{\mu}_C = \mu_C = \bar{\mu}_C$ because $\psi(\mu) = (\bar{\mu}_C, 0, 1-\bar{\mu}_C)$ holds by Step 1. Consider any other $\mu'' = (\bar{\mu}_C, \mu''_{CDD}, \mu''_{DDD})$ with $\mu''_{CDD} \geq \frac{d-c}{a-d}\mu''_{DDD}$. Then, $\mu_{DDD} = 0 < \mu''_{DDD}$. Therefore, by Step 3, it is never the case that $0 < \psi_C(\mu) = \psi_C(\mu'') = \bar{\mu}_C < 1$. This μ'' fails to satisfy the required property for $\hat{\mu}$ that $\psi_C(\hat{\mu}) = \hat{\mu}_C = \mu_C = \bar{\mu}_C$. Thus, (2) is established.

To show (3), consider $\mu_C \in (0, \bar{\mu}_C)$. Note that $(1 - \mu_C - \mu_{DDD}) \geq \frac{d-c}{a-d}\mu_{DDD}$ if and only if $\mu_{DDD} \leq \frac{a-d}{a-c}(1-\mu_C)$. For each $\mu_{DDD} \in [0, \frac{a-d}{a-c}(1-\mu_C)]$, let $\kappa(\mu_{DDD}) = \psi_C((\mu_C, 1-\mu_C-\mu_{DDD}, \mu_{DDD})) - \mu_C$. Then, $\kappa(0) = \psi_C((\mu_C, 1-\mu_C, 0)) - \mu_C > 0$

follows from (1) and $\kappa(\frac{a-d}{a-c}(1-\mu_C)) = -\mu_C < 0$ follows from (2). The continuity of ψ guarantees that there exists $\hat{\mu} \in \Delta$ with $0 < \hat{\mu}_{DDD}$ and $\hat{\mu}_{CDD} > \frac{d-c}{a-d}\hat{\mu}_{DDD}$ such that $\kappa(\hat{\mu}_{DDD}) = 0$. Then, $\hat{\mu} = (\mu_C, 1-\mu_C-\hat{\mu}_{DDD}, \hat{\mu}_{DDD})$ satisfies the required property for $\hat{\mu}$ that $\psi_C(\hat{\mu}) = \hat{\mu}_C = \mu_C$. Consider any other $\mu'' = (\mu_C, \mu''_{CDD}, \mu''_{DDD})$ with $\mu''_{CDD} \geq \frac{d-c}{a-d}\mu''_{DDD}$. Then, either $\hat{\mu}_{CDD} > \mu''_{CDD}$ or $\hat{\mu}_{CDD} < \mu''_{CDD}$. In either case, by Step 3, it is never the case that $0 < \psi_C(\hat{\mu}) = \psi_C(\mu'') = \mu_C < 1$. This μ'' fails to satisfy the required property for $\hat{\mu}$ that $\psi_C(\hat{\mu}) = \hat{\mu}_C = \mu_C$. Thus, (3) is established. ||

[**Step 5**] Note that $\bar{\mu}_C$ established in Step1 depends on (a, d). Taking this fact into account, let us denote $\bar{\mu}_C(a, d)$ explicitly. Then, $\bar{\mu}_C(a, d)$ is continuous.
(Proof)

Suppose that $\bar{\mu}_C(a, d)$ is not continuous at some (a, d). Then, there exists a sequence $\{(a^n, d^n)\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty}(a^n, d^n) = (a, d)$ and $\{\bar{\mu}_C(a^n, d^n)\}_{n=1}^{\infty}$ does not converge to $\bar{\mu}_C(a, d)$. Then, we can take a convergent subsequence $\{\bar{\mu}_C(a^{n_m}, d^{n_m})\}_{m=1}^{\infty}$ such that there exists $\bar{\mu}^*_C = \lim_{m\to\infty} \bar{\mu}_C(a^{n_m}, d^{n_m})$ and $\bar{\mu}^*_C \neq \bar{\mu}_C(a, d)$ because $\{\bar{\mu}_C(a^n, d^n)\}_{n=1}^{\infty}$ is a bounded sequence. By Step 1, $\bar{\mu}_C(a, d)$ is a unique solution $\bar{\mu}_C$ to

$$\bar{\mu}_C - \phi\Big(\bar{\mu}_C(b-c)\boldsymbol{\beta} \geq \bar{\mu}_C(b-d) - (a-d)\Big) = 0. \tag{28}$$

Therefore, $\bar{\mu}^*_C \neq \bar{\mu}_C(a, d)$ means that

$$\bar{\mu}^*_C - \phi\Big(\bar{\mu}^*_C(b-c)\boldsymbol{\beta} \geq \bar{\mu}^*_C(b-d) - (a-d)\Big) \neq 0.$$

Then, there exists $\epsilon > 0$ such that

$$\Big|\bar{\mu}_C - \phi\Big(\bar{\mu}_C(b-c)\boldsymbol{\beta} \geq \bar{\mu}_C(b-\tilde{d}) - (a-\tilde{d})\Big)\Big| > \frac{1}{2}\Big|\bar{\mu}^*_C - \phi\Big(\bar{\mu}^*_C(b-c)\boldsymbol{\beta} \geq \bar{\mu}^*_C(b-d) - (a-d)\Big)\Big| > 0.$$

for any $(\bar{\mu}_C, \tilde{a}, \tilde{d}) \in (\bar{\mu}^*_C - \epsilon, \bar{\mu}^*_C + \epsilon) \times (a-\epsilon, a+\epsilon) \times (d-\epsilon, d+\epsilon)$ because the left hand side of (28) is continuous in (μ_C, a, d). Then, there exists m such that $(\bar{\mu}_C(a^{n_m}, d^{n_m}), a^{n_m}, d^{n_m}) \in (\bar{\mu}^*_C - \epsilon, \bar{\mu}^*_C + \epsilon) \times (a-\epsilon, a+\epsilon) \times (d-\epsilon, d+\epsilon)$ because $\lim_{m\to\infty} \bar{\mu}_C(a^{n_m}, d^{n_m}) = \bar{\mu}^*_C$ and $\lim_{m\to\infty}(a^{n_m}, d^{n_m}) = (a, d)$. Then,

$$\Big|\bar{\mu}_C(a^{n_m}, d^{n_m}) - \phi\Big(\bar{\mu}_C(a^{n_m}, d^{n_m})(b-c)\boldsymbol{\beta} \geq \bar{\mu}_C(a^{n_m}, d^{n_m})(b-d^{n_m}) - (a^{n_m}-d^{n_m})\Big)\Big| > 0.$$

This is a contradiction because $\bar{\mu}_C(a^{n_m}, d^{n_m})$ is a unique solution $\bar{\mu}_C$ to

$$\bar{\mu}_C - \phi\Big(\bar{\mu}_C(b-c)\boldsymbol{\beta} \geq \bar{\mu}_C(b-d^{n_m}) - (a^{n_m}-d^{n_m})\Big) = 0. \qquad ||$$

[**Step 6**] Note that $\hat{\mu}(\mu_C)$ depends on (a, d). Taking this fact into account, let us denote $\hat{\mu}(\mu_C, a, d)$ explicitly. Then, $\hat{\mu}(\mu_C, a, d)$ is continuous.
(Proof)

Consider a sequence $\{(\mu^n_C, a^n, d^n)\}_{n=1}^{\infty}$ in $[0,1] \times [c, b] \times [c, b]$ such that $0 \leq \mu^n_C \leq \bar{\mu}_C(a^n, d^n)$, $c < a^n < b$, and $c < d^n < a^n$ for each n. Suppose that the sequence converges. Let $(\mu^*_C, a^*, d^*) = \lim_{n\to\infty}(\mu^n_C, a^n, d^n)$. Then, it follows from Step 5 that

$$0 \leq \mu^*_C = \lim_{n\to\infty} \mu^n_C \leq \lim_{n\to\infty} \bar{\mu}_C(a^n, d^n) = \bar{\mu}_C(\lim_{n\to\infty} a^n, \lim_{n\to\infty} d^n) = \bar{\mu}_C(a^*, d^*).$$

Then, there exists the $\hat{\mu}(\mu^*_C, a^*, d^*)$ established in Step 4.

Examine the corresponding sequence $\{\hat{\mu}(\mu_C^n, a^n, d^n)\}_{n=1}^{\infty}$ generated by $\hat{\mu}$. Suppose a convergent subsequence $\{\hat{\mu}(\mu_C^{n_m}, a^{n_m}, d^{n_m})\}_{m=1}^{\infty}$ of it arbitrarily. Let $\hat{\mu}^* = \lim_{m\to\infty} \hat{\mu}(\mu_C^{n_m}, a^{n_m}, d^{n_m})$. Then,

$$\hat{\mu}_C^* = \lim_{m\to\infty} \hat{\mu}_C(\mu_C^{n_m}, a^{n_m}, d^{n_m}) = \lim_{m\to\infty} \mu_C^{n_m} = \mu_C^*$$

where the second equality holds because $\hat{\mu}_C(\mu_C^{n_m}, a^{n_m}, d^{n_m}) = \mu_C^{n_m}$ holds for each n_m by the definition of $\hat{\mu}(\mu_C, a, d)$. Additionally,

$$\begin{aligned}
\psi_C(\hat{\mu}^*, a^*, d^*) &= \lim_{m\to\infty} \psi_C(\hat{\mu}(\mu_C^{n_m}, a^{n_m}, d^{n_m}), a^{n_m}, d^{n_m}) \\
&= \lim_{m\to\infty} \hat{\mu}_C(\mu_C^{n_m}, a^{n_m}, d^{n_m}) \\
&= \lim_{n\to\infty} \mu_C^{n_m} \\
&= \mu_C^*
\end{aligned}$$

where the first equality holds because $\psi(\mu, a, d)$ is continuous by the proof of Lemma 1, and the second and third equalities hold because $\psi_C(\hat{\mu}(\mu_C^{n_m}, a^{n_m}, d^{n_m}), a^{n_m}, d^{n_m}) = \hat{\mu}_C(\mu_C^{n_m}, a^{n_m}, d^{n_m}) = \mu_C^{n_m}$ holds for each n_m by the definition of $\hat{\mu}(\mu_C, a, d)$. Thus,

$$\psi_C(\hat{\mu}^*, a^*, d^*) = \hat{\mu}_C^* = \mu_C^*.$$

Then, $\hat{\mu}(\mu_C^*, a^*, d^*) = \hat{\mu}^*$ because $\hat{\mu}(\mu_C^*, a^*, d^*)$ is the unique $\hat{\mu}$ that satisfies $\psi_C(\hat{\mu}) = \hat{\mu}_C = \mu_C^*$. Thus, we know that any convergent subsequence of $\{\hat{\mu}(\mu_C^n, a^n, d^n)\}_{n=1}^{\infty}$ converges to $\hat{\mu}(\mu_C^*, a^*, d^*)$. Then, the sequence $\{\hat{\mu}(\mu_C^n, a^n, d^n)\}_{n=1}^{\infty}$ itself converges to $\hat{\mu}(\mu_C^*, a^*, d^*)$ because it is a sequence in a compact set Δ. This means that $\hat{\mu}(\mu_C, a, d)$ is continuous. (Q.E.D.)

Proof of Lemma 5

Suppose that a belief μ is a three-mode equilibrium distribution. Then, $\mu = \psi(\mu)$ and $\mu_C > 0$ by Lemma 1. By the way of construction of $\hat{\mu}(\mu_C)$ and $\lambda(\mu_C)$, it holds for μ with $\mu = \psi(\mu)$ that $\lambda(\mu_C) = 0$ and $\mu = \hat{\mu}(\mu_C)$.

Conversely, suppose that μ satisfies conditions (1) through (3) in Lemma 5. Then, $\mu = \hat{\mu}(\mu_C)$ means that $\mu_C = \psi_C(\mu)$. Together with $\mu = \hat{\mu}(\mu_C)$, $\lambda(\mu_C) = 0$ means that $\mu_{DDD} = \psi_{DDD}(\mu)$. Therefore, $\mu = \psi(\mu)$. Finally, by Lemma 4-(2),

$$\lambda(\bar{\mu}_C) = \psi_{DDD}(\hat{\mu}(\bar{\mu}_C)) - \hat{\mu}_{DDD}(\bar{\mu}_C) = (1 - \bar{\mu}_C) - 0 > 0.$$

This implies that $\mu_C \neq \bar{\mu}_C$. Then, by Lemma 4-(3), $\mu_C \in (0, \bar{\mu}_C)$ implies that $\mu_{CDD} > 0$ and $\mu_{DDD} > 0$. Thus, this μ is a three-mode equilibrium. (Q.E.D.)

Proof of Theorem 1

Suppose that condition (14) holds. By Lemma 4, $\hat{\mu}(0) = (0, \frac{d-c}{a-c}, \frac{a-d}{a-c})$ and $\psi(\hat{\mu}(0)) = (0, \phi(\boldsymbol{\beta} > \beta^*), \phi(\boldsymbol{\beta} < \beta^*))$. Then, condition (14) guarantees

$$\lambda(0) = \psi_{DDD}(\hat{\mu}(0)) - \hat{\mu}_{DDD}(0) = \phi(\boldsymbol{\beta} < \beta^*) - \frac{a-d}{a-c} < 0.$$

On the other hand, by Lemma 4, $\bar{\mu}_C < 1$ and $\hat{\mu}(\bar{\mu}_C) = (\bar{\mu}_C, 1 - \bar{\mu}_C, 0)$ and $\psi(\hat{\mu}(\bar{\mu}_C)) = (\bar{\mu}_C, 0, 1 - \bar{\mu}_C)$. Then,

$$\lambda(\bar{\mu}_C) = \psi_{DDD}(\hat{\mu}(\bar{\mu}_C)) - \hat{\mu}_{DDD}(\bar{\mu}_C) = (1 - \bar{\mu}_C) - 0 > 0.$$

Finally, $\lambda(\mu_C)$ is continuous because $\psi(\mu)$ is continuous by Lemma 1 and $\hat{\mu}(\mu_C)$ is continuous by Lemma 4.

These establish that there exists $\mu_C \in (0, \bar{\mu}_C)$ such that $\lambda(\mu_C) = 0$. Set $\mu = \hat{\mu}(\mu_C)$. Then, this μ satisfies conditions (1), (2), and (3) in Lemma 5. Hence, this μ is a three-mode equilibrium distribution. (Q.E.D.)

Proof of Corollary 1

Corollary 1-(2) is straightforward from Theorem 1 and the definition of $\bar{d}(a)$. To see 1-(1), recall that $\beta^* = \frac{b-a}{b-c}$. Then, $c < \bar{d}(a) < a$ is immediate by the definition of $\bar{d}(a)$, because $0 < \beta^* < 1$ so that $0 < \phi(\boldsymbol{\beta} > \beta^*) < 1$. It holds that $\frac{d}{da}\bar{d}(a) = \phi(\boldsymbol{\beta} > \beta^*) + (a-c)(\frac{d}{d\beta^*}\phi(\boldsymbol{\beta} > \beta^*))\frac{d\beta^*}{da} > 0$, because $\frac{d}{d\beta^*}\phi(\boldsymbol{\beta} > \beta^*) < 0$ and $\frac{d\beta^*}{da} < 0$. From $c < \bar{d}(a) < a$, it immediately follows that $\lim_{a \to c} \bar{d}(a) = c$. Finally, it holds that $\lim_{a \to b} \bar{d}(a) = b$ because $\lim_{a \to b} \beta^* = 0$ so that $\lim_{a \to b} \phi(\boldsymbol{\beta} > \beta^*) = 1$. (Q.E.D.)

Proof of Lemma 6

To show Lemma 6-(1), recall from the proof of Lemma 4 that $\bar{\mu}_C(a, d)$ is a unique solution $\bar{\mu}_C$ to

$$\bar{\mu}_C - \phi\Big(\bar{\mu}_C(b-c)\boldsymbol{\beta} \geq \bar{\mu}_C(b-d) - (a-d)\Big) = 0.$$

Note that $\phi\Big(\mu_C(b-c)\boldsymbol{\beta} \geq \mu_C(b-d) - (a-d)\Big)$ is 1 for $\mu_C \in [0, \frac{a-d}{b-d}]$ and strictly decreasing in μ_C over $\mu_C \in (\frac{a-d}{b-d}, 1)$ so that the solution $\bar{\mu}_C$ lies in the latter region $(\frac{a-d}{b-d}, 1)$. Then, the solution $\bar{\mu}_C$ is strictly increasing in a and strictly decreasing in d because

$$\frac{\partial}{\partial a}\phi\Big(\bar{\mu}_C(b-c)\boldsymbol{\beta} \geq \bar{\mu}_C(b-d) - (a-d)\Big) > 0$$
$$\frac{\partial}{\partial d}\phi\Big(\bar{\mu}_C(b-c)\boldsymbol{\beta} \geq \bar{\mu}_C(b-d) - (a-d)\Big) < 0.$$

To show Lemma 6-(2), fix $d \in (c, b)$ arbitrarily and compare arbitrary a', a'' with $c < d < a' < a'' < b$. By Lemma 6-(1) established above, $\bar{\mu}_C(a', d) < \bar{\mu}_C(a'', d)$ so that both $\lambda(\mu_C, a', d)$ and $\lambda(\mu_C, a'', d)$ are defined over $[0, \bar{\mu}_C(a', d)]$. We will show in two steps below that for each $\mu_C \in [0, \bar{\mu}_C(a', d)]$

1. $\hat{\mu}_{DDD}(\mu_C, a', d) < \hat{\mu}_{DDD}(\mu_C, a'', d)$, and
2. $\psi_{DDD}(\hat{\mu}(\mu_C, a', d), a', d) \geq \psi_{DDD}(\hat{\mu}(\mu_C, a'', d), a'', d)$.

Then, it is established for any $\mu_C \in [0, \bar{\mu}_C(a', d)]$ that

$$\begin{aligned}\lambda(\mu_C, a', d) &= \psi_{DDD}(\hat{\mu}(\mu_C, a', d), a', d) - \hat{\mu}_{DDD}(\mu_C, a', d) \\ &> \psi_{DDD}(\hat{\mu}(\mu_C, a'', d), a'', d) - \hat{\mu}_{DDD}(\mu_C, a'', d) \\ &= \lambda(\mu_C, a'', d).\end{aligned}$$

It is proved in a similar way that $\lambda(\mu_C, a, d)$ is strictly increasing in d. (The proof is simpler because β^* is independent of d.)

[Step 1] $\hat{\mu}_{DDD}(\mu_C, a', d) < \hat{\mu}_{DDD}(\mu_C, a'', d)$ for each $\mu_C \in [0, \bar{\mu}_C(a', d)]$.

(Proof)

Consider $\mu_C = 0$. Then, by Lemma 4, $\hat{\mu}(0, a', d) = (0, \frac{d-c}{a'-c}, \frac{a'-d}{a'-c})$ and $\hat{\mu}(0, a'', d) = (0, \frac{d-c}{a''-c}, \frac{a''-d}{a''-c})$. Therefore, $\hat{\mu}_{DDD}(\mu_C, a', d) = \frac{a'-d}{a'-c} < \frac{a''-d}{a''-c} = \hat{\mu}_{DDD}(\mu_C, a'', d)$ when $a' < a''$.

Consider $\mu_C \in (0, \bar{\mu}_C(a', d)]$. Suppose to the contrary that $\hat{\mu}_{DDD}(\mu_C, a', d) \geq \hat{\mu}_{DDD}(\mu_C, a'', d)$. Then, on one hand, we have

$$\begin{aligned}\psi_C(\hat{\mu}(\mu_C, a'', d), a', d) \geq \psi_C(\hat{\mu}(\mu_C, a', d), a', d) &= \mu_C \\ &= \psi_C(\hat{\mu}(\mu_C, a'', d), a'', d) \\ &= \hat{\mu}_C(\mu_C, a'', d)\end{aligned}$$

where the first inequality follows from $\hat{\mu}_{DDD}(\mu_C, a', d) \geq \hat{\mu}_{DDD}(\mu_C, a'', d)$ by Step 3 of the proof of Lemma 4.

On the other hand, by $a' < a''$, we have

$$\psi_C(\hat{\mu}(\mu_C, a'', d), a'', d) > \psi_C(\hat{\mu}(\mu_C, a'', d), a', d)$$

for the following reason. Take $(\alpha, \beta) \in T^*_C(\hat{\mu}(\mu_C, a'', d), a', d)$. Then,

$$\begin{aligned}U_{(\alpha,\beta)}(C, \hat{\mu}(\mu_C, a'', d), a', d) &\geq U_{(\alpha,\beta)}(CDD, \hat{\mu}(\mu_C, a'', d), a', d) \\ U_{(\alpha,\beta)}(C, \hat{\mu}(\mu_C, a'', d), a', d) &\geq U_{(\alpha,\beta)}(DDD, \hat{\mu}(\mu_C, a'', d), a', d).\end{aligned}$$

Then, together with $a' < a''$, we have

$$\begin{aligned}&U_{(\alpha,\beta)}(C, \hat{\mu}(\mu_C, a'', d), a'', d) - U_{(\alpha,\beta)}(CDD, \hat{\mu}(\mu_C, a'', d), a'', d) \\ &= U_{(\alpha,\beta)}(C, \hat{\mu}(\mu_C, a'', d), a', d) - U_{(\alpha,\beta)}(CDD, \hat{\mu}(\mu_C, a'', d), a', d) \\ &\quad + \hat{\mu}_{CDD}(\mu_C, a'', d)(a'' - a') \\ &\geq U_{(\alpha,\beta)}(C, \hat{\mu}(\mu_C, a'', d), a', d) - U_{(\alpha,\beta)}(CDD, \hat{\mu}(\mu_C, a'', d), a', d) \\ &\geq 0\end{aligned}$$

and

$$\begin{aligned}&U_{(\alpha,\beta)}(C, \hat{\mu}(\mu_C, a'', d), a'', d) - U_{(\alpha,\beta)}(DDD, \hat{\mu}(\mu_C, a'', d), a'', d) \\ &= U_{(\alpha,\beta)}(C, \hat{\mu}(\mu_C, a'', d), a', d) - U_{(\alpha,\beta)}(DDD, \hat{\mu}(\mu_C, a, d), a', d) \\ &\quad + \hat{\mu}_C(\mu_C, a'', d)(a'' - a') + \hat{\mu}_{CDD}(\mu_C, a'', d)(a'' - a') \\ &\geq U_{(\alpha,\beta)}(C, \hat{\mu}(\mu_C, a'', d), a', d) - U_{(\alpha,\beta)}(DDD, \hat{\mu}(\mu_C, a'', d), a', d) \\ &\geq 0.\end{aligned}$$

Therefore, $(\alpha, \beta) \in T^*_C(\hat{\mu}(\mu_C, a'', d), a'', d)$. Hence, $T^*_C(\hat{\mu}(\mu_C, a'', d), a', d) \subseteq T^*_C(\hat{\mu}(\mu_C, a'', d), a'', d)$. Furthermore, $T^*_C(\hat{\mu}(\mu_C, a'', d), a', d) \subsetneqq T^*_C(\hat{\mu}(\mu_C, a'', d), a'', d)$ by a similar argument to Step 3 of the proof of Lemma 4. Hence, $\psi_C(\hat{\mu}(\mu_C, a'', d), a'', d) > \psi_C(\hat{\mu}(\mu_C, a'', d), a', d)$.

By combining the above inequalities, we have

$$\psi_C(\hat{\mu}(\mu_C, a'', d), a'', d) > \psi_C(\hat{\mu}(\mu_C, a'', d), a', d) \geq \hat{\mu}_C(\mu_C, a'', d).$$

This contradicts the supposition that $\psi_C(\hat{\mu}(\mu_C, a'', d), a'', d) = \hat{\mu}_C(\mu_C, a'', d)$. Hence, it must be the case that $\hat{\mu}_{DDD}(\mu_C, a', d) < \hat{\mu}_{DDD}(\mu_C, a'', d)$.

[**Step 2**] $\psi_{DDD}(\hat{\mu}(\mu_C, a', d), a', d) \geq \psi_{DDD}(\hat{\mu}(\mu_C, a'', d), a'', d)$ for each $\mu_C \in [0, \bar{\mu}_C(a', d)]$.

(Proof)

First, we show that $\alpha^*(\hat{\mu}(\mu_C, a', d), a', d) > \alpha^*(\hat{\mu}(\mu_C, a'', d), a'', d)$. Suppose to the contrary that $\alpha^*(\hat{\mu}(\mu_C, a', d), a', d) \leq \alpha^*(\hat{\mu}(\mu_C, a'', d), a'', d)$. Recall from (6) that

$$\begin{aligned}H(\alpha|\hat{\mu}(\mu_C, a', d), a', d) &= \frac{\hat{\mu}_{DDD}(\mu_C, a', d)}{\hat{\mu}_C(\mu_C, a', d)}(\alpha - \alpha^*(\hat{\mu}(\mu_C, a', d), a', d)) + \beta^*(a') \\ H(\alpha|\hat{\mu}(\mu_C, a'', d), a'', d) &= \frac{\hat{\mu}_{DDD}(\mu_C, a'', d)}{\hat{\mu}_C(\mu_C, a'', d)}(\alpha - \alpha^*(\hat{\mu}(\mu_C, a'', d), a'', d)) + \beta^*(a'')\end{aligned}$$

where we denote $\beta^*(a') = \frac{b-a'}{b-c}$ and $\beta^*(a'') = \frac{b-a''}{b-c}$ to express the fact that β^* depends on a. Note that the slopes $\frac{\hat{\mu}_{DDD}(\mu_C,a',d)}{\hat{\mu}_C(\mu_C,a',d)}$ and $\frac{\hat{\mu}_{DDD}(\mu_C,a'',d)}{\hat{\mu}_C(\mu_C,a'',d)}$ are related as $\frac{\hat{\mu}_{DDD}(\mu_C,a',d)}{\hat{\mu}_C(\mu_C,a',d)} < \frac{\hat{\mu}_{DDD}(\mu_C,a'',d)}{\hat{\mu}_C(\mu_C,a'',d)}$ because $\hat{\mu}_C(\mu_C,a',d) = \mu_C = \hat{\mu}_C(\mu_C,a'',d)$ and $\hat{\mu}_{DDD}(\mu_C,a',d) < \hat{\mu}_{DDD}(\mu_C,a'',d)$ by Step 1. Note also that $\beta^*(a') = \frac{b-a'}{b-c} > \frac{b-a''}{b-c} = \beta^*(a'')$ when $a' < a''$. Then, under the supposition that $\alpha^*(\hat{\mu}(\mu_C,a',d),a',d) \leq \alpha^*(\hat{\mu}(\mu_C,a'',d),a'',d)$, we have

$$\begin{aligned}
T^*_C(\hat{\mu}(\mu_C,a',d),a',d) &= T \cap \{(\alpha,\beta)|\alpha \leq \alpha^*(\hat{\mu}(\mu_C,a',d),a',d), \beta \geq H(\alpha|\hat{\mu}(\mu_C,a',d),a',d)\} \\
&\subsetneqq T \cap \{(\alpha,\beta)|\alpha \leq \alpha^*(\hat{\mu}(\mu_C,a'',d),a'',d), \beta \geq H(\alpha|\hat{\mu}(\mu_C,a'',d),a'',d)\} \\
&= T^*_C(\hat{\mu}(\mu_C,a'',d),a'',d).
\end{aligned}$$

This means that $\hat{\mu}_C(\mu_C,a',d) = \phi(T^*_C(\hat{\mu}(\mu_C,a',d),a',d) < \phi(T^*_C(\hat{\mu}(\mu_C,a'',d),a'',d) = \hat{\mu}_C(\mu_C,a'',d)$. This contradicts $\hat{\mu}_C(\mu_C,a',d) = \mu_C = \hat{\mu}_C(\mu_C,a'',d)$.

Then, together with $\beta^*(a') = \frac{b-a'}{b-c} > \frac{b-a''}{b-c} = \beta^*(a'')$, the relation $\alpha^*(\hat{\mu}(\mu_C,a',d),a',d) > \alpha^*(\hat{\mu}(\mu_C,a'',d),a'',d)$ thus established implies that

$$\begin{aligned}
\psi_{CDD}(\hat{\mu}(\mu_C,a',d),a',d) &= \phi(T^*_{CDD}(\hat{\mu}(\mu_C,a',d),a',d)) \\
&= \phi(T \cap \{(\alpha,\beta)|\alpha > \alpha^*(\hat{\mu}(\mu_C,a',d),a',d), \beta > \beta^*(a')\}) \\
&\leq \phi(T \cap \{(\alpha,\beta)|\alpha > \alpha^*(\hat{\mu}(\mu_C,a'',d),a'',d), \beta > \beta^*(a'')\}) \\
&= \phi(T^*_{CDD}(\hat{\mu}(\mu_C,a'',d),a'',d)) \\
&= \psi_{CDD}(\hat{\mu}(\mu_C,a'',d),a'',d).
\end{aligned}$$

Hence, together with $\psi_C(\hat{\mu}(\mu_C,a',d),a',d) = \mu_C = \psi_C(\hat{\mu}_C(\mu_C,a'',d),a'',d)$, we conclude that

$$\begin{aligned}
\psi_{DDD}(\hat{\mu}(\mu_C,a',d),a',d) &= 1 - \psi_C(\hat{\mu}(\mu_C,a',d),a',d) - \psi_{CDD}(\hat{\mu}(\mu_C,a',d),a',d) \\
&\geq 1 - \psi_C(\hat{\mu}(\mu_C,a'',d),a'',d) - \psi_{CDD}(\hat{\mu}(\mu_C,a'',d),a'',d) \\
&= \psi_{DDD}(\hat{\mu}(\mu_C,a'',d),a'',d).
\end{aligned}$$

(Q.E.D.)

Proof of Theorem 2

[**Step 1**] Fix $a \in (c,b)$. We will show how to find $\hat{d}(a)$ with the desired properties.

First, recall from Lemma 6-(1) that $\bar{\mu}_C(a,d)$ is strictly decreasing in $d \in (c,a)$. This allows us to define $\bar{\bar{\mu}}_C(a) = \lim_{d \downarrow c} \bar{\mu}_C(a,d)$ because $\bar{\mu}_C(a,d)$ is a probability and bounded from above by 1.

For each $\mu_C \in [0, \bar{\bar{\mu}}_C(a))$, there exists $d' \in (c,a)$ such that $\mu_C \in [0, \bar{\mu}_C(a,d')]$ so that $\lambda(\mu_C,a,d')$ is defined. Then, $\lambda(\mu_C,a,d)$ is defined for every $d \in (c,d')$ because $d < d'$ means $\bar{\mu}_C(a,d) > \bar{\mu}_C(a,d')$. Over the interval (c,d'), $\lambda(\mu_C,a,d)$ is increasing in d by Lemma 6-(2). This allows us to define $\underline{\lambda}(\mu_C,a) = \lim_{d \downarrow c} \lambda(\mu_C,a,d)$ because $\lambda(\mu_C,a,d)$ is a difference in two probabilities and bounded from below by -1.

Define $M_C(a) = \{\mu_C \in [0, \bar{\bar{\mu}}_C(a))|\underline{\lambda}(\mu_C,a) < 0\}$. Note from the proof of Theorem 1 that $\lambda(0,a,d) < 0$ for any $d \in (c, \bar{d}(a))$. This implies that $\underline{\lambda}(0,a) < 0$. Therefore, $0 \in M_C(a)$. Thus, $M_C(a)$ is nonempty.

Consider $\mu_C \in M_C(a) \setminus 0$. Then, by the way of construction of $\underline{\lambda}(\mu_C,a)$, $\underline{\lambda}(\mu_C,a) < 0$ guarantees that there exists some $d' \in (c,a)$ such that $\lambda(\mu_C,a,d') < 0$. On the other hand, we can show the following fact, for which we postpone the proof to Step 3 below.

Claim 1: *For $\mu_C \in M_C(a) \setminus 0$, there exists some $d'' \in (c,a)$ such that $\mu_C \leq \bar{\mu}_C(a,d'')$ and $\lambda(\mu_C,a,d'') > 0$.*

For d' and d'' thus selected for the given $\mu_C \in M_C(a) \setminus 0$, it immediately follows from $\lambda(\mu_C, a, d') < 0$ and $\lambda(\mu_C, a, d'') > 0$ that $d' < d''$ because $\lambda(\mu_C, a, d)$ is increasing in d by Lemma 6-(2). Then, there exists a unique $d^* \in (d', d'')$ in the interval (c, a) such that $\lambda(\mu_C, a, d^*) = 0$ because $\lambda(\mu_C, a, d)$ is continuous and strictly increasing in d. We write $d^*(\mu_C, a)$ to express the fact that this d^* depends on (μ_C, a).

For $\mu_C = 0 \in M_C(a)$, note from the proof of Theorem 1 that $\lambda(0, a, \bar{d}(a)) = 0$. So we set $d^*(0, a) = \bar{d}(a)$.

Now define $\hat{d}(a) \equiv \sup_{\mu_C \in M_C(a)} d^*(\mu_C, a)$. We can show the following fact, for which we postpone the proof to Step 3 below.

Claim 2: *There exists* $\hat{\mu}_C \in M_C(a)$ *that achieves* $\sup_{\mu_C \in M_C(a)} d^*(\mu_C, a)$ *so that* $\hat{d}(a) = \max_{\mu_C \in M_C(a)} d^*(\mu_C, a) = d^*(\hat{\mu}_C, a)$.

We will show that this $\hat{d}(a)$ is what Theorem 2 claims.

[**Step 2**] We will show that $\hat{d}(a)$ has the properties claimed in Theorem 2.

We show Theorem 2-(1). Fix $d \in (c, \hat{d}(a))$. Then, $d < \hat{d}(a)$ guarantees that there exists some $\mu'_C \in M_C(a)$ such that $d < d^*(\mu'_C, a) \leq \hat{d}(a)$. This means $\lambda(\mu'_C, a, d) < 0$ because $\lambda(\mu'_C, a, d^*(\mu'_C, a)) = 0$ and $\lambda(\mu'_C, a, \tilde{d})$ is strictly increasing in $\tilde{d}$. Recall from the proof of Theorem 1 that $\lambda(\bar{\mu}_C(a, d), a, d) > 0$. Then, there exists $\mu_C \in (\mu'_C, \bar{\mu}_C(a, d))$ such that $\lambda(\mu_C, a, d) = 0$ because $\lambda(\tilde{\mu}_C, a, d)$ is continuous in $\tilde{\mu}_C$. Set $\mu = \hat{\mu}(\mu_C)$. Then, this μ is a three-mode equilibrium distribution in a prisoner's dilemma $PD((a, b, c, d), f)$.

We show Theorem 2-(2). It is immediate from the way of construction of $\hat{d}(a)$ that if $\hat{d}(a) < d < a$, then there is no $\mu_C \in M_C(a) \setminus 0$ that satisfies $\lambda(\mu_C, a, d) = 0$. If $\mu_C \notin M_C(a)$, $\lambda(\mu_C, a, d) = 0$ never holds because if it were the case that $\lambda(\mu_C, a, d) = 0$, it must be the case that $\lambda(\mu_C, a, \frac{1}{2}c + \frac{1}{2}d) < 0$, a contradiction to $\mu_C \notin M_C(a)$. Hence, if $\hat{d}(a) < d < a$, there is no three-mode equilibrium in a prisoner's dilemma $PD((a, b, c, d), f)$.

We show the remaining properties of $\hat{d}(a)$. First, we show that $\bar{d}(a) \leq \hat{d}(a) < a$. Note that $d^*(0, a) = \bar{d}(a)$. Therefore, $\hat{d}(a) = \sup_{\mu_C \in M_C(a)} d^*(\mu_C, a) \geq d^*(0, a) = \bar{d}(a)$. If $\bar{d}(a) = \hat{d}(a)$, it is also immediate that $\hat{d}(a) < a$ because $\bar{d}(a) < a$ by Corollary 1. Consider the other case of $\bar{d}(a) < \hat{d}(a)$. Then, $\hat{d}(a) = \max_{\mu_C \in M_C(a)} d^*(\mu_C, a) = d^*(\hat{\mu}_C, a)$ for some $\hat{\mu}_C \in M_C(a) \setminus 0$ by Claim 2. Hence, it follows that $\hat{d}(a) < a$ because $c < d^*(\hat{\mu}_C, a) < a$.

Second, from the fact that $c < \bar{d}(a) \leq \hat{d}(a) < a < b$, which we have just established, it immediately follows that $\lim_{a \to c} \hat{d}(a) = c$ and it also follows that $\lim_{a \to b} \hat{d}(a) = b$, because $\lim_{a \to b} \bar{d}(a) = b$ by Corollary 1.

Third, we show that $\hat{d}(a)$ is strictly increasing. Fix $a', a'' \in (c, b)$ such that $a' < a''$. We will show below that

1. $M_C(a') \subseteq M_C(a'')$, and
2. $d^*(\mu_C, a') < d^*(\mu_C, a'')$ for every $\mu_C \in M_C(a')$.

Let us show the first fact that $M_C(a') \subseteq M_C(a'')$. Recall from Lemma 6-(1) that $a' < a''$ implies $\bar{\mu}_C(a', d) < \bar{\mu}_C(a'', d)$ for each $d \in (c, a')$. This means that $\bar{\bar{\mu}}_C(a') = \lim_{d \downarrow c} \bar{\mu}_C(a', d) \leq \lim_{d \downarrow c} \bar{\mu}_C(a'', d) = \bar{\bar{\mu}}_C(a'')$. Therefore, $[0, \bar{\bar{\mu}}_C(a')) \subseteq [0, \bar{\bar{\mu}}_C(a''))$ and $\underline{\lambda}(\mu_C, a')$ is defined over $\mu_C \in [0, \bar{\bar{\mu}}_C(a'))$, while $\underline{\lambda}(\mu_C, a'')$ is defined over $\mu_C \in [0, \bar{\bar{\mu}}_C(a''))$. Now, take $\mu_C \in M_C(a')$. Then, $\mu_C \in [0, \bar{\bar{\mu}}_C(a'))$ and $\underline{\lambda}(\mu_C, a') < 0$. It follows from $\mu_C \in [0, \bar{\bar{\mu}}_C(a'))$ that $\mu_C \in [0, \bar{\bar{\mu}}_C(a''))$ and $\underline{\lambda}(\mu_C, a'')$ is defined. Furthermore, if $\lambda(\mu_C, a', d)$ is defined for d such that $\mu_C \in [0, \bar{\mu}_C(a', d)]$, then $\bar{\mu}_C(a', d) < \bar{\mu}_C(a'', d)$ guarantees that $\lambda(\mu_C, a'', d)$ is also defined for the d, and Lemma 6-(2)

implies that $\lambda(\mu_C, a', d) > \lambda(\mu_C, a'', d)$. Hence, $\underline{\lambda}(\mu_C, a') = \lim_{d \downarrow c} \lambda(\mu_C, a', d) \geq \lim_{d \downarrow c} \lambda(\mu_C, a'', d) = \underline{\lambda}(\mu_C, a'')$. This leads us to conclude that $0 > \underline{\lambda}(\mu_C, a') \geq \underline{\lambda}(\mu_C, a'')$. Hence $\mu_C \in M_C(a'')$.

Let us show the second fact that $d^*(\mu_C, a') < d^*(\mu_C, a'')$ for every $\mu_C \in M_C(a')$. By way of construction, it holds that $\lambda(\mu_C, a', d^*(\mu_C, a')) = 0$ and $\lambda(\mu_C, a'', d^*(\mu_C, a'')) = 0$. It follows from Lemma 6-(2) that $a' < a''$ implies $0 = \lambda(\mu_C, a', d^*(\mu_C, a')) > \lambda(\mu_C, a'', d^*(\mu_C, a'))$. Then, it follows from Lemma 6-(2) that $\lambda(\mu_C, a'', d^*(\mu_C, a'')) = 0 > \lambda(\mu_C, a'', d^*(\mu_C, a'))$ implies $d^*(\mu_C, a'') > d^*(\mu_C, a')$.

Then, it follows from the second fact that $\hat{d}(a') = \sup_{\mu_C \in M_C(a')} d^*(\mu_C, a') = d^*(\hat{\mu}_C, a') < d^*(\hat{\mu}_C, a'')$ where $\hat{\mu}_C$ is the $\hat{\mu}_C$ for $a = a'$ stated in Claim 2. It follows from the first fact that $d^*(\hat{\mu}_C, a'') \leq \sup_{\mu_C \in M_C(a')} d^*(\mu_C, a'') \leq \sup_{\mu_C \in M_C(a'')} d^*(\mu_C, a'') = \hat{d}(a'')$. These establish that $\hat{d}(a') < \hat{d}(a'')$.

Finally, we show that $\hat{d}(a)$ is continuous. Fix $a \in (c, b)$. Consider a sequence $\{a^n\}_{n=1}^{\infty}$ such that $a^n \leq a$ and $\lim_{n\to\infty} a^n = a$. We show that $\lim_{n\to\infty} \hat{d}(a^n) = \hat{d}(a)$. Without loss of generality, we consider an increasing sequence $\{a^n\}_{n=1}^{\infty}$; that is, $a^n \leq a^{n+1}$. Then, $\lim_{n\to\infty} \hat{d}(a^n)$ exists and $\lim_{n\to\infty} \hat{d}(a^n) \leq \hat{d}(a)$ because $\hat{d}(\tilde{a})$ is strictly increasing so that $\hat{d}(a^n) \leq \hat{d}(a^{n+1}) \leq \hat{d}(a)$. We show that $\lim_{n\to\infty} \hat{d}(a^n) \geq \hat{d}(a)$. Let $\hat{\mu}_C$ achieve $\hat{d}(a) = \max_{\mu_C \in M_C(a)} d^*(\mu_C, a) = d^*(\hat{\mu}_C, a)$ as in Claim 2. Note that $\underline{\lambda}(\hat{\mu}_C, a) < 0$ because $\hat{\mu}_C \in M_C(a)$. Therefore, there exists $d \in (c, a)$ such that $\lambda(\hat{\mu}_C, a, d)$ is defined and $\lambda(\hat{\mu}_C, a, d) < 0$. Then, there exists N such that $\lambda(\hat{\mu}_C, a^n, d) < 0$ for every $n \geq N$ because $\lim_{n\to\infty} a^n = a$ and $\lambda(\hat{\mu}_C, \tilde{a}, d)$ is continuous in $\tilde{a}$. This means that $\underline{\lambda}(\hat{\mu}_C, a^n) < 0$ for every $n \geq N$ so that $\hat{\mu}_C \in M_C(a^n)$ for every $n \geq N$. Then, $d^*(\hat{\mu}_C, a^n)$ is defined for every $n \geq N$. Note that $d^*(\tilde{\mu}_C, \tilde{a})$ is continuous in $(\tilde{\mu}_C, \tilde{a})$, because $d^*(\tilde{\mu}_C, \tilde{a})$ is a unique solution to $\lambda(\tilde{\mu}_C, \tilde{a}, d) = 0$ and $\lambda(\tilde{\mu}_C, \tilde{a}, d)$ is continuous in $(\tilde{\mu}_C, \tilde{a}, d)$. Therefore, $\lim_{n\to\infty} d^*(\hat{\mu}_C, a^n) = d^*(\hat{\mu}_C, a)$. Hence, $\lim_{n\to\infty} \hat{d}(a^n) = \lim_{n\to\infty} \sup_{\mu_C \in M_C(a^n)} d^*(\mu_C, a^n) \geq \lim_{n\to\infty} d^*(\hat{\mu}_C, a^n) = d^*(\hat{\mu}_C, a) = \hat{d}(a)$.

Consider a sequence $\{a^n\}_{n=1}^{\infty}$ such that $a^n \geq a$ and $\lim_{n\to\infty} a^n = a$. Without loss of generality, we consider a decreasing sequence $\{a^n\}_{n=1}^{\infty}$; that is, $a^n \geq a^{n+1}$. Then, $\lim_{n\to\infty} \hat{d}(a^n)$ exists and $\lim_{n\to\infty} \hat{d}(a^n) \geq \hat{d}(a)$ because $\hat{d}(\tilde{a})$ is strictly increasing so that $\hat{d}(a^n) \geq \hat{d}(a^{n+1}) \geq \hat{d}(a)$. We show by contradiction that $\lim_{n\to\infty} \hat{d}(a^n) = \hat{d}(a)$. Suppose that $\lim_{n\to\infty} \hat{d}(a^n) > \hat{d}(a)$. Let $\hat{\mu}_C^n$ achieve $\hat{d}(a^n) = \max_{\mu_C \in M_C(a^n)} d^*(\mu_C, a^n) = d^*(\hat{\mu}_C^n, a^n)$ for $a = a^n$ in Claim 2. Then, a sequence $\{\hat{\mu}_C^n\}_{n=1}^{\infty}$ has a convergent subsequence because each $\hat{\mu}_C^n$ is in a compact set $[0, 1]$. Without loss of generality, we assume that $\{\hat{\mu}_C^n\}_{n=1}^{\infty}$ itself is convergent, and we denote $\lim_{n\to\infty} \hat{\mu}_C^n = \hat{\mu}_C$. Then, $\hat{\mu}_C \in [0, \bar{\mu}_C(a, \lim_{n\to\infty} d^*(\hat{\mu}_C^n, a))]$ because $\hat{\mu}_C^n \in [0, \bar{\mu}_C(a^n, d^*(\hat{\mu}_C^n, a^n))]$, $\lim_{n\to\infty} \hat{\mu}_C^n = \hat{\mu}_C$, $\lim_{n\to\infty} a^n = a$, and $\bar{\mu}_C(\tilde{a}, \tilde{d})$ is continuous in $(\tilde{a}, \tilde{d})$. Therefore, $\lambda(\hat{\mu}_C, a, \lim_{n\to\infty} d^*(\hat{\mu}_C^n, a))$ is defined. It must be the case that $\lambda(\hat{\mu}_C, a, \lim_{n\to\infty} d^*(\hat{\mu}_C^n, a)) \leq 0$ because if it were the case that $\lambda(\hat{\mu}_C, a, \lim_{n\to\infty} d^*(\hat{\mu}_C^n, a)) > 0$, there must exist N such that $\lambda(\hat{\mu}_C^N, a^N, d^*(\hat{\mu}_C^N, a)) > \frac{1}{2}\lambda(\hat{\mu}_C^N, a^N, \lim_{n\to\infty} d^*(\hat{\mu}_C^n, a)) > 0$, a contradiction to the condition that $\lambda(\hat{\mu}_C^N, a^N, d^*(\hat{\mu}_C^N, a)) = 0$. We supposed that $\lim_{n\to\infty} d^*(\hat{\mu}_C^n, a^n) = \lim_{n\to\infty} \hat{d}(a^n) > \hat{d}(a)$. Therefore, $\lambda(\hat{\mu}_C, a, \lim_{n\to\infty} d^*(\hat{\mu}_C^n, a)) \leq 0$ implies $\lambda(\hat{\mu}_C, a, \hat{d}(a)) < 0$ by Lemma 6-(2). Hence, $\underline{\lambda}(\hat{\mu}_C, a) < 0$ so that $\hat{\mu}_C \in M_C(a)$. Therefore, $d^*(\hat{\mu}_C, a)$ is defined. Then, $\lim_{n\to\infty} \hat{d}(a^n) = \lim_{n\to\infty} d^*(\hat{\mu}_C^n, a^n) = d^*(\hat{\mu}_C, a)$ because, as we noted above, $d^*(\tilde{\mu}_C, \tilde{a})$ is continuous in $(\tilde{\mu}_C, \tilde{a})$. This means that $\hat{d}(a) = \sup_{\mu_C \in M_C(a)} d^*(\mu_C, a) \geq d^*(\hat{\mu}_C, a) = \lim_{n\to\infty} \hat{d}(a^n) > \hat{d}(a)$. This is a contradiction.

[**Step 3**] We prove two claims presented in Step 1.

First, we prove Claim 1. Consider the behavior of $\bar{\mu}_C(a, d)$ with respect to d. Recall from Lemma 6-(1) that $\bar{\mu}_C(a, d)$ is strictly decreasing in $d \in (c, a)$. To find its limit as d goes to a, recall from the proof of Lemma 4 (Step 5) that $\bar{\mu}_C(a, d)$ is a unique

solution $\bar{\mu}_C$ to the equation (28), which is rearranged into

$$\bar{\mu}_C = \phi\Big(\bar{\mu}_C\boldsymbol{\beta} \geq \bar{\mu}_C\frac{b-d}{b-c} - \frac{a-d}{b-c}\Big).$$

This means that as d goes to a, $\bar{\mu}_C(a,d)$ approaches

$$\lim_{d\uparrow a}\bar{\mu}_C(a,d) = \phi\Big((\lim_{d\uparrow a}\bar{\mu}_C(a,d))\boldsymbol{\beta} \geq (\lim_{d\uparrow a}\bar{\mu}_C(a,d))\frac{b-a}{b-c} - \frac{a-a}{b-c}\Big) = \phi(\boldsymbol{\beta} \geq \beta^*).$$

Hence, if the given $\mu_C \in M_C(a)\setminus 0$ is $\phi(\boldsymbol{\beta}\geq\beta^*) < \mu_C < \bar{\bar{\mu}}_C(a)$, there exists some $d'' \in (c,a)$ such that $\mu_C = \bar{\mu}_C(a,d'')$ because $\bar{\mu}_C(a,d)$ is continuous in d by the proof of Lemma 4 (Step 5). Recall from the proof of Theorem 1 that $\lambda(\bar{\mu}_C(a,d''),a,d'') > 0$. Therefore, $\lambda(\mu_C,a,d'') > 0$.

Consider μ_C in the remaining case of $0 < \mu_C \leq \phi(\boldsymbol{\beta}\geq\beta^*)$. In this case, it holds for every $d\in(c,a)$ that $\mu_C \leq \bar{\mu}_C(a,d)$ so that $\lambda(\mu_C,a,d)$ is defined. Recall from Lemma 4 that $\hat{\mu}(\mu_C,a,d)$ lies in Δ' so that $\frac{a-d}{d-c}\hat{\mu}_{CDD}(\mu_C,a,d) \geq \hat{\mu}_{DDD}(\mu_C,a,d)$ holds for any $d\in(c,a)$. This means that $\lim_{d\uparrow a}\hat{\mu}_{DDD}(\mu_C,a,d) = 0$. On the other hand, note by Lemma 2 that

$$\begin{aligned}
&\psi_{DDD}(\hat{\mu}(\mu_C,a,d))\\
&= \phi(T^*_{DDD}(\hat{\mu}(\mu_C,a,d)))\\
&= \phi(T\cap\{(\alpha,\beta)|\beta \leq H(\alpha|\hat{\mu}(\mu_C,a,d)),\beta\leq\beta^*\})\\
&= \phi(T\cap\{(\alpha,\beta)|\beta \leq \frac{\hat{\mu}_{DDD}(\mu_C,a,d)}{\hat{\mu}_C(\mu_C,a,d)}\alpha + \beta^* + \frac{\hat{\mu}_{DDD}(\mu_C,a,d)(d-c) - \hat{\mu}_{CDD}(\mu_C,a,d)(a-d)}{\hat{\mu}_C(\mu_C,a,d)(b-c)},\beta\leq\beta^*\})\\
&= \phi(T\cap\{(\alpha,\beta)|\beta \leq \frac{\hat{\mu}_{DDD}(\mu_C,a,d)}{\mu_C}\alpha + \beta^* + \frac{\hat{\mu}_{DDD}(\mu_C,a,d)(d-c) - \hat{\mu}_{CDD}(\mu_C,a,d)(a-d)}{\mu_C(b-c)},\beta\leq\beta^*\})
\end{aligned}$$

where the last equality follows from $\mu_C = \hat{\mu}_C(\mu_C,a,d)$. This means that $\lim_{d\uparrow a}\psi_{DDD}(\hat{\mu}(\mu_C,a,d)) = \phi(\boldsymbol{\beta}\leq\beta^*)$, because $\lim_{d\uparrow a}\hat{\mu}_{DDD}(\mu_C,a,d) = 0$ guarantees that

$$\lim_{d\uparrow a}\Big[\frac{\hat{\mu}_{DDD}(\mu_C,a,d)}{\mu_C}\alpha + \beta^* + \frac{\hat{\mu}_{DDD}(\mu_C,a,d)(d-c) - \hat{\mu}_{CDD}(\mu_C,a,d)(a-d)}{\mu_C(b-c)}\Big] = \beta^*.$$

Hence

$$\lim_{d\uparrow a}\lambda(\mu_C,a,d) = \lim_{d\uparrow a}\psi_{DDD}(\hat{\mu}(\mu_C,a,d)) - \lim_{d\uparrow a}\hat{\mu}_{DDD}(\mu_C,a,d) = \phi(\boldsymbol{\beta}\leq\beta^*) > 0.$$

Therefore, there exists d'' close enough to a such that $\lambda(\mu_C,a,d'') > 0$.

Second, we prove Claim 2. Recall from Step 2 that we noticed that $d^*(0,a) = \bar{d}(a)$ and we showed that $\hat{d}(a) = \sup_{\mu_C\in M_C(a)} d^*(\mu_C,a) \geq d^*(0,a) = \bar{d}(a)$. If $\sup_{\mu_C\in M_C(a)} d^*(\mu_C,a) = \bar{d}(a)$, then $\hat{\mu}_C = 0$ achieves $\max_{\mu_C\in M_C(a)} d^*(\mu_C,a)$ and $\hat{d}(a) = \bar{d}(a)$.

Consider the other case of $\sup_{\mu_C\in M_C(a)} d^*(\mu_C,a) > \bar{d}(a)$. Then, there must exist $\{\mu_C^n\}_{n=1}^\infty$ such that $\mu_C^n \in M_C(a)$ and $\lim_{n\to\infty} d^*(\mu_C^n,a) = \sup_{\mu_C\in I} d^*(\mu_C,a)$. The sequence contains a convergent subsequence because each μ_C^n is a probability and contained in a compact set $[0,1]$. Without loss of generality, we assume that $\{\mu_C^n\}_{n=1}^\infty$ itself is convergent, and we denote $\hat{\mu}_C = \lim_{n\to\infty}\mu_C^n$.

Note that $\lim_{n\to\infty} d^*(\mu_C^n,a) > \bar{d}(a)$ implies that there exists N such that $d^*(\mu_C^n,a) > \bar{d}(a)$ for any $n\geq N$. Therefore, $[0,\bar{\mu}_C(a,d^*(\mu_C^n,a))] \subset [0,\bar{\mu}_C(a,\bar{d}(a))]$ for any $n\geq N$ by Lemma 6-(1). This guarantees that $\hat{\mu}_C \in [0,\bar{\mu}_C(a,\bar{d}(a))]$ because $\mu_C^n \in [0,\bar{\mu}_C(a,d^*(\mu_C^n,a))]$ for each n. This means that $\lambda(\hat{\mu}_C,a,\bar{d}(a))$ is defined. Therefore, $\underline{\lambda}(\hat{\mu}_C,a)$ is defined.

Now, we show that $\underline{\lambda}(\hat{\mu}_C, a) < 0$; that is, $\hat{\mu}_C \in M_C(a)$ so that $\hat{\mu}_C$ achieves $\max_{\mu_C \in M_C(a)} d^*(\mu_C, a)$. Suppose to the contrary that $\underline{\lambda}(\hat{\mu}_C, a) \geq 0$. Then, it must be the case that $\lambda(\hat{\mu}_C, a, d) > 0$ for every $d \in (c, a)$ for which $\lambda(\hat{\mu}_C, a, d)$ is defined. To see this, suppose that $\lambda(\hat{\mu}_C, a, d)$ is defined and $\lambda(\hat{\mu}_C, a, d) \leq 0$ for some $d \in (c, a)$. Then, $\hat{\mu}_C \in [0, \bar{\mu}_C(a, d)]$ implies $\hat{\mu}_C \in [0, \bar{\mu}_C(a, \frac{1}{2}c + \frac{1}{2}d)]$ by Lemma 6-(1) so that $\lambda(\hat{\mu}_C, a, \frac{1}{2}c + \frac{1}{2}d)$ is defined, and $\lambda(\hat{\mu}_C, a, d) \leq 0$ implies $\lambda(\hat{\mu}_C, a, \frac{1}{2}c + \frac{1}{2}d) < 0$ by Lemma 6-(2). This means $\underline{\lambda}(\hat{\mu}_C, a) < 0$, a contradiction. Hence, $\lambda(\hat{\mu}_C, a, d) > 0$ for every $d \in (c, a)$ for which $\lambda(\hat{\mu}_C, a, d)$ is defined. Take $d = \bar{d}(a)$, for which we showed above that $\lambda(\hat{\mu}_C, a, \bar{d}(a))$ is defined. Then, $\lambda(\hat{\mu}_C, a, \bar{d}(a)) > 0$. Then, there exists N' such that $\lambda(\mu_C^{N'}, a, d^*(\mu_C^{N'}, a)) > \lambda(\mu_C^{N'}, a, \bar{d}(a)) > \frac{1}{2}\lambda(\hat{\mu}_C, a, \bar{d}(a)) > 0$ because $\lim_{n\to\infty} d^*(\mu_C^n, a) > \bar{d}(a)$, $\hat{\mu}_C = \lim_{n\to\infty} \mu_C^n$, and $\lambda(\tilde{\mu}_C, a, \tilde{d})$ is continuous in $\tilde{\mu}_C$ and strictly increasing in $\tilde{d}$. This contradicts the condition that $\lambda(\mu_C^{N'}, a, d^*(\mu_C^{N'}, a)) = 0$. Hence, $\underline{\lambda}(\hat{\mu}_C, a) < 0$ must hold. (Q.E.D.)

Proof of Lemma 7

Lemma 7-(2) is immediate because $\beta^* = \frac{b-a}{b-c}$. As for Lemma 7-(1), we show that $\alpha^*(\mu^{\min}(a, d))$ is strictly increasing in a for a range of (a, d) with $c < d < \bar{d}(a)$. The remaining part of Lemma 7-(1) is proved similarly.

Fix $d \in (c, b)$ and consider a', a'' such that $a' < a''$. Suppose that (a', d) lies in a range with $c < d < \bar{d}(a')$. Then, by Corollary 1, (a'', d) also lies in a range with $c < d < \bar{d}(a') < \bar{d}(a'')$. Therefore, we have $\mu_C^{\min}(a', d) < \mu_C^{\min}(a'', d)$ by Theorem 3. We show that $\alpha^*(\mu^{\min}(a', d)) < \alpha^*(\mu^{\min}(a'', d))$. Suppose to the contrary that $\alpha^*(\mu^{\min}(a', d)) \geq \alpha^*(\mu^{\min}(a'', d))$. Note by Lemma 7-(2) that $\beta^*(a') = \frac{b-a'}{b-c} > \frac{b-a''}{b-c} = \beta^*(a'')$ when $a' < a''$. Then, together with the fact that $\mu_{CDD}^{\min}(a', d) = \psi_{CDD}(\mu^{\min}(a', d), a', d)$ and $\mu_{CDD}^{\min}(a'', d) = \psi_{CDD}(\mu^{\min}(a'', d), a'', d)$ in the equilibria, $\alpha^*(\mu^{\min}(a', d)) \geq \alpha^*(\mu^{\min}(a'', d))$ and $\beta^*(a') > \beta^*(a'')$ imply that

$$\begin{aligned}
\mu_{CDD}^{\min}(a', d) &= \psi_{CDD}(\mu^{\min}(a', d), a', d) \\
&= \phi(T_{CDD}^*(\mu^{\min}(a', d), a', d)) \\
&= \phi(T \cap \{(\alpha, \beta) | \alpha > \alpha^*(\mu^{\min}(a', d)), \beta > \beta^*(a')\}) \\
&< \phi(T \cap \{(\alpha, \beta) | \alpha > \alpha^*(\mu^{\min}(a'', d)), \beta > \beta^*(a'')\}) \\
&= \phi(T_{CDD}^*(\mu^{\min}(a'', d), a'', d)) \\
&= \psi_{CDD}(\mu^{\min}(a'', d), a'', d) \\
&= \mu_{CDD}^{\min}(a'', d).
\end{aligned}$$

Then, $\mu_C^{\min}(a', d) < \mu_C^{\min}(a'', d)$ and $\mu_{CDD}^{\min}(a', d) < \mu_{CDD}^{\min}(a'', d)$ imply that

$$\mu_{DDD}^{\min}(a', d) = 1 - \mu_C^{\min}(a', d) - \mu_{CDD}^{\min}(a', d) > 1 - \mu_C^{\min}(a'', d) - \mu_{CDD}^{\min}(a'', d) = \mu_{DDD}^{\min}(a'', d).$$

Hence, $\frac{\mu_{CDD}^{\min}(a', d)}{\mu_{DDD}^{\min}(a', d)} < \frac{\mu_{CDD}^{\min}(a'', d)}{\mu_{DDD}^{\min}(a'', d)}$. Then, it must hold for the expression (4) of α^* that

$$\alpha^*(\mu^{\min}(a', d)) = \frac{a' - d}{b - c} \frac{\mu_{CDD}^{\min}(a', d)}{\mu_{DDD}^{\min}(a', d)} - \frac{d - c}{b - c} < \frac{a'' - d}{b - c} \frac{\mu_{CDD}^{\min}(a'', d)}{\mu_{DDD}^{\min}(a'', d)} - \frac{d - c}{b - c} = \alpha^*(\mu^{\min}(a'', d))$$

when $a' < a''$. This contradicts the supposition $\alpha^*(\mu^{\min}(a', d)) \geq \alpha^*(\mu^{\min}(a'', d))$. (Q.E.D.)

Proof of Theorem 5

We show that $\mu_{DDD}^{\min}(a, d)$ is strictly decreasing in a for a range of (a, d) with $\alpha^*(\mu^{\min}(a, d)) \leq \beta^*$ and $c < d < \bar{d}(a)$. The remaining part of Theorem 5 is proved similarly.

Fix $d \in (c, b)$ and consider a', a'' such that $a' < a''$. Suppose that (a', d) and (a'', d) lie in a range with $c < d < \bar{d}(a') < \bar{d}(a'')$ and $\alpha^*(\mu^{\min}(a', d)) < \alpha^*(\mu^{\min}(a'', d)) \leq \beta^*(a'') < \beta^*(a')$ where Corollary 1 guarantees $\bar{d}(a') < \bar{d}(a'')$ and Lemma 7 guarantees $\alpha^*(\mu^{\min}(a', d)) < \alpha^*(\mu^{\min}(a'', d))$ and $\beta^*(a'') < \beta^*(a')$. Note by Lemma 2 that $\alpha^*(\mu^{\min}(a', d)) < \beta^*(a')$ and $\alpha^*(\mu^{\min}(a'', d)) \leq \beta^*(a'')$ means that

$$\{(\alpha, \beta) \in T | \beta < \beta^*(a')\} = T^*_C(\mu^{\min}(a', d), a', d) \cup T^*_{DDD}(\mu^{\min}(a', d), a', d)$$
$$\{(\alpha, \beta) \in T | \beta < \beta^*(a'')\} = T^*_C(\mu^{\min}(a'', d), a'', d) \cup T^*_{DDD}(\mu^{\min}(a'', d), a'', d).$$

Here, $\beta^*(a'') < \beta^*(a')$ implies that

$$\phi(\{(\alpha, \beta) \in T | \beta < \beta^*(a')\}) > \phi(\{(\alpha, \beta) \in T | \beta < \beta^*(a'')\})$$

so that

$$\begin{aligned}&\phi(T^*_C(\mu^{\min}(a', d), a', d)) + \phi(T^*_{DDD}(\mu^{\min}(a', d), a', d))\\ &> \phi(T^*_C(\mu^{\min}(a'', d), a'', d)) + \phi(T^*_{DDD}(\mu^{\min}(a'', d), a'', d)).\end{aligned}$$

In a range with $c < d < \bar{d}(a') < \bar{d}(a'')$, Theorem 3 applies and we have

$$\phi(T^*_C(\mu^{\min}(a', d), a', d)) = \mu^{\min}_C(a', d) < \mu^{\min}_C(a'', d) = \phi(T^*_C(\mu^{\min}(a'', d), a'', d)).$$

Therefore,

$$\begin{aligned}\mu^{\min}_{DDD}(a', d) - \mu^{\min}_{DDD}(a'', d) &= \phi(T^*_{DDD}(\mu^{\min}(a', d), a', d)) - \phi(T^*_{DDD}(\mu^{\min}(a'', d), a'', d))\\ &> \phi(T^*_C(\mu^{\min}(a'', d), a'', d)) - \phi(T^*_C(\mu^{\min}(a', d), a', d))\\ &> 0.\end{aligned}$$

(Q.E.D.)

Proof of Lemma 8

The condition (21) that a type (α, β) prefers C to CDD and DDD under a belief μ is rewritten as

$$\frac{\mu_{CDD}}{\mu_C}(a - d) > \gamma(d - (c - \alpha(b - c))) + (\max[a, b - \beta(b - c)] - a).$$

Consider $(\alpha, \beta), (\alpha', \beta')$ such that

$$\gamma(d-(c-\alpha(b-c)))+(\max[a, b-\beta(b-c)]-a) < \gamma(d-(c-\alpha'(b-c)))+(\max[a, b-\beta'(b-c)]-a).$$

Then, if

$$\gamma(d - (c - \alpha'(b - c))) + (\max[a, b - \beta'(b - c)] - a) < \frac{\mu_{CDD}}{\mu_C}(a - d),$$

then

$$\gamma(d - (c - \alpha(b - c))) + (\max[a, b - \beta(b - c)] - a) < \frac{\mu_{CDD}}{\mu_C}(a - d).$$

Furthermore, we can take $\frac{\mu_{CDD}}{\mu_C}$ such that

$$\gamma(d-(c-\alpha(b-c)))+(\max[a, b-\beta(b-c)]-a) < \frac{\mu_{CDD}}{\mu_C}(a-d) < \gamma(d-(c-\alpha'(b-c)))+(\max[a, b-\beta'(b-c)]-a).$$

because we can choose any positive value for $\frac{\mu_{CDD}}{\mu_C}$ given $\gamma = \frac{\mu_{DDD}}{\mu_C}$ when we compare the incentives to lead between (α, β) and (α', β'). These mean that (α, β) has a stronger

incentive to lead under the belief ratio γ than (α', β'). Therefore, a type (α, β) has the strongest incentive to lead under the belief ratio γ if and only if (α, β) is a solution to

$$\min_{(\tilde{\alpha},\tilde{\beta})\in T} \gamma(d-(c-\tilde{\alpha}(b-c)))+(\max[a, b-\tilde{\beta}(b-c)]-a).$$

Consider the minimum of $\gamma(d-(c-\tilde{\alpha}(b-c)))+(\max[a, b-\tilde{\beta}(b-c)]-a)$ in a subset of T with $\tilde{\beta} \geq \beta^*$. Note that $\max[a, b-\tilde{\beta}(b-c)]-a = 0$ if $\tilde{\beta} \geq \beta^*$ so that the objective for the minimization reduces to $\gamma(d-(c-\tilde{\alpha}(b-c)))$. The minimum over the set of $(\tilde{\alpha}, \tilde{\beta})$ with $\tilde{\alpha} \geq \tilde{\beta} \geq \beta^*$ is attained at (β^*, β^*). This means that the minimum of $\gamma(d-(c-\tilde{\alpha}(b-c)))+(\max[a, b-\tilde{\beta}(b-c)]-a)$ over T is attained by (α, β) in a subset of T with $\tilde{\beta} \leq \beta^*$. Note that $\max[a, b-\tilde{\beta}(b-c)]-a = b-\tilde{\beta}(b-c)-a$ if $\tilde{\beta} \leq \beta^*$ so that the objective for the minimization reduces to

$$\gamma(d-(c-\tilde{\alpha}(b-c)))+(b-\tilde{\beta}(b-c)-a) = (b-c)(\gamma\tilde{\alpha}-\tilde{\beta})+\gamma(d-c)+(b-a).$$

The minimum of this objective over the set of $(\tilde{\alpha}, \tilde{\beta})$ with $0 \leq \tilde{\beta} \leq \beta^*$ and $\tilde{\beta} \leq \tilde{\alpha}$ is attained by $(\alpha, \beta) = (0, 0)$ if $\gamma > 1$, (β^*, β^*) if $\gamma < 1$, and $\{(\alpha, \beta) = t(0, 0) + (1-t)(\beta^*, \beta^*) | 0 \leq t \leq 1\}$ if $\gamma = 1$. (Q.E.D.)

Proof of Lemma 9

Lemmas 9-(1) and (2) are restatements of Lemma 7. We prove Lemma 9-(3). In particular, we prove that $\gamma^*(\mu^{\min}(a, d))$ is strictly decreasing in a for a range of (a, d) with $c < d < \bar{d}(a)$. The remaining parts of Lemma 9-(3) are proved similarly.

Fix $d \in (c, b)$ and consider a', a'' such that $a' < a''$. Suppose that (a', d) and (a'', d) lie in a range with $c < d < \bar{d}(a') < \bar{d}(a'')$ where Corollary 1 guarantees $\bar{d}(a') < \bar{d}(a'')$. We show that $\gamma^*(\mu^{\min}(a', d)) = \frac{\mu^{\min}_{DDD}(a', d)}{\mu^{\min}_{C}(a', d)} > \frac{\mu^{\min}_{DDD}(a'', d)}{\mu^{\min}_{C}(a'', d)} = \gamma^*(\mu^{\min}(a'', d))$. Suppose to the contrary that $\frac{\mu^{\min}_{DDD}(a', d)}{\mu^{\min}_{C}(a', d)} \leq \frac{\mu^{\min}_{DDD}(a'', d)}{\mu^{\min}_{C}(a'', d)}$. Note by Theorem 3 that $\mu^{\min}_{C}(a', d) < \mu^{\min}_{C}(a'', d)$ when $a' < a''$ for a range with $c < d < \bar{d}(a') < \bar{d}(a'')$. Then, $\frac{\mu^{\min}_{DDD}(a', d)}{\mu^{\min}_{C}(a', d)} \leq \frac{\mu^{\min}_{DDD}(a'', d)}{\mu^{\min}_{C}(a'', d)}$ implies that $\mu^{\min}_{DDD}(a', d) < \mu^{\min}_{DDD}(a'', d)$. On the other hand, note by Lemma 2 that

$$\mu^{\min}_{DDD}(a', d) = \phi(T \cap \{(\alpha, \beta) | \beta \leq H(\alpha | \mu^{\min}(a', d), a', d), \beta \leq \beta^*(a')\})$$
$$\mu^{\min}_{DDD}(a'', d) = \phi(T \cap \{(\alpha, \beta) | \beta \leq H(\alpha | \mu^{\min}(a'', d), a'', d), \beta \leq \beta^*(a'')\})$$

where

$$H(\alpha | \mu^{\min}(a', d), a', d) = \frac{\mu^{\min}_{DDD}(a', d)}{\mu^{\min}_{C}(a', d)}(\alpha - \alpha^*(\mu^{\min}(a', d), a', d)) + \beta^*(a')$$
$$H(\alpha | \mu^{\min}(a'', d), a'', d) = \frac{\mu^{\min}_{DDD}(a'', d)}{\mu^{\min}_{C}(a'', d)}(\alpha - \alpha^*(\mu^{\min}(a'', d), a'', d)) + \beta^*(a'')$$

and $\beta^*(a') = \frac{b-a'}{b-c}$, $\beta^*(a'') = \frac{b-a''}{b-c}$. By Lemma 9-(1) and (2), $\alpha^*(\mu^{\min}(a', d), a', d) < \alpha^*(\mu^{\min}(a'', d), a'', d)$ and $\beta^*(a') > \beta^*(a'')$ when $a' < a''$ for a range with $c < d < \bar{d}(a') < \bar{d}(a'')$. Then, together with $\frac{\mu^{\min}_{DDD}(a', d)}{\mu^{\min}_{C}(a', d)} \leq \frac{\mu^{\min}_{DDD}(a'', d)}{\mu^{\min}_{C}(a'', d)}$, we must have

$$\{(\alpha, \beta) | \beta \leq H(\alpha | \mu^{\min}(a', d), a', d), \beta \leq \beta^*(a')\} \supsetneq \{(\alpha, \beta) | \beta \leq H(\alpha | \mu^{\min}(a'', d), a'', d), \beta \leq \beta^*(a'')\}$$

so that $\mu^{\min}_{DDD}(a', d) > \mu^{\min}_{DDD}(a'', d)$. This is a contradiction. (Q.E.D.)

Proof of Theorem 7

We prove Theorem 7-(2). Theorem 7-(1) is proved similarly.

Consider (a,d) and (a',d') such that $a > a'$, $d < d'$, and $c < d' < \bar{d}(a')$. Then, (a,d) also lies in a range with $c < d < \bar{d}(a)$ because $\bar{d}(a') < \bar{d}(a)$ by Corollary 1 when $a > a'$. Therefore, both $PD((a,b,c,d),f)$ and $PD((a',b,c,d'),f)$ admit a three-mode equilibrium. Recall from Lemma 2 that

$$T_C^*(\mu^{\min}(a,d),a,d) = \{(\alpha,\beta)|\beta \geq H(\alpha|\mu^{\min}(a,d),a,d), \alpha \leq \alpha^*(\mu^{\min}(a,d),a,d)\}$$
$$T_C^*(\mu^{\min}(a',d'),a',d') = \{(\alpha,\beta)|\beta \geq H(\alpha|\mu^{\min}(a',d'),a',d'), \alpha \leq \alpha^*(\mu^{\min}(a,d),a,d)\}$$

where

$$H(\alpha|\mu^{\min}(a,d),a,d) = \frac{\mu_{DDD}^{\min}(a,d)}{\mu_C^{\min}(a,d)}(\alpha - \alpha^*(\mu^{\min}(a,d),a,d)) + \beta^*(a)$$
$$H(\alpha|\mu^{\min}(a'',d),a'',d) = \frac{\mu_{DDD}^{\min}(a',d')}{\mu_C^{\min}(a',d')}(\alpha - \alpha^*(\mu^{\min}(a',d'),a',d')) + \beta^*(a').$$

We show that $T_C^*(\mu^{\min}(a,d),a,d)$ is a leadership pattern of a more equity concerned leader than $T_C^*(\mu^{\min}(a',d'),a',d')$.

To compare $T_C^*(\mu^{\min}(a,d),a,d)$ with $T_C^*(\mu^{\min}(a',d'),a',d')$, consider two lines $\beta = H(\alpha|\mu^{\min}(a,d),a,d)$ and $\beta = H(\alpha|\mu^{\min}(a',d'),a',d')$, which define the boundaries of $T_C^*(\mu^{\min}(a,d),a,d)$ and $T_C^*(\mu^{\min}(a',d'),a',d')$ respectively. Given that $c < d < \bar{d}(a)$ and $c < d' < \bar{d}(a')$, Lemma 9 applies and we have $\alpha^*(\mu^{\min}(a,d),a,d) > \alpha^*(\mu^{\min}(a',d'),a',d')$, $\beta^*(a) < \beta^*(a')$, and $\gamma^*(\mu^{\min}(a,d)) = \frac{\mu_{DDD}^{\min}(a,d)}{\mu_C^{\min}(a,d)} < \frac{\mu_{DDD}^{\min}(a',d')}{\mu_C^{\min}(a',d')} = \gamma^*(\mu^{\min}(a',d'))$. Then, $\gamma^*(\mu^{\min}(a,d)) < \gamma^*(\mu^{\min}(a',d'))$ implies that there exists a unique $(\tilde{\alpha},\tilde{\beta})$ in R^2 that satisfies $\tilde{\beta} = H(\tilde{\alpha}|\mu^{\min}(a,d),a,d)$ and $\tilde{\beta} = H(\tilde{\alpha}|\mu^{\min}(a'',d),a'',d)$ simultaneously. Furthermore, $\alpha^*(\mu^{\min}(a,d),a,d) > \alpha^*(\mu^{\min}(a',d'),a',d')$ and $\beta^*(a) < \beta^*(a')$ imply that $(\alpha^*(\mu^{\min}(a,d),a,d),\beta^*(a))$ is located south east of $(\alpha^*(\mu^{\min}(a',d'),a',d'),\beta^*(a'))$. This in turn implies that $\tilde{\alpha} < \alpha^*(\mu^{\min}(a,d),a,d)$, $\tilde{\beta} < \beta^*(a)$, $\tilde{\alpha} < \alpha^*(\mu^{\min}(a',d'),a',d')$, and $\tilde{\beta} < \beta^*(a')$.

The three types $(\tilde{\alpha},\tilde{\beta})$, $(\alpha^*(\mu^{\min}(a,d),a,d),\beta^*(a))$, and $(\alpha^*(\mu^{\min}(a',d'),a',d'),\beta^*(a'))$ identified above may or may not be included in the type space T. One of the possibilities is a case in which all of them are in T. Figure 15 demonstrates this case. Then, $(\alpha^*(\mu^{\min}(a,d),a,d),\beta^*(a)) \in T$ means that when we define $(\hat{\alpha},\hat{\beta})$ by $\hat{\alpha} = \alpha^*(\mu^{\min}(a,d),a,d)$ and $\hat{\beta} = \max\{\beta|(\alpha^*(\mu^{\min}(a,d),a,d),\beta) \in T\}$, it holds that $(\hat{\alpha},\hat{\beta}) \in T_C^*(\mu^{\min}(a,d),a,d)$ and that $\alpha \leq \hat{\alpha}$ and $\beta \leq \hat{\beta}$ for any $(\alpha,\beta) \in T_C^*(\mu^{\min}(a,d),a,d)$; that is, the type $(\hat{\alpha},\hat{\beta})$ is the most inequity concerned in both envy and guilt among all the leader types in $T_C^*(\mu^{\min}(a,d),a,d)$. Similarly, $(\hat{\alpha}',\hat{\beta}')$ such that $\hat{\alpha}' = \alpha^*(\mu^{\min}(a',d'),a',d')$ and $\hat{\beta}' = \max\{\beta'|(\alpha^*(\mu^{\min}(a',d'),a',d'),\beta') \in T\}$ is the most inequity concerned in both envy and guilt among all the leader types in $T_C^*(\mu^{\min}(a',d'),a',d')$. Furthermore, $(\hat{\alpha}',\hat{\beta}') \leq (\hat{\alpha},\hat{\beta})$ holds by $\alpha^*(\mu^{\min}(a,d),a,d) > \alpha^*(\mu^{\min}(a',d'),a',d')$. Hence, the condition (1) for $T_C^*(\mu^{\min}(a,d),a,d)$ to be a leadership pattern of more equity concerned leader than $T_C^*(\mu^{\min}(a',d'),a',d')$ is satisfied.

Note also in this case that $T_C^*(\mu^{\min}(a',d'),a',d') \setminus T_C^*(\mu^{\min}(a,d),a,d) \neq \emptyset$ and $T_C^*(\mu^{\min}(a,d),a,d) \setminus T_C^*(\mu^{\min}(a',d'),a',d') \neq \emptyset$. Take $(\alpha',\beta') \in T_C^*(\mu^{\min}(a',d'),a',d') \setminus T_C^*(\mu^{\min}(a,d),a,d)$ and $(\alpha,\beta) \in T_C^*(\mu^{\min}(a,d),a,d) \setminus T_C^*(\mu^{\min}(a',d'),a',d')$ arbitrarily. Then, it must hold for (α',β') that $\beta' \geq H(\alpha'|\mu^{\min}(a',d'),a',d')$ and $\beta' < H(\alpha'|\mu^{\min}(a,d),a,d)$. This means that $\alpha' < \tilde{\alpha}$ and $\beta' < \tilde{\beta}$. On the other hand, it must hold for (α,β) that $\beta \geq H(\alpha|\mu^{\min}(a,d),a,d)$ and $\beta < H(\alpha|\mu^{\min}(a',d'),a',d')$. This means that $\tilde{\alpha} < \alpha$ and $\tilde{\beta} < \beta$. Hence, we have $(\alpha',\beta') \leq (\tilde{\alpha},\tilde{\beta}) \leq (\alpha,\beta)$. Thus, the condition (2) for $T_C^*(\mu^{\min}(a,d),a,d)$ to be a leadership pattern of more equity concerned leader than $T_C^*(\mu^{\min}(a',d'),a',d')$ is also satisfied.

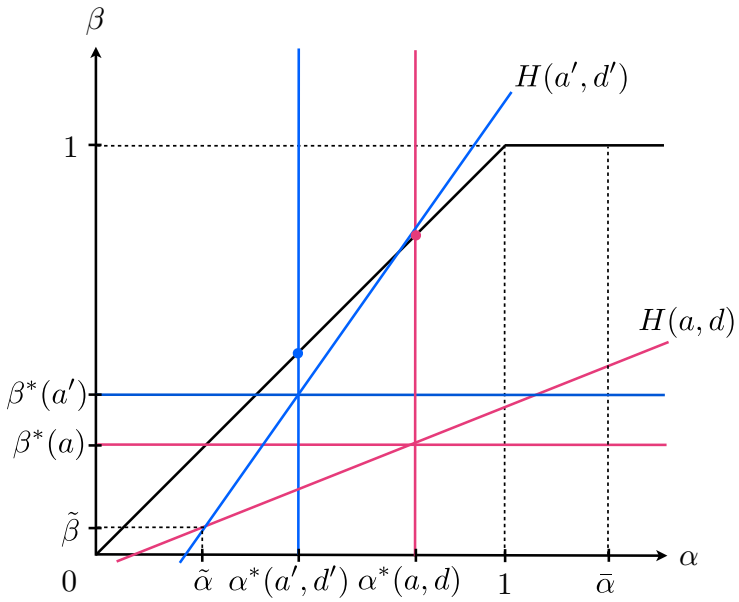


Figure 15: ordered-leader-set

Similarly, we can show that $T_C^*(\mu^{\min}(a,d),a,d)$ is a leadership pattern of a more equity concerned leader than $T_C^*(\mu^{\min}(a',d'),a',d')$ in all the remaining cases of relations of the three types $(\hat{\alpha},\hat{\beta})$, $(\alpha^*(\mu^{\min}(a,d),a,d),\beta^*(a))$, and $(\alpha^*(\mu^{\min}(a',d'),a',d'),\beta^*(a'))$ to the type set T. (Q.E.D.)

Proof of Theorem 8

We show Theorem 8-(2). Theorem 8-(1) is proved similarly. Fix $d \in (c,b)$. Note that there exists a three-mode equilibrium with the minimum leadership distribution $\mu^{\min}(a,d)$ for any $a \in (\bar{d}^{-1}(d),b)$ because $c < d = \bar{d}(\bar{d}^{-1}(d)) < \bar{d}(a)$ by Corollary 1.

Suppose that $\gamma^*(\mu^{\min}(\tilde{a},d)) < 1$ for any $a \in (\bar{d}^{-1}(d),b)$. Then, the type of the strongest incentive to lead in the three-mode equilibrium distribution $\mu^{\min}(a,d)$ is the (β^*,β^*)-type for any $a \in (\bar{d}^{-1}(d),b)$ by Lemma 8. Hence, Theorem 8-(2) holds for $a_L^{\min}(d) = \bar{d}^{-1}(d)$.

Suppose that $\gamma^*(\mu^{\min}(a',d)) \geq 1$ for some $a' \in (\bar{d}^{-1}(d),b)$. Then, $\gamma^*(\mu^{\min}(\frac{1}{2}\bar{d}^{-1}(d)+\frac{1}{2}a',d)) > 1$ by Lemma 9. Examine $\gamma^*(\mu^{\min}(a,d)) = \frac{\mu_{DDD}^{\min}(a,d)}{\mu_C^{\min}(a,d)}$ for $a \in (a',b)$. Theorem 3 guarantees that $\mu_C^{\min}(a,d) > \mu_C^{\min}(a',d)$ for any $a \in (a',b)$. On the other hand, $T_{DDD}^*(\mu^{\min}(a,d),a,d) \subseteq \{(\tilde{\alpha},\tilde{\beta})|\tilde{\beta} \leq \beta^*(a)\}$ by Lemma 2. Note that $\lim_{a\uparrow b}\beta^*(a) = \lim_{a\uparrow b}\frac{b-a}{b-c} = 0$. Hence, $\lim_{a\uparrow b}\mu_{DDD}^{\min}(a,d) = \lim_{a\uparrow b}\phi(T_{DDD}^*(\mu^{\min}(a,d),a,d)) \leq \lim_{a\uparrow b}\phi(\{(\tilde{\alpha},\tilde{\beta})|\tilde{\beta} \leq \beta^*(a)\}) = 0$. This means that there exists $a \in (a',b)$ such that $\mu_{DDD}^{\min}(a,d) < \mu_C^{\min}(a',d)$. Then, $\gamma^*(\mu^{\min}(a,d)) = \frac{\mu_{DDD}^{\min}(a,d)}{\mu_C^{\min}(a,d)} < \frac{\mu_C^{\min}(a',d)}{\mu_C^{\min}(a',d)} = 1$. Thus, we have a',a with $\bar{d}^{-1}(d) < a' < a < b$, for which $\gamma^*(\mu^{\min}(\frac{1}{2}\bar{d}^{-1}(d)+\frac{1}{2}a',d)) > 1 > \gamma^*(\mu^{\min}(a,d))$. Then, when we set $a_L^{\min}(d) \equiv \sup\{a \in (\bar{d}^{-1}(d),b)|\gamma^*(\mu^{\min}(a,d)) \geq 1\}$, it immediately follows that $\bar{d}^{-1}(d) < a_L^{\min}(d) < b$, and it holds by Lemma 9 that $\gamma^*(\mu^{\min}(\tilde{a},d)) > 1$ for $\tilde{a} \in (\bar{d}^{-1}(d),a_L^{\min}(d))$, while $\gamma^*(\mu^{\min}(\tilde{a},d)) < 1$ for $\tilde{a} \in (a_L^{\min}(d),b)$. This means that the type of the strongest incentive to lead in the three-mode equilibrium distribution $\mu^{\min}(a,d)$ is Materialist if $a < a_L^{\min}(d)$ and the type (β^*,β^*) if $a > a_L^{\min}(d)$.

We show Theorem 8-(3). We show that $a_L^{\min}(d)$ is increasing. Fix d',d'' such that $c < d' < d'' < b$. We show that $a_L^{\min}(d') \leq a_L^{\min}(d'')$. Suppose to the contrary that $a_L^{\min}(d') > a_L^{\min}(d'')$. Recall that we set $a_L^{\min}(d') = \sup\{a \in (\bar{d}^{-1}(d'),b)|\gamma^*(\mu^{\min}(a,d')) \geq$

1} if the set on the right-hand side is not empty, and we set $a_L^{\min}(d') = \bar{d}^{-1}(d')$ otherwise. $a_L^{\min}(d'')$ is parallel. Then, the supposition $a_L^{\min}(d') > a_L^{\min}(d'')$ implies that in the definition of $a_L^{\min}(d')$, even $\{a \in (\bar{d}^{-1}(d''), b) | \gamma^*(\mu^{\min}(a, d')) \geq 1\}$ is not empty and $a_L^{\min}(d') = \sup\{a \in (\bar{d}^{-1}(d''), b) | \gamma^*(\mu^{\min}(a, d')) \geq 1\}$ because $\bar{d}^{-1}(d') < \bar{d}^{-1}(d'')$ by Corollary 1. Note by Lemma 9 that $\gamma^*(\mu^{\min}(a, d')) < \gamma^*(\mu^{\min}(a, d''))$ for any $a \in (\bar{d}^{-1}(d''), b)$. Hence, $\{a \in (\bar{d}^{-1}(d''), b) | \gamma^*(\mu^{\min}(a, d'')) \geq 1\}$ is not empty and $\sup\{a \in (\bar{d}^{-1}(d''), b) | \gamma^*(\mu^{\min}(a, d')) \geq 1\} \leq \sup\{a \in (\bar{d}^{-1}(d''), b) | \gamma^*(\mu^{\min}(a, d'')) \geq 1\} = a_L^{\min}(d'')$. This is a contradiction.

Finally, we show that $a_L^{\max}(d) \leq a_L^{\min}(d)$ for any $d \in (c, b)$. Fix $d \in (c, b)$ and consider $a \in (\bar{d}^{-1}(d), b)$. We show that $\gamma^*(\mu^{\min}(a, d)) \geq \gamma^*(\mu^{\max}(a, d))$. Suppose to the contrary that $\gamma^*(\mu^{\min}(a, d)) < \gamma^*(\mu^{\max}(a, d))$. Then, it must hold that $\mu_{DDD}^{\min}(a, d) < \mu_{DDD}^{\max}(a, d)$ because $\gamma^*(\mu^{\min}(a, d)) = \frac{\mu_{DDD}^{\min}(a,d)}{\mu_C^{\min}(a,d)}$, $\gamma^*(\mu^{\max}(a, d)) = \frac{\mu_{DDD}^{\max}(a,d)}{\mu_C^{\max}(a,d)}$, and $\mu_C^{\min}(a, d) \leq \mu_C^{\max}(a, d)$. On the other hand, we claim that $\alpha^*(\mu_C^{\min}(a, d), a, d) < \alpha^*(\mu_C^{\max}(a, d), a, d)$ when $\gamma^*(\mu^{\min}(a, d)) < \gamma^*(\mu^{\max}(a, d))$. Suppose to the contrary that $\alpha^*(\mu_C^{\min}(a, d), a, d) \geq \alpha^*(\mu_C^{\max}(a, d), a, d)$. Then, it must follow by Lemma 2 that $\mu_{CDD}^{\min}(a, d) \leq \mu_{CDD}^{\max}(a, d)$. Then, it follows that

$$\mu_{DDD}^{\min}(a, d) = 1 - \mu_C^{\min}(a, d) - \mu_{CDD}^{\min}(a, d) \geq 1 - \mu_C^{\max}(a, d) - \mu_{CDD}^{\max}(a, d) = \mu_{DDD}^{\max}(a, d).$$

This contradicts $\mu_{DDD}^{\min}(a, d) < \mu_{DDD}^{\max}(a, d)$. Hence, we must have $\alpha^*(\mu_C^{\min}(a, d), a, d) < \alpha^*(\mu_C^{\max}(a, d), a, d)$. Note by Lemma 2 that

$$\mu_{DDD}^{\min}(a, d) = \phi(T \cap \{(\alpha, \beta) | \beta \leq H(\alpha | \mu^{\min}(a, d), a, d), \beta \leq \beta^*\})$$
$$\mu_{DDD}^{\max}(a, d) = \phi(T \cap \{(\alpha, \beta) | \beta \leq H(\alpha | \mu^{\max}(a, d), a, d), \beta \leq \beta^*\})$$

where

$$H(\alpha | \mu^{\min}(a, d), a, d) = \frac{\mu_{DDD}^{\min}(a, d)}{\mu_C^{\min}(a, d)}(\alpha - \alpha^*(\mu^{\min}(a, d), a, d)) + \beta^*$$
$$H(\alpha | \mu^{\max}(a, d), a, d) = \frac{\mu_{DDD}^{\max}(a, d)}{\mu_C^{\max}(a, d)}(\alpha - \alpha^*(\mu^{\max}(a, d), a, d)) + \beta^*.$$

Here, $\frac{\mu_{DDD}^{\min}(a,d)}{\mu_C^{\min}(a,d)} = \gamma^*(\mu^{\min}(a, d)) < \gamma^*(\mu^{\max}(a, d)) = \frac{\mu_{DDD}^{\max}(a,d)}{\mu_C^{\max}(a,d)}$ and $\alpha^*(\mu_C^{\min}(a, d), a, d) < \alpha^*(\mu_C^{\max}(a, d), a, d)$ imply that $\mu_{DDD}^{\min}(a, d) > \mu_{DDD}^{\max}(a, d)$. This contradicts $\mu_{DDD}^{\min}(a, d) < \mu_{DDD}^{\max}(a, d)$. Hence, it must be the case that $\gamma^*(\mu^{\min}(a, d)) \geq \gamma^*(\mu^{\max}(a, d))$. This means that $\gamma^*(\mu^{\max}(a, d)) < 1$ for any a with $a > a_L^{\min}(d)$ because $\gamma^*(\mu^{\min}(a, d)) < 1$ if $a > a_L^{\min}(d)$. Hence, $a_L^{\max}(d) \leq \sup\{a \in (\bar{d}^{-1}(d), b) | \gamma^*(\mu^{\max}(a, d)) \geq 1\} \leq a_L^{\min}(d)$. (Q.E.D.)

Proof of Theorem 9

We show Theorem 9-(1). First, $\frac{\partial}{\partial a}\phi(\boldsymbol{\beta} < \frac{b-a}{b-c} | \alpha^* \leq \alpha) < 0$ for any $\alpha^* \in [0, \bar{\alpha}]$. Therefore, $\frac{\partial}{\partial a} \min_{\alpha^* \in [0,\bar{\alpha}]} \phi(\boldsymbol{\beta} < \frac{b-a}{b-c} | \alpha^* \leq \alpha) < 0$. Hence, $(\bar{d}_{ml}(a))' = 1 - (b - c)\frac{\partial}{\partial a} \min_{\alpha^* \in [0,\bar{\alpha}]} \phi(\boldsymbol{\beta} < \frac{b-a}{b-c} | \alpha^* \leq \alpha) > 0$. Second, $\bar{d}_{ml}(a) < a$ because $\phi(\boldsymbol{\beta} < \beta^* | \alpha^* \leq \alpha)$ is continuous in α^* and $\phi(\boldsymbol{\beta} < \beta^* | \alpha^* \leq \alpha) > 0$ for each $\alpha \in [0, \bar{\alpha}]$ so that $\min_{\alpha^* \in [0,\bar{\alpha}]} \phi(\boldsymbol{\beta} < \beta^* | \alpha^* \leq \alpha) > 0$. Third, $\lim_{a \to c} \bar{d}_{ml}(a) = c - (b - c)$ because $\lim_{a \to c} \phi(\boldsymbol{\beta} < \beta^* | \alpha^* \leq \alpha) = \phi(\boldsymbol{\beta} < 1 | \alpha^* \leq \alpha) = 1$. Fourth, $\lim_{a \to b} \bar{d}_{ml}(a) = b$ because $\lim_{a \to b} \phi(\boldsymbol{\beta} < \beta^* | \alpha^* \leq \alpha) = \phi(\boldsymbol{\beta} < 0 | \alpha^* \leq \alpha) = 0$. Finally, there exists $\bar{a}_{ml} \in (c, b)$ such that $\bar{d}_{ml}(\bar{a}_{ml}) = c$ because $\lim_{a \to c} \bar{d}_{ml}(a) = c - (b - c) < c < \lim_{a \to b} \bar{d}_{ml}(a) = b$ and $\bar{d}_{ml}(a)$ is continuous and strictly increasing.

We show Theorem 9-(2). Fix (a,d) such that $a \in (c, \bar{a}'_{ml})$ and $\max(c, \bar{d}_{ml}(a)) < d < \hat{d}(a)$. Suppose to the contrary that there exists a three-mode equilibrium and it is either an inequity concerned leader pattern or a hybrid leader pattern. Then, the three-mode equilibrium distribution μ must satisfy

$$\frac{b-a}{b-c} = \beta^* \leq \alpha^*(\mu) = \frac{\mu_{CDD}}{\mu_{DDD}} \frac{a-d}{b-c} - \frac{d-c}{b-c}$$

where the first equality is the definition of β^* and the second equality is by (4). Hence, it must hold that $\frac{b-c}{a-c} \leq \frac{a-d}{a-c}(1 + \frac{\mu_{CDD}}{\mu_{DDD}})$. However, $\bar{d}_{ml}(a) < d$ implies

$$\frac{b-c}{a-c} > \frac{a-d}{a-c} \frac{1}{\phi(\boldsymbol{\beta} < \beta^* | \alpha^*(\mu) \leq \boldsymbol{\alpha})} \geq \frac{a-d}{a-c}\Big(1 + \frac{\mu_{CDD}}{\mu_{DDD}}\Big)$$

where the second inequality holds for the reason that, as Lemma 2 shows,

$$\frac{\mu_{CDD}}{\mu_{DDD}} \leq \frac{\phi(\boldsymbol{\beta} > \beta^* | \alpha^*(\mu) \leq \boldsymbol{\alpha})}{\phi(\boldsymbol{\beta} < \beta^* | \alpha^*(\mu) \leq \boldsymbol{\alpha})}$$

in a three-mode equilibrium. This is a contradiction. Therefore, there should be no three-mode equilibrium of either an inequity concerned leader pattern or a hybrid leader pattern. (Q.E.D.)

Proof of Theorem 10

[Step 1] Fix $a \in (c,b)$. Then, we show that under condition (23), there exists $\bar{d}_{icl}(a) \in (c, \bar{d}(a))$ such that $\beta^* \phi(\boldsymbol{\beta} > \beta^*) > \mu_{CDD}$ for any three-mode equilibrium distribution μ in any $PD((a,b,c,d),f)$ with $d \in (c, \bar{d}_{icl}(a))$.

(Proof)

Consider the supremum of $\frac{\mu_{CDD}}{\mu_{DDD}}$ when we consider all the three-mode equilibrium distributions μ in $PD((a,b,c,d),f)$ over $d \in (c, \bar{d}(a))$. Then, it is finite. (We postpone the proof of this fact to Step 3 below.) This fact implies that

$$\begin{aligned}
&\lim_{d \downarrow c} \Big[\mu_{CDD} - \phi\Big(\boldsymbol{\alpha} > \frac{\mu_{CDD}}{\mu_{DDD}}(1-\beta^*) \text{ and } \boldsymbol{\beta} > \beta^*\Big)\Big] \\
&= \lim_{d \downarrow c} \Big[\phi(\boldsymbol{\alpha} > \alpha^*(\mu) \text{ and } \boldsymbol{\beta} > \beta^*) - \phi\Big(\boldsymbol{\alpha} > \frac{\mu_{CDD}}{\mu_{DDD}}(1-\beta^*) \text{ and } \boldsymbol{\beta} > \beta^*\Big)\Big] \\
&= \lim_{d \downarrow c} \Big[\phi\Big(\boldsymbol{\alpha} > \frac{\mu_{CDD}}{\mu_{DDD}}(1-\beta^*) - \frac{d-c}{b-c}\Big(1 + \frac{\mu_{CDD}}{\mu_{DDD}}\Big) \text{ and } \boldsymbol{\beta} > \beta^*\Big) - \phi\Big(\boldsymbol{\alpha} > \frac{\mu_{CDD}}{\mu_{DDD}}(1-\beta^*) \text{ and } \boldsymbol{\beta} > \beta^*\Big)\Big] \\
&= 0
\end{aligned}$$

where the first equality follows from Lemma 2 and the second equality follows from the expression (4) of $\alpha^*(\mu)$. Then, there exists $\bar{d}_{icl}(a) \in (c, \bar{d}(a))$ such that

$$\mu_{CDD} - \phi\Big(\boldsymbol{\alpha} > \frac{\mu_{CDD}}{\mu_{DDD}}(1-\beta^*) \text{ and } \boldsymbol{\beta} > \beta^*\Big) < \beta^* \phi(\boldsymbol{\beta} > \beta^*) - \phi\Big(\boldsymbol{\alpha} > \frac{\beta^* \phi(\boldsymbol{\beta} > \beta^*)}{1 - \phi(\boldsymbol{\beta} > \beta^*)}(1-\beta^*) \text{ and } \boldsymbol{\beta} > \beta^*\Big)$$

for any $d \in (c, \bar{d}_{icl}(a))$, because the condition (23) implies that the right-hand side of the above inequality is a positive number. Then,

$$\mu_{CDD} - \phi\Big(\boldsymbol{\alpha} > \frac{\mu_{CDD}}{\mu_{DDD}}(1-\beta^*) \text{ and } \boldsymbol{\beta} > \beta^*\Big) < \beta^* \phi(\boldsymbol{\beta} > \beta^*) - \phi\Big(\boldsymbol{\alpha} > \frac{\beta^* \phi(\boldsymbol{\beta} > \beta^*)}{\mu_{CDD}}(1-\beta^*) \text{ and } \boldsymbol{\beta} > \beta^*\Big)$$

for any $d \in (c, \bar{d}_{icl}(a))$, because $\mu_{CDD} < 1 - \phi(\boldsymbol{\beta} > \beta^*)$ by Lemma 2. This inequality implies that $\beta^* \phi(\boldsymbol{\beta} > \beta^*) > \mu_{CDD}$, because

$$x - \phi\Big(\boldsymbol{\alpha} > \frac{x}{\mu_{DDD}}(1-\beta^*) \text{ and } \boldsymbol{\beta} > \beta^*\Big)$$

is strictly increasing in x.
[Step 2] Fix $a \in (c, b)$. Take a three-mode equilibrium distribution μ in $PD((a, b, c, d), f)$ with $d \in (c, \bar{d}_{icl}(a))$. Consider the type $(\alpha, \beta) = (0, 0)$. Then,

$$\begin{aligned} U_{(0,0)}(DDD, \mu) - U_{(0,0)}(C, \mu) &= (\mu_C b + \mu_{CDD} d + \mu_{DDD} d) - (\mu_C a + \mu_{CDD} a + \mu_{DDD} c) \\ &> \mu_C (b - a) - \mu_{CDD}(a - c) \\ &> (\phi(\boldsymbol{\beta} > \beta^*) - \mu_{CDD})(b - a) - \mu_{CDD}(a - c) \\ &= (b - c)(\beta^* \phi(\boldsymbol{\beta} > \beta^*) - \mu_{CDD}) \\ &> 0 \end{aligned}$$

where the first inequality holds by $d > c$, the second inequality holds because $\mu_C > \phi(\boldsymbol{\beta} > \beta^*) - \mu_{CDD}$ by Lemma 2, and the last inequality holds by Step 1. This means that C is not a best response for the type $(\alpha, \beta) = (0, 0)$ so that $(0, 0) \notin T^*_C(\mu)$. Hence, the three-mode equilibrium must be an inequity concerned leader pattern.
[Step 3] We prove the fact that the supremum of $\frac{\mu_{CDD}}{\mu_{DDD}}$ is finite. Suppose to the contrary that we can find a three-mode equilibrium with an arbitrarily large $\frac{\mu_{CDD}}{\mu_{DDD}}$. Then, there exists a sequence of $\{d^n\}_{n=1}^{\infty}$ in $(c, \bar{d}(a))$ and a corresponding sequence of $\{\mu^n\}_{n=1}^{\infty}$ such that μ^n is a three-mode equilibrium distribution in $PD((a, b, c, d^n), f)$ and $\lim_{n\to\infty} \mu^n_{DDD} = 0$. Note that $\lim_{n\to\infty} \mu^n_{DDD} = 0$ implies $\lim_{n\to\infty} \phi(\alpha^*(\mu^n, a, d^n) \leq \boldsymbol{\alpha}$ and $\boldsymbol{\beta} < \beta^*) = 0$ because $0 \leq \phi(\alpha^*(\mu^n, a, d^n) \leq \boldsymbol{\alpha}$ and $\boldsymbol{\beta} < \beta^*) < \phi(T^*_{DDD}(\mu^n, a, d^n)) = \mu^n_{DDD}$ by Lemma 2. Note also that $\lim_{n\to\infty} \phi(\alpha^*(\mu^n, a, d^n) \leq \boldsymbol{\alpha}$ and $\boldsymbol{\beta} < \beta^*) = 0$ implies that $\lim_{n\to\infty} \mu^n_{CDD} = \lim_{n\to\infty} \phi(\alpha^*(\mu^n, a, d^n) \leq \boldsymbol{\alpha}$ and $\boldsymbol{\beta} > \beta^*) = 0$. Hence, we must have $\lim_{n\to\infty} \mu^n_C = 1 - \lim_{n\to\infty} \mu^n_{CDD} - \lim_{n\to\infty} \mu^n_{DDD} = 1$. On the other hand, recall that $\mu^n_C < \bar{\mu}(a, d^n)$ where $\bar{\mu}(a, d^n)$ is a unique solution $\bar{\mu}_C$ to the equation (28) with $d = d^n$ in the proof of Lemma 4. Consider the equation (28) with $d = c$. Then, it is rewritten as

$$\bar{\mu}_C = \phi\Big(\frac{a - c}{b - c} \geq (1 - \boldsymbol{\beta})\bar{\mu}_C\Big). \tag{29}$$

It follows by a similar argument to the proof of Lemma 4 that a unique solution $\bar{\mu}_C(a, c)$ to this equation (29) is located as $\frac{a-c}{b-c} < \bar{\mu}_C(a, c) < 1$. Furthermore, $\bar{\mu}_C(a, d^n) < \bar{\mu}_C(a, c)$ follows from $c < d^n$ by a similar argument to the proof of Lemma 6-(1). Hence, $\mu^n_C < \bar{\mu}_C(a, d^n) < \bar{\mu}_C(a, c) < 1$ for any n. This contradicts $\lim_{n\to\infty} \mu^n_C = 1$. Hence, it is established that $\sup \frac{\mu_{CDD}}{\mu_{DDD}}$ is finite. (Q.E.D.)

Proof of Theorem 11
Consider a three-mode strategy $\boldsymbol{s}$ in a prisoner's dilemma with two timings and consider the corresponding two-timing three-mode strategy in a prisoner's dilemma with K timings via the procedure $T_{C_{k^*}} = T_C(\boldsymbol{s})$, $T_{CDD_{k^*+1}} = T_{CDD}(\boldsymbol{s})$, and $T_{DDD_{k^*+1}} = T_{DDD}(\boldsymbol{s})$.

The necessity is straightforward, because the timing k^* along the equilibrium path in the prisoner's dilemma with K timings is equivalent to the beginning of play in the prisoner's dilemma with two timings and no deviation incentive from the two-timing three-mode strategy within timings k^* and $k^* + 1$ in the prisoner's dilemma with K timings is equivalent to no deviation incentive from the corresponding three-mode strategy in the prisoner's dilemma with two timings.

We show the sufficiency. Suppose that the three-mode strategy $\boldsymbol{s}$ is a sequential equilibrium in the prisoner's dilemma with two timings. Then, by the same argument as the necessity part, a deviation from the two-timing three-mode strategy within timings k^* and $k^* + 1$ in the prisoner's dilemma with K timings is not beneficial for

any type. Furthermore, a deviation to moving at timing $k \neq k^*, k^*+1$ is not beneficial either. First, consider a deviation to choosing C at timing $k < k^*$. The expected utility from this deviation is given by

$$\phi(T_{C_{k^*}} \cap \{(\alpha,\beta) \in T | \beta < \beta^*\})(c - \alpha(b-c)) + \phi(T_{C_{k^*}} \cap \{(\alpha,\beta) \in T | \beta \geq \beta^*\})a$$
$$+ \phi(T_{CDD_{k^*+1}})a + \phi(T_{DDD_{k^*+1}})(c - \alpha(b-c))$$

where the first term corresponds to the case in which his opponent is a type $(\alpha,\beta) \in T_{C_{k^*}}$ with $\beta < \beta^*$, he observes the deviation to C at the earlier timing k, and he responds with D at the later timing k^* to the choice of C. Compare it with the expected utility from following C_{k^*}-mode, which is given by

$$\phi(T_{C_{k^*}})a + \phi(T_{CDD_{k^*+1}})a + \phi(T_{DDD_{k^*+1}})(c - \alpha(b-c)).$$

The latter utility is higher than the former by $\phi(T_{C_{k^*}} \cap \{(\alpha,\beta) \in T | \beta < \beta^*\})[a - (c - \alpha(b-c))] > 0$ for any type $(\alpha,\beta) \in T$. Thus, C_{k^*}-mode is strictly better than a deviation to choosing C at timing $k < k^*$, and the equilibrium behavior is at least as good as C_{k^*}-mode by the optimality of the equilibrium behavior within timings k^* and k^*+1. Hence, a deviation to choosing C at timing $k < k^*$ is not beneficial for any type.

Second, consider a deviation to choosing D at timing $k < k^*$. The expected utility from this deviation is d because his opponent will observe the deviation to D and will respond with D for sure. Compare this with the expected utility from following CDD_{k^*+1}-mode, which is given by $\phi(T_{C_{k^*}})a + \phi(T_{CDD_{k^*+1}})d + \phi(T_{DDD_{k^*+1}})d$. The latter utility is higher than the former by $\phi(T_{C_{k^*}})(a-d) > 0$ for any type $(\alpha,\beta) \in T$. Thus, CDD_{k^*+1}-mode is strictly better than a deviation to choosing D at timing $k < k^*$, and the equilibrium behavior is at least as good as CDD_{k^*+1}-mode by the optimality of the equilibrium behavior within timings k^* and k^*+1. Hence, a deviation to choosing D at timing $k < k*$ is not beneficial for any type.

Finally, consider a deviation to moving at $k > k^*+1$. The optimal way of deviation is CDD_k-mode for a type $(\alpha,\beta) \in T$ with $\beta \geq \beta^*$ and DDD_k for a type $(\alpha,\beta) \in T$ with $\beta \leq \beta^*$ because his opponent is supposed to finish choosing either C or D before timing k and he simply chooses his best response to the observed choice by the opponent. Note that the expected utilities from CDD_k-mode and DDD_k-mode are the same as the expected utilities from CDD_{k^*+1}-mode and DDD_{k^*+1}-mode respectively. The equilibrium behavior is at least as good as CDD_{k^*+1}-mode and DDD_{k^*+1}-mode by the optimality of the equilibrium behavior within timings k^* and k^*+1 and hence it is at least as good as deviating to moving at timing $k > k^*+1$. (Q.E.D.)

Proof of Theorem 12

Consider a three-mode strategy with a threshold k^* in a prisoner's dilemma with K timings. Suppose that a three-mode strategy with a threshold k^* is a sequential equilibrium in a prisoner's dilemma with K timings. Let T_{C_k} $(1 \leq k \leq k^*)$, T_{CDD_k} $(k^*+1 \leq k \leq K)$, and T_{DDD_k} $(k^*+1 \leq k \leq K)$ denote the sets of types who follow C_k-mode, CDD_k-mode, and DDD_k-mode respectively in this equilibrium.

[Step 1] First, we show that $\phi(\cup_{1 \leq k \leq k^*} T_{C_k} \cap \{(\alpha,\beta) \in T | \beta < \beta^*\}) > 0$. Suppose to the contrary that $\phi(\cup_{1 \leq k \leq k^*} T_{C_k} \cap \{(\alpha,\beta) \in T | \beta < \beta^*\}) = 0$. Then, any type $(\alpha,\beta) \in T_{C_k}$ with $1 \leq k \leq k^*$ is a type with $\beta \geq \beta^*$. This type responds with C to a choice of C by his opponent at an earlier timing $k' < k$ than him. This means that choosing C at timing k with $1 \leq k \leq k^*$ induces the same outcome; (C,C) and (C,D) are realized with probabilities $\phi((\cup_{1 \leq k \leq k^*} T_{C_k}) \cup (\cup_{k^*+1 \leq k \leq K} T_{CDD_k}))$ and $\phi(\cup_{k^*+1 \leq k \leq K} T_{DDD_k})$ respectively. Note also that following CDD_k with $k^*+1 \leq k \leq K$ induces the

same outcome; (C,C) and (D,D) are realized with probabilities $\phi(\cup_{1\leq k\leq k^*}T_{C_k})$ and $\phi((\cup_{k^*+1\leq k\leq K}T_{CDD_k})\cup(\cup_{k^*+1\leq k\leq K}T_{DDD_k}))$ respectively. Similarly, following DDD_k with $k^*+1\leq k\leq K$ induces the same outcome; (D,C) and (D,D) are realized with probabilities $\phi(\cup_{1\leq k\leq k^*}T_{C_k})$ and $\phi((\cup_{k^*+1\leq k\leq K}T_{CDD_k})\cup(\cup_{k^*+1\leq k\leq K}T_{DDD_k}))$ respectively. The sequential rationality implies that a type $(\alpha,\beta)\in T_{C_k}$ with $1\leq k\leq k^*$ prefers the outcome in which (C,C) and (C,D) are realized with probabilities $\phi((\cup_{1\leq k\leq k^*}T_{C_k})\cup(\cup_{k^*+1\leq k\leq K}T_{CDD_k}))$ and $\phi(\cup_{k^*+1\leq k\leq K}T_{DDD_k})$ to both the outcome in which (C,C) and (D,D) are realized with probabilities $\phi(\cup_{1\leq k\leq k^*}T_{C_k})$ and $\phi((\cup_{k^*+1\leq k\leq K}T_{CDD_k})\cup(\cup_{k^*+1\leq k\leq K}T_{DDD_k}))$ and the outcome in which (D,C) and (D,D) are realized with probabilities $\phi(\cup_{1\leq k\leq k^*}T_{C_k})$ and $\phi((\cup_{k^*+1\leq k\leq K}T_{CDD_k})\cup(\cup_{k^*+1\leq k\leq K}T_{DDD_k}))$. The parallel implications hold for a type $(\alpha,\beta)\in T_{CDD_k}$ with $k^*+1\leq k\leq K$ and a type $(\alpha,\beta)\in T_{DDD_k}$ with $k^*+1\leq k\leq K$. Now, induce a two-timing three-mode strategy by setting $T'_{C_{k^*}}=\cup_{1\leq k\leq k^*}T_{C_k}$, $T'_{CDD_{k^*+1}}=\cup_{k^*+1\leq k\leq K}T_{CDD_k}$, and $T'_{DDD_{k^*+1}}=\cup_{k^*+1\leq k\leq K}T_{DDD_k}$ where $T'_{C_{k^*}}$, $T'_{CDD_{k^*+1}}$, and $T'_{DDD_{k^*+1}}$ denote the sets of types who follow C_{k^*}-mode, CDD_{k^*+1}-mode, and DDD_{k^*+1}-mode respectively in the induced strategy. Then, following C_{k^*}-mode, CDD_{k^*+1}-mode, and DDD_{k^*+1}-mode against the two-timing three-mode strategy induces the same outcome in the equilibrium. Hence, the implications of the sequential rationality in the original sequential equilibrium guarantee that this two-timing three-mode strategy is also a sequential equilibrium. Apply Theorem 11 to the induced two-timing three-mode strategy. Then, the corresponding three-mode strategy $\boldsymbol{s}$ in a prisoner's dilemma with two timings where $T_C(\boldsymbol{s})=T'_{C_{k^*}}$, $T_{CDD}(\boldsymbol{s})=T'_{CDD_{k^*+1}}$, and $T_{DDD}(\boldsymbol{s})=T'_{DDD_{k^*+1}}$ must be a sequential equilibrium. In this three-mode equilibrium, we must have $\phi(T_C(\boldsymbol{s})\cap\{(\alpha,\beta)\in T|\beta<\beta^*\})=0$. This contradicts Lemma 2.

[Step 2] Second, we show that $\phi(T_{C_{k^*}})>0$. Actually, we show a stronger result that $\phi(T_{C_{k^*}}\cap\{(\alpha,\beta)\in T|\beta<\beta^*\})>0$. Suppose to the contrary that $\phi(C_{k^*}\cap\{(\alpha,\beta)\in T|\beta<\beta^*\})=0$. Then, by Step 1, we have $\phi(\cup_{1\leq k\leq k^*-1}T_{C_k}\cap\{(\alpha,\beta)\in T|\beta<\beta^*\})>0$. Hence, there is a timing k' with $1\leq k'\leq k^*-1$ such that $\phi(T_{C_{k'}}\cap\{(\alpha,\beta)\in T|\beta<\beta^*\})>0$. Take a type $(\alpha,\beta)\in T_{C_{k'}}\cap\{(\alpha,\beta)\in T|\beta<\beta^*\}$. His expected utility from following $C_{k'}$-mode is

$$\begin{aligned}&\phi(\cup_{1\leq k\leq k'-1}T_{C_k})(b-\beta(b-c))+\phi(\cup_{k'\leq k\leq k^*-1}T_{C_k})a\\&+\phi(\cup_{k^*+1\leq k\leq K}T_{CDD_k})a+\phi(\cup_{k^*+1\leq k\leq K}T_{DDD_k})(c-\alpha(b-c))\end{aligned}$$

where the first term corresponds to the case in which his opponent is a type $(\alpha,\beta)\in T_{C_k}$ with $1\leq k\leq k'-1$, the opponent chooses C at an earlier timing $k<k'$ than him, and he responds with D at the later timing k' because D is the strict best response to C under $\beta<\beta^*$. His expected utility from deviating to C_{k^*}-mode is

$$\phi(\cup_{1\leq k\leq k^*-1}T_{C_k})(b-\beta(b-c))+\phi(\cup_{k^*+1\leq k\leq K}T_{CDD_k})a+\phi(\cup_{k^*+1\leq k\leq K}T_{DDD_k})(c-\alpha(b-c)).$$

The latter utility is higher than the former utility by $\phi(\cup_{k'\leq k\leq k^*-1}T_{C_k})(b-\beta(b-c)-a)>0$ under $\beta<\beta^*$ because $\phi(\cup_{k'\leq k\leq k^*-1}T_{C_k})\geq\phi(T_{C_{k'}}\cap\{(\alpha,\beta)\in T|\beta<\beta^*\})>0$. This is a contradiction to the equilibrium.

[Step 3] Third, we show that $\phi(T_{C_1})=\cdots=\phi(T_{C_{k^*-1}})=0$. Suppose to the contrary that there is a timing k' with $1\leq k'\leq k^*-1$ such that $\phi(T_{C_{k'}})>0$. Take a type $(\alpha,\beta)\in T_{C_{k'}}$. Suppose a case in which $\beta<\beta^*$. Then, his expected utility from following $C_{k'}$-mode is

$$\begin{aligned}&\phi(\cup_{1\leq k\leq k'-1}T_{C_k})(b-\beta(b-c))+\phi(T_{C_{k'}})a\\&+\phi(\cup_{k'+1\leq k\leq k^*}T_{C_k}\cap\{(\alpha,\beta)\in T|\beta\geq\beta^*\})a+\phi(\cup_{k'+1\leq k\leq k^*}T_{C_k}\cap\{(\alpha,\beta)\in T|\beta<\beta^*\})(c-\alpha(b-c))\\&+\phi(\cup_{k^*+1\leq k\leq K}T_{CDD_k})a+\phi(\cup_{k^*+1\leq k\leq K}T_{DDD_k})(c-\alpha(b-c))\end{aligned}$$

where

1. the first term corresponds to the case in which his opponent is a type $(\alpha, \beta) \in T_{C_k}$ with $1 \leq k \leq k' - 1$, the opponent chooses C at an earlier timing $k < k'$ than him, and he responds with D at the later timing k' because D is the strict best response to C under $\beta < \beta^*$,

2. the second term corresponds to the case in which his opponent is a type $(\alpha, \beta) \in T_{C_{k'}}$ and chooses C at timing k' simultaneously with him,

3. the third term corresponds to the case in which his opponent is a type $(\alpha, \beta) \in T_{C_k}$ with $k'+1 \leq k \leq k^*$ such that $\beta \geq \beta^*$, the opponent moves at timing k after he chooses C at timing k' and responds with C because C is a best response to C under $\beta \geq \beta^*$, and

4. the fourth term corresponds to the case in which his opponent is a type $(\alpha, \beta) \in T_{C_k}$ with $k' + 1 \leq k \leq k^*$ such that $\beta < \beta^*$, the opponent moves at timing k after he chooses C at timing k' and responds with D because D is the strict best response to C under $\beta < \beta^*$.

His expected utility from deviating to C_{k^*}-mode is

$$\begin{aligned}&\phi(\cup_{1\leq k\leq k'^*-1} T_{C_k})(b - \beta(b-c)) + \phi(T_{C_{k^*}})a \\ &+ \phi(\cup_{k^*+1\leq k\leq K} T_{CDD_k})a + \phi(\cup_{k^*+1\leq k\leq K} T_{DDD_k})(c - \alpha(b-c)).\end{aligned}$$

The latter utility is higher than the former utility by

$$\begin{aligned}&\phi(T_{C_{k'}})[(b - \beta(b-c)) - a] \\ &+ \phi(\cup_{k'+1\leq k\leq k^*-1} T_{C_k} \cap \{(\alpha,\beta) \in T | \beta \geq \beta^*\})[(b - \beta(b-c)) - a] \\ &+ \phi(\cup_{k'+1\leq k\leq k^*-1} T_{C_k} \cap \{(\alpha,\beta) \in T | \beta < \beta^*\})[(b - \beta(b-c)) - (c - \alpha(b-c))] \\ &+ \phi(T_{C_{k^*}} \cap \{(\alpha,\beta) \in T | \beta < \beta^*\})[a - (c - \alpha(b-c))] \\ &\geq \phi(T_{C_{k'}})[(b - \beta(b-c)) - a] + \phi(T_{C_{k^*}} \cap \{(\alpha,\beta) \in T | \beta < \beta^*\})[a - (c - \alpha(b-c))] \\ &> 0\end{aligned}$$

where the last inequality holds because $\phi(T_{C_{k'}}) > 0$ by the supposition and $\phi(T_{C_{k^*}} \cap \{(\alpha, \beta) \in T | \beta < \beta^*\}) > 0$ by Step 2. This is a contradiction to the equilibrium.

Suppose the remaining case in which $\beta \geq \beta^*$ for the taken type $(\alpha, \beta) \in T_{C_{k'}}$. Then, his expected utility from following $C_{k'}$-mode is

$$\begin{aligned}&\phi(\cup_{1\leq k\leq k'-1} T_{C_k})a + \phi(T_{C_{k'}})a \\ &+ \phi(\cup_{k'+1\leq k\leq k^*} T_{C_k} \cap \{(\alpha,\beta) \in T | \beta \geq \beta^*\})a + \phi(\cup_{k'+1\leq k\leq k^*} T_{C_k} \cap \{(\alpha,\beta) \in T | \beta < \beta^*\})(c - \alpha(b-c)) \\ &+ \phi(\cup_{k^*+1\leq k\leq K} T_{CDD_k})a + \phi(\cup_{k^*+1\leq k\leq K} T_{DDD_k})(c - \alpha(b-c))\end{aligned}$$

where the first term is different from the corresponding utility for the case of $\beta < \beta^*$ because the type with $\beta \geq \beta^*$ responds with C when his opponent is a type $(\alpha, \beta) \in T_{C_k}$ with $1 \leq k \leq k' - 1$ and the opponent chooses C at an earlier timing $k < k'$ than him. His expected utility from deviating to C_{k^*}-mode is

$$\begin{aligned}&\phi(\cup_{1\leq k\leq k'^*-1} T_{C_k})a + \phi(T_{C_{k^*}})a \\ &+ \phi(\cup_{k^*+1\leq k\leq K} T_{CDD_k})a + \phi(\cup_{k^*+1\leq k\leq K} T_{DDD_k})(c - \alpha(b-c))\end{aligned}$$

where the first term is different from the corresponding utility for the case of $\beta < \beta^*$ for the same reason as his expected utility from following $C_{k'}$-mode. The latter utility is higher than the former utility by

$$\begin{aligned}
&\phi(\cup_{k'+1\leq k\leq k^*-1} T_{C_k} \cap \{(\alpha,\beta)\in T|\beta<\beta^*\})[a-(c-\alpha(b-c))]\\
&+\phi(T_{C_{k^*}}\cap\{(\alpha,\beta)\in T|\beta<\beta^*\})[a-(c-\alpha(b-c))]\\
&\geq \phi(T_{C_{k^*}}\cap\{(\alpha,\beta)\in T|\beta<\beta^*\})[a-(c-\alpha(b-c))]\\
&>0
\end{aligned}$$

where the last inequality holds because $\phi(T_{C_{k^*}} \cap \{(\alpha,\beta)\in T|\beta<\beta^*\}) > 0$ by Step 2. This is a contradiction to the equilibrium.

[Step 4] Finally, we show that $\phi(CDD_{k^*+1}) > 0$. Suppose to the contrary that $\phi(CDD_{k^*+1}) = 0$. By Step 1 through 3, we have $\phi(C_1) = \cdots = \phi(C_{k^*-1}) = 0$ and $\phi(C_{k^*} \cap \{(\alpha,\beta)\in T|\beta<\beta^*\}) > 0$. Take a type $(\alpha,\beta) \in C_{k^*} \cap \{(\alpha,\beta)\in T|\beta<\beta^*\}$. Then, his expected utility from following C_{k^*} is

$$\phi(T_{C_{k^*}})a + \phi(\cup_{k^*+2\leq k\leq K} T_{CDD_k})a + \phi(\cup_{k^*+1\leq k\leq K} T_{DDD_k})(c-\alpha(b-c)).$$

Consider a deviation to the following behavior. Choose $\emptyset$ at timing $k = 1,\cdots,k^*$. At timing k^*+1, choose D if the opponent chooses C or D before timing k^*+1 and choose C if the opponent chooses at timing 1 through k^*. His expected utility from this deviation is

$$\phi(T_{C_{k^*}})(b-\beta(b-c)) + \phi(\cup_{k^*+2\leq k\leq K} T_{CDD_k})a + \phi(\cup_{k^*+1\leq k\leq K} T_{DDD_k})(c-\alpha(b-c)).$$

The latter utility is higher than the former utility by $\phi(T_{C_{k^*}})[(b-\beta(b-c))-a] > 0$ under $\beta < \beta^*$. This is a contradiction to the equilibrium. (Q.E.D.)

Proof of Theorem 13

Suppose that there exists a three-mode equilibrium in $PD((a,b,c,d),f)$. Then, $0 < \alpha^*(\mu) < \bar{\alpha}$ and a type (α,β) with $\alpha = \alpha^*(\mu)$ is indifferent between C and CDD. This means that $\mu_C a + \mu_{CDD}a + \mu_{DDD}[c-\alpha^*(\mu)(b-c)] = \mu_C a + \mu_{CDD}d + \mu_{DDD}d$, which is rearranged into

$$\frac{\mu_{CDD}}{\mu_{CDD}+\mu_{DDD}}a + \frac{\mu_{DDD}}{\mu_{CDD}+\mu_{DDD}}[c-\alpha^*(\mu)(b-c)] = d.$$

Note that $\mu_{DDD} \geq \phi(\alpha^*(\mu) \leq \boldsymbol{\alpha}, \boldsymbol{\beta} < \beta^*)$ and $\mu_{CDD} \leq \phi(\alpha^*(\mu) \leq \boldsymbol{\alpha}, \boldsymbol{\beta} > \beta^*)$. This means that

$$\frac{\mu_{DDD}}{\mu_{CDD}+\mu_{DDD}} \geq \frac{\phi(\alpha^*(\mu)\leq\boldsymbol{\alpha},\boldsymbol{\beta}<\beta^*)}{\phi(\alpha^*(\mu)\leq\boldsymbol{\alpha},\boldsymbol{\beta}>\beta^*)+\phi(\alpha^*(\mu)\leq\boldsymbol{\alpha},\boldsymbol{\beta}<\beta^*)} = \phi(\boldsymbol{\beta}<\beta^*|\alpha^*(\mu)\leq\boldsymbol{\alpha}).$$

Hence,

$$(1-\phi(\boldsymbol{\beta}<\beta^*|\alpha^*(\mu)\leq\boldsymbol{\alpha}))a + \phi(\boldsymbol{\beta}<\beta^*|\alpha^*(\mu)\leq\boldsymbol{\alpha})[c-\alpha^*(\mu)(b-c)] \geq d. \quad (30)$$

When condition (27) holds, the inequality (30) cannot be satisfied if $d > \bar{d}(a)$. Suppose to the contrary that the inequality (30) holds. Then,

$$(1-\phi(\boldsymbol{\beta}<\beta^*|\alpha^*(\mu)\leq\boldsymbol{\alpha}))a+\phi(\boldsymbol{\beta}<\beta^*|\alpha^*(\mu)\leq\boldsymbol{\alpha})[c-\alpha^*(\mu)(b-c)] \geq d > (1-\phi(\boldsymbol{\beta}<\beta^*))a+\phi(\boldsymbol{\beta}<\beta^*)c,$$

which is rearranged into

$$\phi(\boldsymbol{\beta}<\beta^*|\alpha^*(\mu)\leq\boldsymbol{\alpha})) < \frac{1}{1+\frac{b-c}{a-c}\alpha^*(\mu)}\phi(\boldsymbol{\beta}<\beta^*).$$

This contradicts the condition (27) at $\tilde{\alpha} = \alpha^*(\mu)$ and $\tilde{\beta} = \beta^*$. (Q.E.D.)

Proof of Lemma 10

Suppose that $\bar{d}(a)$ is the exact bound for the existence of a three-mode equilibrium. Fix $m > 0$ for a given $a \in (c, b)$. Suppose to the contrary that for any $d \in (c, \bar{d}(a))$ there exists a three-mode equilibrium distribution $\mu = (\mu_C, \mu_{CDD}, \mu_{DDD})$ with $\mu_C \geq m$ in a prisoner's dilemma $PD((a, b, c, d), f)$. Then, we can take a sequence $\{d^n\}_{n=1}^{\infty}$ such that (1) $c < d^n < \bar{d}(a)$, (2) $d^n < d^{n+1}$ and $\lim_{n\to\infty} d^n = \bar{d}(a)$, and (3) $\mu_C^n \geq m$ where $\mu_C^n \equiv \mu_C^{\max}(a, d^n)$. Then, there exists $\mu_C^* \equiv \lim_{n\to\infty} \mu_C^n$ because $d^n < d^{n+1}$ implies $\mu_C^n = \mu_C^{\max}(a, d^n) > \mu_C^{\max}(a, d^{n+1}) = \mu_C^{n+1}$ by Theorem 3. Here, $\mu_C^* \geq m$ because $\mu_C^n \geq m$. Then, $\mu_C^* > 0$ implies that $\lambda(\mu_C^*, a, \bar{d}(a)) > 0$ because $\lambda(\mu_C, a, \bar{d}(a)) > 0$ for any $\mu_C \in (0, \bar{\mu}_C(a, \bar{d}(a)))$ when $\bar{d}(a)$ is the exact bound for the existence of three-mode equilibrium so that there is no three-mode equilibrium in $PD((a, b, c, \bar{d}(a)), f)$. Then, the continuity of $\lambda(\mu_C, a, d)$ with respect to (μ_C, d) guarantees that there exists $\epsilon > 0$ such that $\lambda(\mu_C, a, d) > 0$ for any $d \in (\bar{d}(a) - \epsilon, \bar{d}(a))$ and any $\mu_C \in (\mu_C^* - \epsilon, \mu_C^* + \epsilon)$. We can find N such that $d^N \in (\bar{d}(a) - \epsilon, \bar{d}(a))$ and $\mu_C^N \in (\mu_C^* - \epsilon, \mu_C^* + \epsilon)$ because $\lim_{n\to\infty} d^n = \bar{d}(a)$ and $\lim_{n\to\infty} \mu_C^n = \mu_C^*$. We have $\lambda(\mu_C^N, a, d^N) > 0$. This is a contradiction because the three-mode equilibrium distribution $\mu^N = (\mu_C^N, \mu_{CDD}^N, \mu_{DDD}^N)$ must satisfy $\lambda(\mu_C^N, a, d^N) = 0$. (Q.E.D.)

Proof of Theorem 14

Recall that a three-mode equilibrium distribution μ is a materialist leader pattern if and only if $\alpha^*(\mu, a, d) < \beta^*$. Note that a similar argument to the proof of Lemma 7 leads us to conclude that, for (a, d) fixed, when there exist multiple three-mode equilibria $\mu = (\mu_C, \mu_{CDD}, \mu_{DDD})$ and $\mu' = (\mu'_C, \mu'_{CDD}, \mu'_{DDD})$ in a prisoner's dilemma $PD((a, b, c, d), f)$, $\alpha^*(\mu, a, d) < \alpha^*(\mu', a, d)$ if and only if $\mu_C < \mu'_C$. Hence, for (a, d) fixed, all the three-mode equilibria in a prisoner's dilemma $PD((a, b, c, d), f)$ are of a materialist leader pattern if and only if $\alpha^*(\mu^{\max}(a, d), a, d) < \beta^*(a)$.

Now, suppose that $\bar{d}(a)$ is the exact bound for the existence of a three-mode equilibrium. First, we show that for any $a \in (c, b)$ fixed, there exists $d \in (c, \bar{d}(a))$ such that any three-mode equilibrium in the prisoner's dilemma $PD((a, b, c, d), f)$ is of a materialist leader pattern. Suppose to the contrary that for any $d \in (c, \bar{d}(a))$ there exists a three-mode equilibrium with either a hybrid leader pattern or an inequity concerned leader pattern in a prisoner's dilemma $PD((a, b, c, d), f)$. Then, we can take a pair of sequences $\{d^n\}_{n=1}^{\infty}$ and $\{\mu^n\}_{n=1}^{\infty}$ such that (1) $c < d^n < \bar{d}(a)$, (2) $d^n < d^{n+1}$ and $\lim_{n\to\infty} d^n = \bar{d}(a)$, and (3) $\alpha^*(\mu^n, a, d^n) \geq \beta^*$ where $\mu^n = \mu^{\max}(a, d)$. Then, there exists $\mu_C^* \equiv \lim_{n\to\infty} \mu_C^n$ because $d^n < d^{n+1}$ implies $\mu_C^n = \mu_C^{\max}(a, d^n) > \mu_C^{\max}(a, d^{n+1}) = \mu_C^{n+1}$ by Theorem 3. Lemma 10 guarantees that $\mu_C^* = \lim_{n\to\infty} \mu_C^n = 0$. By Lemma 2, this implies that $\mu_{CDD}^* = \lim_{n\to\infty} \mu_{CDD}^n = \phi(\boldsymbol{\beta} \geq \beta^*)$ and $\mu_{DDD}^* = \lim_{n\to\infty} \mu_{DDD}^n = \phi(\boldsymbol{\beta} < \beta^*)$. Hence, the expression (4) of α^* gives

$$\lim_{n\to\infty} \alpha^*(\mu^n, a, d^n) = \lim_{n\to\infty} \left[\frac{a - d^n}{b - c} \frac{\mu_{CDD}^n}{\mu_{DDD}^n} - \frac{d^n - c}{b - c} \right] = \frac{a - \bar{d}(a)}{b - c} \frac{\phi(\boldsymbol{\beta} \geq \beta^*)}{\phi(\boldsymbol{\beta} < \beta^*)} - \frac{\bar{d}(a) - c}{b - c} = 0.$$

This means that there exists N such that $0 \leq \alpha^*(\mu^N, a, d^N) < \beta^*$. This is a contradiction.

Then, for each $a \in (c, b)$ fixed, we can define

$$\hat{d}_{ml}(a) \equiv \inf\{d \in (c, \bar{d}(a)) | \text{ any three-mode equilibrium in } PD((a, b, c, d), f) \text{ is of a ml-pattern}\}.$$

From the definition it follows that $c \leq \hat{d}_{ml}(a) < \bar{d}(a)$. Furthermore, any three-mode equilibrium in any prisoner's dilemma $PD((a,b,c,d),f)$ with $\hat{d}_{ml}(a) < d < \bar{d}(a)$ is of a materialist leader pattern because $\alpha^*(\mu^{\max}(a,d),a,d)$ is strictly decreasing in d and β^* is constant in d by Lemma 7 so that if $\alpha^*(\mu^{\max}(a,d'),a,d') < \beta^*(a)$ holds for $d' \in (\hat{d}_{ml}(a), \bar{d}(a))$, then $\alpha^*(\mu^{\max}(a,d),a,d) < \alpha^*(\mu^{\max}(a,d'),a,d') < \beta^*(a)$ holds for any $d \in (d', \bar{d}(a))$. There exists a three-mode equilibrium with either a hybrid leader pattern or an inequity concerned leader pattern in any prisoner's dilemma $PD((a,b,c,d),f)$ with $c \leq d < \hat{d}_{ml}(a)$ for the same reason.

Second, we show the properties of $\hat{d}_{ml}(a)$. Consider a, a' with $c < a < a' < b$. We show that $\hat{d}_{ml}(a) \leq \hat{d}_{ml}(a')$. Consider $d \in (c, \hat{d}_{ml}(a))$. Then, there exists a three-mode equilibrium with either a hybrid leader pattern or an inequity concerned leader pattern in $PD((a,b,c,d),f)$. This means that $\alpha^*(\mu^{\max}(a,d),a,d) \geq \beta^*(a)$. Lemma 7 states that $\alpha^*(\mu^{\max}(a,d),a,d)$ is strictly increasing in a and that $\beta^*(a)$ is strictly decreasing in a. Therefore, $\alpha^*(\mu^{\max}(a',d),a',d) > \alpha^*(\mu^{\max}(a,d),a,d) \geq \beta^*(a) > \beta^*(a')$. Hence, there exists a three-mode equilibrium with either a hybrid leader pattern or an inequity concerned leader pattern in any prisoner's dilemma $PD((a',b,c,d),f)$ with $d \in (c, \hat{d}_{ml}(a))$. This means that $\hat{d}_{ml}(a) \leq \hat{d}_{ml}(a')$.

Third, the relation $\hat{d}_{ml}(a) \leq \bar{d}_{ml}(a)$ is immediate from Theorem 9-(2).

Finally, recall from Theorem 8 that for any $d \in (c,b)$ there exists $a_L^{\max}(d)$ and, for any $a \in (a_L^{\max}(d), b)$, the type with the strongest incentive to lead under $\mu^{\max}(a,d)$ is the (β^*,β^*)-type so that $\alpha^*(\mu^{\max}(a,d),a,d) \geq \beta^*(a)$. This means that $\lim_{a \to b} \hat{d}_{ml}(a) = b$. (Q.E.D.)

References

Andreoni, James. 1989. "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence," *Journal of Political Economy*, 97(6): 1447–1458.

Andreoni, James. 1990. "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving," *Economic Journal*, 100(401): 464–477.

Andreoni, James, and John Miller. 2002. "According to GARP: An Experimental Test of the Consistency of Preferences for Altruism," *Econometrica*, 70(2): 737–753.

Arbak, Emrah, and Marie-Claire Villeval. 2013. "Voluntary Leadership: Motivation and Influence," *Social Choice and Welfare*, 40(3): 635–662.

Bartling, Börn, and Ferdinand A. von Siemens. 2010. "The Intensity of Incentives in Firms and Markets: Moral Hazard with Envious Agents," *Labour Economics*, 17(3): 598–607.

Bartling, Börn, and Ferdinand A. von Siemens. 2010. "Equal Sharing Rules in Partnerships," *Journal of Institutional and Theoretical Economics*, 166(2): 299–320.

Battigalli, Pierpaolo, and Martin Dufwenberg. 2007. "Guilt in Games," *American Economic Review*, 97(2): 170–176.

Battigalli, Pierpaolo, and Martin Dufwenberg. 2009. "Dynamic Psychological Games," *Journal of Economic Theory*, 144(1): 1–35.

Bolton, Gary E. 1991. "A Comparative Model of Bargaining: Theory and Evidence," *American Economic Review*, 81(5): 1096–1136.

Bolton, Gary E., and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition," *American Economic Review*, 90(1): 166–193.

Bohnet, Iris, Fiona Greig, Benedikt Herrmann, and Richard J. Zeckhauser. 2008. "Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States," *American Economic Review*, 98(1): 294–310.

Bohnet, Iris, and Richard J. Zeckhauser. 2004. "Trust, Risk and Betrayal," *Journal of Economic Behavior & Organization*, 55(4): 467–484.

Camerer, Colin F. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*, Princeton, NJ: Princeton University Press.

Charness, Gary, and Martin Dufwenberg. 2006. "Promises and Partnership," *Econometrica*, 74(6): 1579–1601.

Charness, Gary, and Matthew Rabin. 2002. "Understanding Social Preferences with Simple Tests," *Quarterly Journal of Economics*, 117(3): 817–869.

Cooper, David J., and John H. Kagel. 2013. "Other Regarding Preferences: A Selective Survey of Experimental Results," to appear in John H. Kagel and Alvin E. Roth, eds., *Handbook of Experimental Economics*. Vol. 2. Princeton, NJ: Princeton University Press.

Demougin, Dominique, and Claude Fluet. 2006. "Group vs. Individual Performance Pay When Workers are Envious," in Dominique Demougin and Christian Shade, eds., *An Economic Perspective on Entrepreneurial Decision Making*. Berlin: Duncker & Humblot.

Demougin, Dominique, Claude Fluet, and Carsten Helm. 2006. "Output and Wages with Inequality Averse Agents," *Canadian Journal of Economics*, 39(2): 399–413.

Desiraju, Ramarao, and David E. M. Sappington. 2007. "Equity and Adverse Selection," *Journal of Economics & Management Strategy*, 16(2): 285–318.

Dubey, Pradeep, John Geanakoplos, and Ori Haimanko. 2012. "Prizes versus Wages with Envy and Pride," *Japanese Economic Review*, 64(1): 98–121.

Duffy, John, and Félix Muñoz-García. 2011. "Signaling Concerns about Fairness: Cooperation under Uncertain Social Preferences," Unpublished manuscript, University of Pittsburgh and Washington State University.

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. "A Theory of Sequential Reciprocity," *Games and Economic Behavior*, 47(2): 268–298.

Dur, Robert, and Amihai Glazer. 2008. "Optimal Contracts When a Worker Envies His Boss," *Journal of Law, Economics, & Organization*, 24(1): 120–137.

Englmaier, Florian, and Achim Wambach. 2010. "Optimal Incentive Contracts under Inequity Aversion," *Games and Economic Behavior*, 69(2): 312–328.

Erlei, Mathias. 2008. "Heterogeneous Social Preferences," *Journal of Economic Behavior & Organizations*, 65(3–4): 436–457.

Falk, Armin, and Urs Fischbacher. 2006. "A Theory of Reciprocity," *Games and Economic Behavior*, 54(2): 293–315.

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation," *Quarterly Journal of Economics*, 114(3): 817–868.

Fehr, Ernst, and Klaus M. Schmidt. 2004. "Fairness and Incentives in a Multi-Task Principal-Agent Model," *Scandinavian Journal of Economics*, 106(3): 453–474.

Fehr, Ernst, and Klaus M Schmidt. 2006. "The Economics of Fairness, Reciprocity and Altruism – Experimental Evidence and New Theories," in Serge-Christophe Kolm and Jean Mercier Ythier, eds., *Handbook on the Economics of Giving, Reciprocity and Altruism*. Vol. 1. Amsterdam: Elsevier/North-Holland.

Fehr, Ernst, Alexander Klein, and Klaus M. Schmidt. 2007. "Fairness and Contract Design," *Econometrica*, 75(1): 121–154.

Fehr, Ernst, Susanne Kremhelmer, and Klaus M. Schmidt. 2008. "Fairness and the Optimal Allocation of Property Rights," *Economic Journal*, 118(531): 1262–1284.

Fischbacher, Urs, Simon Gächter, and Ernst Fehr. 2001. "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment," *Economics Letters*, 71(3): 397–404.

Fudenberg, Drew, and David K. Levine. 2012. "Fairness, Risk Preferences and Independence: Impossibility Theorems," *Journal of Economic Behavior & Organization*, 81(2): 606–612.

Gächter, Simon. 2007. "Conditional Cooperation: Behavioral Regularities from the Lab and the Field and Their Policy Implications," in Bruno S. Frey and Alois Stutzer, eds., *Economics and Psychology*. Cambridge: The MIT Press.

Geanakoplos, John, David Pearce, and Ennio Stacchetti. 1989. "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, 1(1): 60–79.

Grund, Christian, and Dirk Sliwka. 2005. "Envy and Compassion in Tournaments," *Journal of Economics & Management Strategy*, 14(1): 187–207.

Gul, Faruk, and Wolfgang Pesendorfer. 2010. "Interdependent Preference Models as a Theory of Intentions," Unpublished manuscript, Princeton University.

Herrmann, Benedikt, and Christian Thöni. 2009. "Measuring Conditional Cooperation: A Replication Study in Russia," *Experimental Economics*, 12(1): 87–92.

Huck, Steffen, and Pedro Rey-Biel. 2006. "Endogenous Leadership in Teams," *Journal of Institutional and Theoretical Economics*, 162(2): 253–261.

Itoh, Hideshi. 2004. "Moral Hazard and Other-Regarding Preferences," *Japanese Economic Review*, 55(1): 18–45.

Kandel, Eugene, and Edward P. Lazear. 1992. "Peer Pressure and Partnerships," *Journal of Political Economy*, 100(4): 801–817.

Kukushkin, Nikolai S. 2013. "Monotone Comparative Statics: Changes in Preferences versus Changes in the Feasible Set," *Economic Theory*, 52(3): 1039–1060.

Li, Jianpei 2009. "Team Production with Inequity Averse Agents," *Portuguese Economic Journal*, 8(2): 119–136.

Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiment," *Review of Economic Dynamics*, 1(3): 593–622.

Marx, Leslie M., and Steven A. Matthews. 2000. "Dynamic Voluntary Contribution to a Public Project," *Review of Economic Studies*, 67(2): 327–358.

Milgrom, Paul R., and Chris Shannon. 1994. "Monotone Comparative Statics," *Econometrica*, 62(1): 157–180.

Milgrom, Paul R., and Robert J. Weber. 1982. "A Theory of Auctions and Competitive Bidding," *Econometrica*, 50(5): 1089–1022.

Neilson, William S. 2006. "Axiomatic Reference-Dependence in Behavior toward Others and toward Risk," *Economic Theory*, 28(3): 681–692.

Neilson, William S., and Jill Stowe. 2010. "Piece-Rate Contracts for Other-Regarding Workers," *Economic Inquiry*, 48(3): 575–586.

Nosenzo, Daniele, and Martin Sefton. 2012. "Endogenous Move Structure and Voluntary Provision of Public Goods: Theory and Experiment," *Journal of Public Economic Theory*, 13(5): 721–754.

Préget, Raphaële, Phu Nguyen-Van, Marc Willinger. 2012. "Who are the Voluntary Leaders? Experimental Evidence from a Sequential Contribution Game," Bureau d'Economie Théorique et Appliquée, Working Papers: No. 2012-21.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics," *American Economic Review*, 83(5): 1281–1302.

Rasch, Alexander, Achim Wambach, and Kristina Wiener. 2012. "Bargaining and Inequity Aversion: On the Efficiency of the Double Auction," *Economics Letters*, 114(2): 178–181.

Rey-Biel, Pedro. 2008. "Inequity Aversion and Team Incentives," *Scandinavian Journal of Economics*, 110(2): 297–320.

Rivas, M. Fernanda, and Matthias Sutter. 2011. "The Benefits of Voluntary Leadership in Experimental Public Goods Games," *Economics Letters*, 112(2): 176–178.

Rohde, Kirsten I. M. 2010. "A Preference Foundation for Fehr and Schmidt's Model of Inequity Aversion," *Social Choice and Welfare*, 34(4): 537–547.

Rotemberg, Julio J. 2008. "Minimally Acceptable Altruism and the Ultimatum Game," *Journal of Economic Behavior & Organization*, 66(3–4): 457–476.

Rotemberg, Julio J. 2011. "Fair Pricing," *Journal of the European Economic Association*, 9(5): 952–981.

Rotemberg, Julio J., and Garth Saloner 1993. "Leadership Style and Incentives," *Management Science*, 39(11): 1299–1318.

Perry, Motty, and Philip J. Reny. 1993. "A Non-cooperative Bargaining Model with Strategically Timed Offers," *Journal of Economic Theory*, 59(1): 50–77.

Saito, Kota. 2010. "Social Preferences under Risk: Equality of Opportunity vs. Equality of Outcome," Unpublished manuscript, California Institute of Technology.

Sandbu, Martin Eiliv. 2008. "Axiomatic Foundations for Fairness-Motivated Preferences," *Social Choice and Welfare*, 31(4): 589–619.

Santos-Pinto, Luís. 2008. "Making Sense of the Experimental Evidence on Endogenous Timing in Duopoly Markets," *Journal of Economic Behavior & Organization*, 68(3–4): 657–666.

Segal, Uzi, and Joel Sobel. 2007. "Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings," *Journal of Economic Theory*, 136(1): 197–216.

Topkis, Donald M. 1998. *Supermodularity and Complementarity*, Princeton, NJ: Princeton University Press.

Varian, Hal R. 1994. "Sequential Contributions to Public Goods," *Journal of Public Economics*, 53(2): 165–186.

[2014.5.19 1156]