Leadership in Prisoner’s Dilemma with Inequity Aversive Preferences

Koji Abe$^*$

Osaka University

Hajime Kobayashi$^+$

Kansai University

Hideo Suehiro$^\S$

Kobe University

February 28, 2012

Abstract

We develop a new economic theory of leadership in which a leader-follower relationship endogenously emerges in consequences of players’ heterogeneous inequity aversions. We study a prisoner’s dilemma in which players are endowed with Fehr and Schmidt preferences with inequity aversions as their private information and then choose cooperation or defection once at one of two timings that they prefer. We provide a sufficient condition for the existence of leadership equilibrium. Who takes the leadership depends on game parameters. We provide a characterization of the equilibrium leadership patterns. We present an application to organizational design.

JEL Classification: C72, D03, D82

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

$^*$We are grateful to Ernst Fehr and Larry Samuelson for their helpful comments.

$^*$Institute of Social and Economic Research, Osaka University, 6-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan; abe@iser.osaka-u.ac.jp.

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

$^\S$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

In organizations, the members face prisoner’s dilemmas in various situations. For example, team works in organizations can be interpreted as one of such prisoner’s dilemmas. However, the standard assumption that members simultaneously choose actions does not generally hold in those prisoner’s dilemmas played in organizations. It is usual that some members wait to see what others do and some members commit to their decisions before other members makes their choices. Examples are omnipresent. The works by shop clerks in retail shops, clerical workers in offices, architects in design projects are team works and they are usually done in parallel works, where each member can advance or delay the timing of his choices in his work relative to the timing of his fellow worker’s choice by his discretion.

When players can decide timings for their moves at their discretion, the leadership comes into play in achieving cooperation in prisoner’s dilemmas. A player may choose cooperation \((C)\) in advance to his opponent’s choice and this may “influence” the opponent’s delayed choice toward \(C\) instead of defection \((D)\). If this leadership process is realized, the dilemma is resolved.

We develop an economic theory of a mechanism by which the above leadership process is realized in prisoner’s dilemmas.\(^1\) Our theory of leadership is built on two hypotheses on human beings in social dilemmas including prisoner’s dilemma. The first hypothesis is that a player has stable and regular preferences over outcomes of social dilemma, and the preferences reflect the inequity between own payoff and his opponent’s payoff in the outcomes in addition to how large his own material payoff is. Particularly, we hypothesize that a player has Fehr and Schmidt (1999) preference with inequity aversion. Furthermore, we assume variety of the preferences among players under which players may differ in their sensitivities to inequity measured by an envy parameter \(\alpha\) and a guilt parameter \(\beta\).

Our second hypothesis is that players behave in a strategically rational manner in social dilemmas. Each player fully understand a game that he faces, that is, a situation in which a social outcome is determined depending on a whole profile of players’ choices and they differ in their preferences over social outcomes. Each player behaves rationally given this understanding. In this hypothesis, we assume that each player does not know his opponent’s preference but the players hold a common prior about types and they commonly recognize the pattern of behaviors which the opponent takes depending on his preference, that is, a Bayesian strategy is common knowledge among the players.

Under these hypotheses, we consider a Bayesian model of prisoner’s dilemma with endogenous moves. A player is endowed with his type that is defined by a pair \((\alpha, \beta)\) of his envy parameter and his guilt parameter. A type is a realization of a continuous random variable and it is his private information. Each player must choose \(C\) or \(D\). There are two timings for move; timing 1 and timing 2. Each player makes his choice between \(C\) and \(D\) once at one of the two timings and at which timing he makes his choice is at his discretion.

We show that leadership is realized with positive probabilities in a play of a particular sequential equilibrium in the Bayesian game. Our equilibrium strategy divides

---

\(^1\)The seminal work of economic theory of leadership is Hermalin (1998). Since then, several economic theories of leadership have been proposed. See Bolton, Brunnermeier and Veldkamp (2010) and Hermalin (2007) for surveys.
player’s types into three categories: Leader types, defector types, and conditional cooperator types. In this equilibrium, a leader type chooses C at timing 1, whereas a defector type and a conditional cooperator type wait. 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 the opponent’s behaviors, while the latter responds to C with C (and with D otherwise.) We can observe the successful leadership on a path in which a leader type is matched with a conditional cooperator type.

Roughly speaking, our equilibrium strategy relates inequity aversions to the aforementioned three categories in the following way. A type with a low or middle envy parameter is a leader type. A type with a high envy parameter and a low guilt parameter is a defector type. A type with a high envy parameter and a high guilt parameter is a conditional cooperator type.

This assignment of the three types to inequity aversions makes them follow the described behaviors optimally as follows. First, the high-envy-high-guilt type prefers to behave as a conditional cooperator. He is willing to choose C if his opponent chooses C because choosing D makes him feel guilty. However, he is not willing to take a leadership to induce his opponent to choose C because he strongly feels envy if his leadership is betrayed and, under the incomplete information about the opponent’s type, he does not preclude a possibility that his opponent is the defector type who chooses D at timing 2. Second, the low-or-middle envy type is ready to use the opportunity of inducing the conditional cooperator to choose C by committing C at timing 1 because he feels little envy even if he is betrayed by the defector type. Finally, the high-envy-low-guilt type has no bravery of taking the leadership because he feels much stronger envy if his choice of C is betrayed. Furthermore, he chooses D without hesitation at timing 2 because he feels little guilt in responding to C with D. Thus, three different types in inequity aversions optimally follow the three different behavior modes.

We proceed to characterize our equilibrium in terms of who takes the leadership behavior. In our model, specific types of players become a leader and whether a player takes the leadership behavior is endogenously determined in equilibrium depending on both his envy parameter and his guilt parameter. We show that there are three types of equilibrium that differ to each other in their patterns of leadership. The first type equilibrium is one in which only middle-envy type takes the leadership behavior. The low-envy type including the extreme type who maximizes own material payoffs does not take the leadership behavior and chooses D at timing 2. The second type equilibrium is one in which both the low-envy type and the middle-envy type take the leadership behavior. The third type is one in which only the low-envy type takes the leadership behavior. We show that given a distribution of types, each equilibrium prevails depending on the payoffs of prisoner’s dilemma.

This characterization of leadership patterns can be applied to a study of organizational design. We consider prisoner’s dilemma in sequential works. A situation of sequential works means a way of work in which there is an exogenous limitation on an order of works so that a work by one member must be done prior to a work by another member. Examples are omnipresent again. In a retail shop, the shop manager organizes works for his shop clerks and then each shop clerk works in the organized work environment. In a project of architecture, the project manager designs a master plan and then each architect designs its details. An organization needs to assign a member to the position of the first mover in sequential works. It is usual that the first

---

2As we mention in the discussion section, a player of defector type does not need to wait in general. Here, we treat the simplest case to illustrate our equilibrium strategy.
mover is selected from a population who have worked in parallel works. The problem is how to select the first mover effectively depending on a work record in parallel works. According to our theory of leadership, a work record of a player is a record of his revealed types. However, the characterization of leadership patterns indicates that how a player’s type is revealed in a play of prisoner’s dilemma depends on the payoff structure of the dilemma. There is no direct relation in which a player who takes a leadership in some prisoner’s dilemma in parallel works is a good leader for sequential works. We develop a method to select the first mover for sequential works effectively.

This paper has three contributions to our understanding of leadership. First, social dilemma is one of the most basic and universal obstacles for advancement of our society and it is of crucial importance to understand how human beings can and do overcome the dilemma. This paper shows that the leadership is an effective way by which social dilemma is resolved. In our theory, if there is a room for the players to move voluntarily in a social dilemma, the leadership emerges endogenously by the players’ own will and cooperation is realized. This is made possible by social preferences endowed to human beings. The leadership mechanism sheds a new light over the problem of resolving a social dilemma in addition to the existing studies based on institutions such as Fehr and Gächter (2000) and Kosfeld, Okada and Riedl (2009) and long term relationships as in the repeated game literature.

The second contribution is that this paper provides a new economic theory of leadership. The leadership needs a mechanism of endogenous sequencing of moves that generates a leader-follower relation. The payoff structure of prisoner’s dilemma is characteristic in two features in this respect. The first feature is behavioral externality; a player’s payoff depends on the action choice of the opponent, specifically, it increases when the opponent chooses $C$ irrespective of his own choice. The second feature is the lack of strategic externality; a player’s optimal action does not depend on the action choices by the opponent, specifically, $D$ is a dominant choice. The behavioral externality gives a player a desire to induce the opponent towards a choice of $C$ but the lack of strategic externality deprives a method to influence the opponent’s choice. This creates the social dilemma.

Our theory of leadership is based on an observation that these two features are modified in terms of utilities when a player is endowed with Fehr and Schmidt preferences. The lack of strategic externality in payoffs is modified into two contrasting cases in utilities. A high-guilt type is subject to strategic externality. His best response to $C$ is $C$ and his best response to $D$ is $D$. A low-guilt type is not subject to strategic externality. His best response is dominantly $D$. This creates a possibility to influence the opponent’s choice with positive probabilities but to the limited extent. The behavioral externality in payoffs is also modified into two cases in utilities. A high-guilt type’s utility from $D$ decreases when the opponent chooses $C$ while a low-guilt type’s utility increases. Furthermore, although all the type’s utility from $C$ increases when the opponent chooses $C$, the sensitivity is high for a high-envy type and low for a low-envy type. This creates a difference in the strength of desire to achieve the Pareto optimal outcome $(C, C)$ across different types. The combination of contrasting strategic externality and contrasting and differing behavioral externality sorts the types of player into three different behavior modes under the incomplete information of preferences. This generates a leader-follower relationship endogenously in prisoner’s dilemma.

There are several works on leadership in market competition that employ a combination of behavioral externality and strategic externality. For example in Cournot competition, Hamilton and Slutsky (1990), Albæk (1990) and Normann (1997) show that there are two equilibria with endogenously formed Stackelberg leader-follower re-
lation. The Cournot outcome is not supported by an equilibrium because a firm has an incentive to commit to an increased production in advance expecting that it induces his opponent to produce less because of the strategic externality and this leads to the leader’s profit increase due to the behavioral externality. A firm takes a leadership behavior as long as his opponent does not. However, this mechanism allows either firm to be a Stackelberg leader in equilibrium. In contrast, the uncertainty of opponent’s behavior over leader, defector, and conditional cooperator plays a central role in our leadership mechanism. A player takes a leadership if and only if he is a type who finds a benefit in spite of the uncertainty of opponent’s behavior.\textsuperscript{3,4}

The third contribution is that our theory of leadership possesses a strong point in testing the theory. The first hypothesis in our theory assumes that each player has a Fehr and Schmidt (1999) preference with inequity aversion. Rohde (2010) axiomatizes this class of social preferences.\textsuperscript{5} When we let a subject manifest his preferences over social outcomes, this axiomatization enables us to test whether the subject’s manifestation is compatible with a Fehr and Schmidt (1999) preference with inequity aversion. When the compatibility is confirmed, we may proceed to estimate his envy parameter \( \alpha \) and guilt parameter \( \beta \). Our theory predicts a player’s behavior depending on the payoff structure of prisoner’s dilemma and the player’s type \( (\alpha, \beta) \). This means that we can prepare a testable hypothesis for a subject with estimated parameters \( \alpha \) and \( \beta \) under our second hypothesis that players play in a strategically rational manner. Thus, our theory of leadership is testable.

In fact, there is some evidence that suggests that leadership is realized in our game setting. Arbak and Villeval (2011) and Rivas and Sutter (2011) conducted experiments of social dilemma in which each subject is requested to make a choice at one of two timings that he prefers. The experiment of Arbak and Villeval (2011) was implemented in a stranger setting and so the subjects in the experiment faced a social dilemma game in the same setting as the current paper.\textsuperscript{6} They found that some of the subjects take leadership behaviors and some subjects among the one who postpone their choices to timing 2 respond with followership behaviors to the leadership behaviors which they observed. Their finding is compatible with the prediction of our theory.

### 1.2 Literature

The most important literature to our theory is the research on social preferences over outcomes in social dilemmas. Particularly, we employ Fehr and Schmidt (1999) preference with inequity aversion. As Fehr and Schmidt (2006) and Gächter (2007) summarize, there are several alternative approach to model social preferences. One of the most convincing alternative approach is a model of so called warm-grown or altruism. This model says that a utility of a player increases with the well being of other players. Andreoni (1990) and Andreoni and Miller (2002) provide experimental evidence for this model. However, we think that a model of warm-grown or altruism is not enough for explaining a mechanism of leadership. In leadership, some player takes a leadership

\textsuperscript{3}Several authors show that one particular leader-follower relationship is supported in equilibrium under some asymmetry between the firms (for example, Mailath (1993) and van Damme and Hurkens (1999)).

\textsuperscript{4}There are several papers that employ incomplete information settings to explain both how a sequencing of decisions emerges endogenously and why a particular payer makes his decision earlier than others (for example, Gul and Lundholm (1995), Bulow and Klemperer (1999), Sahuguet (2006)). However, an essential component of the driving forces of endogenous sequencing in these works is the exogenously imposed waiting cost.

\textsuperscript{5}Related characterizations are given by Neilson (2006), Sandhu (2008).

\textsuperscript{6}Rivas and Sutter (2011) used a repeated game setting for their experiment.
behavior because he expects that other players do not behave cooperatively unless he
takes the leadership behavior. This means that the behavioral nature of conditional
cooperation by those who postpone their choices to timing 2 is essential for the emer-
gence of leadership behavior.\footnote{Fischbacher, Gächter and Fehr (2001) and Herrmann and Thöni (2009) provide experimental
evidence that people exhibit the behavior of conditional cooperation.} However, as Gächter (2007) points out, a model of
warm-grow or altruism fails to explain the conditional cooperation because a player is
either cooperative unconditionally or uncooperative unconditionally depending on his
parameter of altruism.

Another convincing approach is so called relative income models by Bolton (1991)
and Bolton and Ockenfels (2000). These models say that a utility of a player increases
with his relative income to other players’ incomes. An interpretation according to
Fehr and Schmidt (2006) is that for a given level of own income, a player feels envy
when other players receive more income. A relative income model incorporates the
effect of envy in this interpretation. However, we think that this model is not enough
for explaining a mechanism of leadership. As Fehr and Schmidt (2006) points out,
a relative income model implies the opposite to altruism in terms of other player’s
income because a utility of a player decreases with other players’ incomes. This means
that the conditional cooperation is not realized in a relative income model.

In contrast with a model of warm-grow or altruism and a model of relative income,
the Fehr and Schmidt (1999) preference with inequity aversion models the two dimen-
sional aspect of social preference. The envy and the guilt are treated separately. This enables us to explain the leadership by three different behavioral modes that human
beings exhibit in a given prisoner’s dilemma.

As Fehr and Schmidt (2006) surveys, 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) developed such hybrid models. We employ
the Fehr and Schmidt (1999) model because the model incorporates the minimum set of
elements in social preference that we need for explaining the leadership by a mechanism
which we intend to explore and the resulting theory is testable thanks to the relatively
simple structure of the Fehr and Schmidt (1999) model and its axiomatization.

There are some previous attempts that explain leadership by a model of inequity
aversion. Duffy and Muñoz-García (2011) also employ the same model of Fehr and
Schmidt (1999) preference with inequity aversion as in our theory and develop a
Bayesian model of prisoner’s dilemmas. However, they consider prisoner’s dilemmas
in sequential works for their study of leadership. In a later section of this paper, we
analyze the same model as theirs for the purpose of exploring the implications of our
theory to the organizational design.

Huck and Rey-Biel (2006) study the emergence of leadership in the same class of
social dilemmas in parallel works as we investigate in which a player chooses a level in
real number for his effort. However, they do not use a model of social preference but
employ a particular specification of utility function for a player that directly assigns a
negative effect to a difference between his effort level and his opponent’s effort level.
This means that a conformity pressure on effort choice is directly built in the utility
function. Then, they briefly discuss (in appendix) a private information model with
two possible types for the utility function where the negative effect term exists in one
of them and it is zero in the other and claim that the leadership emerges with positive
probabilities in their model. Their theory heavily depends on the specification of the
utility function and the driving force of leadership is the direct and extreme setting of
the effort conformity pressure in one type and no conformity needs in the other type.
In addition to the current paper and Huck and Rey-Biel (2006), there are several works that study emergence of leadership in the setting of parallel works in social dilemma. Kobayashi and Suehiro (2005, 2008) and Abe, Kobayashi and Suehiro (2010) consider a team production in which each agent chooses his level of effort at one of two timings that he prefers. In contrast with the current paper, they assume that an agent is a self-interested player who pursues his material payoffs only. They introduce uncertainty about the productivity of team and assume that each agent receives a partial private information about the productivity. They show that leadership emerges with positive probabilities in their models. The driving force of the leadership is a mechanism of voluntary signaling about the team productivity and different from the one in the current paper.

Andreoni (2006) considers a voluntary contribution game with a similar uncertainty model to Kobayashi and Suehiro (2005, 2008) and Abe, Kobayashi and Suehiro (2010) in which the value of public goods are uncertain. Prior to the stage of decision about contribution, he puts an additional stage in which each player may invest a cost to get some information about the value of the public goods. He shows that leadership emerges with positive probabilities depending on players’ wealth. The driving force of the leadership is another mechanism of voluntary signaling and different from the one in the current paper.

This paper is organized as follows. In the next section, we illustrate by an example the idea of mechanism of leadership in our theory. Section 3 presents our model. In section 4, we show that the leadership is realized with positive probabilities in equilibrium under some condition. Section 5 is a characterization of the equilibrium in terms of who takes the leadership behavior. We also explore an implication of our theory to the organizational design problem of choosing the first mover for sequential works. Section 6 concludes with some discussions.

2 An Illustrative Example

We illustrate by a degenerate numerical example the idea of mechanism by which the leadership is realized through a combination of different behavior modes driven by diversity in player’s Fehr and Schmidt (1999) preference with inequity aversion.

<table>
<thead>
<tr>
<th>i/j</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>2,2</td>
<td>0,3</td>
</tr>
<tr>
<td>D</td>
<td>3,0</td>
<td>1,1</td>
</tr>
</tbody>
</table>

Table 1: Example of prisoner’s dilemma

Consider the prisoner’s dilemma of Table 1. Suppose three possible types of Fehr and Schmidt (1999) preference with inequity aversion: \((\alpha, \beta) = (0, 0), (1, 0), \) and \((1, 1)\). These preferences are represented by the utility functions for player \(i\)

\[
\begin{align*}
    u_{(0,0)}(x_i, x_j) &= x_i \\
    u_{(1,0)}(x_i, x_j) &= x_i - 1 \times \max\{x_j - x_i, 0\} \\
    u_{(1,1)}(x_i, x_j) &= x_i - 1 \times \max\{x_j - x_i, 0\} - 1 \times \max\{x_i - x_j, 0\}
\end{align*}
\]

\(^8\)There are several papers on signaling theory of leadership. Hermalin (1998) and Komai, Stegeman and Hermalin (2007) are theoretical works, and Komai, Grossman and Deters (2007) and Potters, Selton and Vesterlund (2007) are experimental ones. However, in contrast with the current paper, they study exogenous move games.
where $x_i$ and $x_j$ are material payoffs to player $i$ and player $j$ respectively in Table 1. The $(0,0)$ type feels no envy and no guilt and pursues his material payoffs $x_i$. This is an extreme case of low-envy-low-guilt type. The $(1,0)$ type feels envy when $x_i < x_j$ and discounts his material payoff by disadvantageous inequality $x_j - x_i$ but he feels no guilt. This is an extreme case of high-envy-low-guilt type. The $(1,1)$ type feels envy as much as the $(1,0)$ type and also feels guilt equally when $x_i > x_j$ and discounts his material payoff by advantageous inequality $x_i - x_j$. This is an extreme case of high-envy-high-guilt type. These utility functions give the following utilities to player $i$ of $(0,0)$ type, $(1,0)$ type, and $(1,1)$ type in the prisoner’s dilemma.

<table>
<thead>
<tr>
<th>$i/j$</th>
<th>$C$</th>
<th>$D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C$</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>$D$</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

(0,0) type

<table>
<thead>
<tr>
<th>$i/j$</th>
<th>$C$</th>
<th>$D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C$</td>
<td>2</td>
<td>-3</td>
</tr>
<tr>
<td>$D$</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

(1,0) type

<table>
<thead>
<tr>
<th>$i/j$</th>
<th>$C$</th>
<th>$D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C$</td>
<td>2</td>
<td>-3</td>
</tr>
<tr>
<td>$D$</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

(1,1) type

Table 2: utilities in prisoner’s dilemma of Table 1

Assume that a player is the $(0,0)$ type, the $(1,0)$ type, and the $(1,1)$ type with probabilities $\frac{1}{5}$, $\frac{1}{5}$, and $\frac{2}{5}$.

Now consider a Bayesian strategy in which the $(0,0)$ type chooses $C$ at timing 1, the $(1,0)$ type chooses $D$ at timing 2, and the $(1,1)$ type moves at timing 2 by choosing $C$ if his opponent chooses $C$ at timing 1 and by choosing $D$ otherwise. Then, this Bayesian strategy is a sequential equilibrium strategy.

Let us verify that no type has an incentive to mimic the other types’ behaviors. First, suppose that you are the $(0,0)$ type. Suppose that you chooses $C$ at timing 1. When your opponent follows the described Bayesian strategy, he moves at timing 2 and responds with $C$ to your choice of $C$ if the opponent is the $(1,1)$ type. The opponent betrays you by choosing $D$ at timing 2 if he is the $(1,0)$ type. The opponent chooses $C$ at timing 1 independently if he is the $(0,0)$ type. Each path is realized with probabilities $\frac{3}{5}$, $\frac{1}{5}$, and $\frac{1}{5}$. The expected utility from choosing $C$ at timing 1 is

$$\frac{3}{5} \times 2 + \frac{1}{5} \times 0 + \frac{1}{5} \times 2 = \frac{8}{5}.$$  

If you mimic the $(1,0)$ type and chooses $D$ at timing 2, the opponent of the $(1,1)$ type responds with $D$ in stead of $C$. Then, the expected utility is

$$\frac{3}{5} \times 1 + \frac{1}{5} \times 1 + \frac{1}{5} \times 3 = \frac{7}{5}.$$  

If you mimic the $(1,1)$ type, the opponent’s behavior is the same as in the case of mimicking the $(1,0)$ type but now you respond with $C$ to the opponent’s choice of $C$ at timing 1. Therefore, the expected utility is

$$\frac{3}{5} \times 1 + \frac{1}{5} \times 1 + \frac{1}{5} \times 2 = \frac{6}{5}.$$  

Hence, it is rational for the $(0,0)$ type to choose $C$ at timing 1. Mimicking the $(1,0)$ type is better than mimicking the $(1,1)$ type because he feels no guilt and it is better to respond with $D$ to $C$ at timing 2 than with $C$. The $(0,0)$ type further prefers choosing $C$ at timing 1 to choosing $D$ at timing 2 because the benefit $2 - 1 = 1$ from

---

9The reader may verify that he has no incentive of any other deviation.
the response with C by the (1, 1) type outweighs the cost $0 - 1 = -1$ of the betrayal by the (1, 0) type (plus the cost of giving up the benefit of $3 - 2 = 1$ against the (0, 0) type).

Second, suppose that you are the (1, 0) type. Then, the expected utility of choosing D at timing 2 is

$$\frac{3}{5} \times 1 + \frac{1}{5} \times 1 + \frac{1}{5} \times 3 = \frac{7}{5}$$

while the expected utilities from mimicking the (0, 0) type and the (1, 1) type are

$$\frac{3}{5} \times 2 + \frac{1}{5} \times (-3) + \frac{1}{5} \times 2 = \frac{5}{5}$$

$$\frac{3}{5} \times 1 + \frac{1}{5} \times 1 + \frac{1}{5} \times 2 = \frac{6}{5}.$$ 

Hence, it is rational for the (1, 0) type to choose D at timing 2. It is the same as for the (0, 0) type that it is better for the (1, 0) type to respond with D to C at timing 2 than with C. In contrast with the (0, 0) type, however, it is not worth for the (1, 0) type to choose C at timing 1 because the cost $-3 - 1 = -4$ of the betrayal by the (1, 0) type is much higher.

Finally, suppose that you are the (1, 1) type. Then, the expected utility of moving at timing 2 by responding in the described manner to your opponent’s choice at timing 1 is

$$\frac{3}{5} \times 1 + \frac{1}{5} \times 1 + \frac{1}{5} \times 2 = \frac{6}{5}$$

while the expected utilities from mimicking the (0, 0) type and the (1, 0) type are

$$\frac{3}{5} \times 2 + \frac{1}{5} \times (-3) + \frac{1}{5} \times 2 = \frac{5}{5}$$

$$\frac{3}{5} \times 1 + \frac{1}{5} \times 1 + \frac{1}{5} \times 0 = \frac{4}{5}.$$ 

Hence, it is rational for the (1, 1) type to move at timing 2 by choosing C if your opponent chooses C at timing 1 and by choosing D otherwise. It is not worth for the (1, 1) type to choose C at timing 1 for a similar reason to the (1, 0) type. In contrast with the (1, 0) type, however, it is better for the (1, 1) type to respond with C to C at timing 2 than with D because the (1, 1) type feels guilt.

Thus, we establish that a player of any type plays rationally according to the Bayesian strategy. Suppose that a player of the (0, 0) type and a player of the (1, 1) type are paired and play the prisoner’s dilemma. Then, the (0, 0) type player chooses C at timing 1. The (1, 1) type player moves at timing 2 and chooses C in response to the choice of C by his opponent at the timing 1. Thus, leadership is realized.

### 3 The Model

#### 3.1 The rule of the game PD

We consider a version of prisoner’s dilemma game in which players choose their timings of moves under incomplete information about 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
Table 3: prisoner’s dilemma

<table>
<thead>
<tr>
<th>i / j</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>a, a</td>
<td>c, b</td>
</tr>
<tr>
<td>D</td>
<td>b, c</td>
<td>d, d</td>
</tr>
</tbody>
</table>

Table 3: prisoner’s dilemma

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 $(C, C)$.

We hypothesize that each player $i$ has the Fehr–Schmidt inequity aversive preference over the outcomes described by the pairs of material payoffs in Table 3. The preference is represented by a utility function

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

where $x_i$ and $x_j$ are material payoffs to player $i$ and player $j$ 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 to reflect 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 to reflect a sense of guilt of player $i$.

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. It is shown in Table 4.

Table 4: utility representation of 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 $\alpha_i, \beta_i$ in his utility function but he does not know his opponent’s parameters $\alpha_j, \beta_j$.

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

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

The parameter $\alpha$ is the upper bound of envy parameters and it can be either finite or infinite. The upper bound of guilty parameters is assumed to be 1 because a guilty parameter larger than 1 would mean that more material payoff is undesirable when he is receiving more than his opponent. The envy parameter $\alpha$ is assumed to be no less than the guilty parameter $\beta$ because inequality in material payoffs matters more sensitively when a player receives less than his opponent. We assume that a type $(\alpha, \beta)$ is realized according to a density $f(\alpha, \beta)$ with the full support over $T$. We also assume

\footnote{We do not make a restriction $2a > b + c$ commonly imposed in the repeated prisoner’s dilemma literatures that guarantees $(C, C)$ to be value-maximizing efficient, since it is irrelevant to our analysis.}
that a realization is independent across players. All of these about initial beliefs are assumed to be common knowledge.

Under the incomplete information about their utility functions, the players play the prisoner's dilemma of 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 $\emptyset$ is to choose neither $C$ nor $D$ and postpone his choice until timing 2. 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.

### 3.2 The notion of strategy and equilibrium

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 $s : T \to S$. It assigns to each type $(\alpha, \beta) \in T$ a strategy

$$s(\alpha, \beta) = (a_1(\alpha, \beta), a_C(\alpha, \beta), a_D(\alpha, \beta), a_\emptyset(\alpha, \beta)) \in S$$

which this type follows in a play of the $PD$ where $a_\bullet(\alpha, \beta)$ is the prescribed $a_\bullet$ for each information set for type $(\alpha, \beta)$.

We consider symmetric sequential equilibrium in pure Bayesian strategies in $PD$. This means that both players adopt a pure Bayesian strategy $s : T \to S$ such that, for any type $(\alpha, \beta) \in T$ of player, the strategy $s(\alpha, \beta)$ must be sequentially rational at all of the four information sets in the $PD$ given that the opponent follows $s$. Formally, let $\delta = \delta(s_i, s_j)$ denote the profile of choices made by player $i$ and player $j$ when player $i$ follows the strategy $s_i$ and player $j$ follows the strategy $s_j$. Let $x = x(\delta)$ denote the outcome corresponding to the profile $\delta$ of choices as in Table 3. Then, the expected utility for a type $(\alpha, \beta)$ to play a strategy $s \in S$ given that the opponent follows a Bayesian strategy $s : T \to S$ is given by

$$U_{(\alpha, \beta)}(s, s) = \int_T u_{(\alpha, \beta)}(x(\delta(s, s_j))) f(\alpha_j, \beta_j) d(\alpha_j, \beta_j).$$  \hspace{1cm} (3)

The sequential rationality of $s(\alpha, \beta)$ at timing 1 means

$$s(\alpha, \beta) \in \arg \max_{s \in S} U_{(\alpha, \beta)}(s, s).$$  \hspace{1cm} (4)

If the opponent has chosen $C$ at timing 1, there remains no strategic uncertainty at timing 2. Therefore, the sequential rationality at this information set simply means

$$a_C(\alpha, \beta) \in \arg \max_{a_C \in \{C, D\}} u_{(\alpha, \beta)}(x(a_C, C)).$$  \hspace{1cm} (5)
Simililarly, if the opponent has chosen $D$, 

$$a_D(\alpha, \beta) \in \arg \max_{a_D \in \{C,D\}} u_{(\alpha,\beta)}(x(a_D, D)). \quad (6)$$

Finally, let $\phi_0$ denote a probability measure over $T$ that represents the player’s consistent belief when the opponent has chosen $\emptyset$. Then, the sequential rationality means 

$$a_\emptyset(\alpha, \beta) \in \arg \max_{a_\emptyset \in \{C,D\}} \int_T u_{(\alpha,\beta)}(x(a_\emptyset, a_\emptyset(\alpha_j, \beta_j))) d\phi_0(\alpha_j, \beta_j). \quad (7)$$

We consider the following Bayesian strategy for $PD$.

**Definition 1.** A Bayesian strategy $s : T \rightarrow S$ is a sequential equilibrium strategy if it is a sequential equilibrium for both players to play $s$ with some consistent belief $\phi_0$ given the opponent’s choice $\emptyset$, that is, if the conditions (4) through (7) are satisfied for any $(\alpha,\beta) \in T$.

3.3 The emergence of leadership

We would like to examine a possibility that players achieve Pareto efficient decisions $(C,C)$ in $PD$ through emergence of leadership between the players.

The leadership emerges when one of the players takes the leadership behavior to choose $C$ at timing 1 while the other player waits by choosing $\emptyset$ at timing 1 and, after seeing the opponent’s leadership behavior, he responds with the followership behavior to choose $C$ at timing 2. Formally, for a Bayesian strategy $s : T \rightarrow S$, let 

$$T_L(s) = \{ (\alpha,\beta) \in T | a_1(\alpha,\beta) = C \} \quad (8)$$

denote the set of types who take the leadership behavior, and 

$$T_F(s) = \{ (\alpha,\beta) \in T | a_1(\alpha,\beta) = \emptyset, a_C(\alpha,\beta) = C \} \quad (9)$$

denote the set of types who take the followership behavior. When players play the Bayesian strategy $s : T \rightarrow S$, the leadership emerges along a play path if and only if one player’s type is in $T_L(s)$ and the other player’s type is in $T_F(s)$.

Let $\phi$ denote the probability measure over $T$ induced by the prior density $f$. For a Borel subset $B$ of $T$, it assigns a probability of a player being of a type in $B$ by

$$\phi(B) = \int_B f(\alpha,\beta) d(\alpha,\beta). \quad (10)$$

Then,

**Definition 2.** For a $PD$, we say that leadership emerges in the $PD$ with a positive probability if there exists a sequential equilibrium strategy $s : T \rightarrow S$ such that $\phi(T_L(s)) > 0$ and $\phi(T_F(s)) > 0$.

4 The Emergence of Leadership by Three-mode Equilibrium

4.1 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 defined as the following strategies in $S$ where $C$-mode is described as a reduced strategy with $a_C, a_D, a_\emptyset$ for unreached information sets left unspecified.

$$C = (C, a_C, a_D, a_\emptyset)$$

$$CDD = (\emptyset, C, D, D)$$

$$DDD = (\emptyset, D, D, D)$$

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 $s : T \rightarrow S$, let

$$T_C(s) = \{ (\alpha, \beta) \in T \mid s(\alpha, \beta) = C \}$$

$$T_{CDD}(s) = \{ (\alpha, \beta) \in T \mid s(\alpha, \beta) = CDD \}$$

$$T_{DDD}(s) = \{ (\alpha, \beta) \in T \mid s(\alpha, \beta) = DDD \}$$

denote types who follows $C$, $CDD$, and $DDD$ respectively. Then, the three-mode strategy is defined as follows.

**Definition 3.** A Bayesian strategy $s : T \rightarrow S$ is a three-mode strategy if

$$\phi(T_C(s)) > 0, \phi(T_{CDD}(s)) > 0, \phi(T_{DDD}(s)) > 0, \text{ and } \phi(T_C(s)) + \phi(T_{CDD}(s)) + \phi(T_{DDD}(s)) = 1.$$

When players play a three-mode strategy $s$, four kinds of play paths will be realized with positive probabilities. They 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 corresponds to the emergence of leadership with $T_L(s) = T_C(s)$ and $T_F(s) = T_{CDD}(s)$.

<table>
<thead>
<tr>
<th>$s_i / s_j$</th>
<th>$C$</th>
<th>$CDD$</th>
<th>$DDD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>$C$, $C$</td>
<td>$C$, $\emptyset \rightarrow C$</td>
<td>$C$, $\emptyset \rightarrow D$</td>
</tr>
<tr>
<td>CDD</td>
<td>$\emptyset \rightarrow C$, $C$</td>
<td>$\emptyset \rightarrow D$, $\emptyset \rightarrow D$</td>
<td>$\emptyset \rightarrow D$, $\emptyset \rightarrow D$</td>
</tr>
<tr>
<td>DDD</td>
<td>$\emptyset \rightarrow D$, $C$</td>
<td>$\emptyset \rightarrow D$, $\emptyset \rightarrow D$</td>
<td>$\emptyset \rightarrow D$, $\emptyset \rightarrow D$</td>
</tr>
</tbody>
</table>

Table 5: play paths of three-mode strategies

### 4.2 The three-mode equilibrium

We consider a symmetric sequential equilibrium in three-mode strategies. We call this equilibrium three-mode equilibrium. To characterize three-mode equilibrium, we derive a necessary and sufficient condition for a three-mode strategy to be a sequential equilibrium strategy. Especially, we show that it is enough for each type of players to compare three strategies $C$, $CDD$, $DDD$ for their optimal choice of strategy.

The sequential rationality conditions at timing 2 are straightforward. At timing 2, a player chooses his best response actions to the opponent’s sunk choices $C$ and $D$ (or surely expected $D$ after $\emptyset$). The best response to $D$ is $D$ irrespective of a player’s inequity aversion. On the other hand, a best response to $C$ depends on his parameter $\beta$ that reflects his sense of guilt. The higher his $\beta$ is, the more inclined he is to prefer $C$ given the opponent’s $C$. Formally, we obtain the following result.

**Lemma 1.** A three-mode strategy $s : T \rightarrow S$ satisfies the conditions (5) though (7) if and only if $a_C(\alpha, \beta) = C$ for those types with $\beta > \beta^*$ and $a_C(\alpha, \beta) = D$ for those types with $\beta < \beta^*$ where we define the threshold $\beta^* = \frac{b-a}{2c}$. 

13
For the sequential rationality condition (4) at timing 1, the optimality of a strategy among those with \( a_1 = \emptyset \) is equivalent to the sequential rationality at timing 2. Lemma 1 implies that either CDD or DDD is optimal against a three-mode strategy among those with \( a_1 \neq \emptyset \). The strategies with \( a_1 \neq \emptyset \) are C and \( D = (D_a, a_C, a_D, a_D) \). DDD is as good as D given a three-mode strategy because \( \delta(DDD, s_j) = \delta(D, s_j) = (D, C) \) if \( s_j \) is C, and \( \delta(DDD, s_j) = (D, D) \) if \( s_j \) is either CDD or DDD. Hence, if a strategy is optimal among C, CDD, and DDD, then it is optimal in S. Thus, we establish the following fact.

**Lemma 2.** For a three-mode strategy \( s : T \to S \), the condition (4) is equivalent to

\[
s(\alpha, \beta) \in \arg \max_{s \in \{C, CDD, DDD\}} U_{(\alpha, \beta)}(s, s). \tag{11}
\]

Now, let \( \mu = (\mu_C, \mu_{CDD}, \mu_{DDD}) \) denote a probability distribution that the opponent follows strategies C, CDD, and DDD. Then, the expected utility (3) for a type \((\alpha, \beta)\) to follow C-mode given \( \mu \) is given by

\[
U_{(\alpha, \beta)}(C, \mu) = \mu_C u_{(\alpha, \beta)}(x(C, C)) + \mu_{CDD} u_{(\alpha, \beta)}(x(C, CDD)) + \mu_{DDD} u_{(\alpha, \beta)}(x(C, DDD)) = \mu_C a + \mu_{CDD} a + \mu_{DDD}(c - \alpha(b - c)).
\]

Similarly, the expected utilities from CDD-mode and DDD-mode are:

\[
U_{(\alpha, \beta)}(CDD, \mu) = \mu_C a + \mu_{CDD} d + \mu_{DDD}d
\]

\[
U_{(\alpha, \beta)}(DDD, \mu) = \mu_C(b - \beta(b - c)) + \mu_{CDD}d + \mu_{DDD}d.
\]

By Lemma 2, we can define a set of types for whom C is a best response to \( \mu \) by

\[
T_C^\mu(\mu) = \{(\alpha, \beta) \in T | U_{(\alpha, \beta)}(C, \mu) \geq U_{(\alpha, \beta)}(CDD, \mu), U_{(\alpha, \beta)}(DDD, \mu)\}.
\]

The best response types \( T_C^\mu(\mu) \) and \( T_{DDD}^\mu(\mu) \) are defined in the same manner. Then, we have the following characterization of three-mode equilibrium.

**Lemma 3.** A three-mode strategy \( s : T \to S \) is a sequential equilibrium strategy if and only if

\[
T_C(s) \subseteq T_C^\mu(\mu), \quad T_{CDD}(s) \subseteq T_{CDD}^\mu(\mu), \quad \text{and} \quad T_{DDD}(s) \subseteq T_{DDD}^\mu(\mu)
\]

for \( \mu = (\phi(T_C(s)), \phi(T_{CDD}(s)), \phi(T_{DDD}(s))) \). \tag{12}

Lemma 3 means the following. Consider a three-mode sequential equilibrium strategy \( s : T \to S \). Then, the sequential rationality of the strategy at timing 1 stated as the condition (11) in Lemma 2 is equivalent to the requirement that each player induces consistently from \( s \) a belief \( \mu \) about his opponent’s choice over three behavior modes C, CDD, DDD and chooses a best response to the belief \( \mu \) according to \( s \). A type \((\alpha, \beta) \in T_C(s)\) optimally chooses C. Therefore, \( T_C(s) \subseteq T_C^\mu(\mu) \) for \( \mu = (\phi(T_C(s)), \phi(T_{CDD}(s)), \phi(T_{DDD}(s))) \). The similar explanation applies to types \((\alpha, \beta) \in T_{CDD}(s)\) and \((\alpha, \beta) \in T_{DDD}(s)\). Hence, the condition (12) holds.

Conversely, when a three-mode strategy \( s : T \to S \) satisfies the condition (12), it satisfies the condition (11). Furthermore, the requirements of \( \mu_C > 0, \mu_{CDD} > 0, \) and \( \mu_{DDD} > 0 \) for a thee-mode strategy implies that an information sets at timing 2 after an opponent’s choice of C and \( \emptyset \) at timing 1 are on its play paths. Therefore, the best response properties given \( \mu \) stated in the condition (12) also implies the sequential rationality at timing 2 stated in Lemma 1. Hence, \( s : T \to S \) is a sequential equilibrium strategy.
4.3 A sufficient condition for the existence of the three-mode equilibrium

We develop a sufficient condition for the existence of a three-mode strategy that satisfies the condition (12). Lemma 3 shows that a three-mode equilibrium prevails when players’ expectations about \( C, CDD, DDD \) coincide with a probability distribution over \( C, CDD, DDD \) generated by their best responses to the expectations. We show by a fixed point theorem that there exists an expectation with this property under some condition on the parameters \((a, b, c, d, f)\).

Let

\[
\Delta = \{ \mu = (\mu_C, \mu_{CDD}, \mu_{DDD}) | 0 \leq \mu_C, \mu_{CDD}, \mu_{DDD} \leq 1, \mu_C + \mu_{CDD} + \mu_{DDD} = 1 \}
\]

be the space of probability distributions over \( C, CDD, DDD \). Take \( \mu \in \Delta \). Then, consider a Bayesian strategy \( s : T \to \{ C, CDD, DDD \} \) that is a best response given \( \mu \). This strategy induces a new belief \( \mu' \) over \( C, CDD, DDD \). Denote by \( \Psi(\mu) \) a correspondence that generates the set of \( \mu' \) from \( \mu \) according to this procedure. Formally,

\[
\Psi(\mu) := \left\{ \mu' \in \Delta | \exists s : T \to \{ C, CDD, DDD \} \text{ such that} \\
(1) T_C(s) \subseteq T^*_C(\mu), T_{CDD}(s) \subseteq T^*_{CDD}(\mu), T_{DDD}(s) \subseteq T^*_{DDD}(\mu), \\
(2) \mu'_C = \phi(T_C(s)), \mu'_{CDD} = \phi(T_{CDD}(s)), \mu'_{DDD} = \phi(T_{DDD}(s)) \right\}
\]

Then, by Lemma 3, the existence of a three-mode sequential equilibrium strategy reduces to the existence of a particular kind of fixed point of \( \Psi \).

**Lemma 4.** A three-mode sequential equilibrium strategy exists if there exists \( \mu \in \Delta \) such that

1. \( \mu \in \Psi(\mu) \)
2. \( \mu_C > 0, \mu_{CDD} > 0, \mu_{DDD} > 0 \).

For \( \mu \) that satisfies Condition 1 in Lemma 4, there is an associated Bayesian strategy \( s : T \to \{ C, CDD, DDD \} \) and it satisfies the condition (12). Condition 2 guarantees that it is a three-mode strategy.

To develop a sufficient condition for the correspondence \( \Psi \) to have the properties in Lemma 4, let us study the best response types \( T^*_C(\mu), T^*_{CDD}(\mu), \) and \( T^*_{DDD}(\mu) \). We explore them by examining preferences over \( C, CDD, \) and \( DDD \) because the best response types are defined by the preferences. Take a case of \( \mu \) with \( 0 < \mu_C < 1 \). First, consider a preference over \( C \) and \( CDD \). A player of \((\alpha, \beta)\) 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

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

where we set \( \alpha^*(\mu) = \infty \) when \( \mu_{DDD} = 0 \). Then, \((\alpha, \beta)\) 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

\[
\frac{d - c}{a - d} < \frac{\mu_C}{\mu_{DDD}} < \frac{d - c}{a - d} + \frac{b - c}{a - d} \bar{\alpha}.
\]

\[11^{11}\]The Bayesian strategy \( \alpha \) selects a best response from \( \{ C, CDD, DDD \} \) but is not required to be a three-mode strategy. Therefore, the belief \( \mu' \) is not necessarily \( \mu'_C > 0, \mu'_{CDD} > 0, \) or \( \mu'_{DDD} > 0 \).
Second, consider a preference over $CDD$ and $DDD$. A player of $(\alpha, \beta)$ 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 $\beta^*$ is defined in Lemma 1. He prefers the opposite when $\beta < \beta^*$, and he is indifferent when $\beta = \beta^*$. Note that the threshold $\beta^*$ is $0 < \beta^* = \frac{b - a}{b - c} < 1$.

Finally, consider a preference over $C$ and $DDD$. A player of $(\alpha, \beta)$ 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

$$H(\alpha|\mu) = \frac{\mu_{DDD}}{\mu_C} \alpha + \left[\frac{b - a}{b - c} + \frac{\mu_{DDD} (d - c) - \mu_{CDD} (a - d)}{\mu_C (b - c)}\right].$$

(15)

Then, $(\alpha, \beta)$ 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 $\alpha^*(\mu)$, $\beta^*$ 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 types to $\mu$ with $0 < \mu_C < 1$ are characterized as follows.

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

(16)

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

(17)

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

(18)

An example of the best response types are illustrated in Figure 1. This example is obtained when $\mu$ satisfies $\beta^* < \alpha^*(\mu) < \bar{\alpha}$ and $H(0|\mu) > 0$. All of the best response types are non-degenerate in this example.
\( \mu_{CDD} = \phi(T_{CDD}(\mu)) \), and \( \mu_{DDD} = \phi(T_{DDD}(\mu)) \) and this relation is continuous because the lines \( \alpha = \alpha^*(\mu) \) and \( \beta = H(\alpha(\mu)) \) that give the best response types in (16), (17), and (18) continuously depend on \( \mu \) with \( 0 < \mu < 1 \). This guarantees (together with the upper hemi continuity property of \( \Psi \) at \( \mu \) with \( \mu_C = 0 \) or \( \mu_C = 1 \) that we will prove in Appendix) that there exists a fixed point \( \mu \in \Psi(\mu) \) in \( \Delta \).

Without restriction on \( \phi \), however, there is no guarantee that the fixed point \( \mu \) is in the interior of \( \Delta \) to generate non-degenerate best response types \( T_{C}^*(\mu) \neq \emptyset \), \( T_{CDD}^*(\mu) \neq \emptyset \), and \( T_{DDD}^*(\mu) \neq \emptyset \). We find the following sufficient condition for the correspondence \( \Psi \) to have a fixed point \( \mu \) with \( \mu_C > 0, \mu_{CDD} > 0, \mu_{DDD} > 0 \) so that there exists a three-mode equilibrium associated with \( \mu \).

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

\[
\phi(\beta > \beta^*) > \frac{d - c}{a - c}.
\]  

(19)

The condition (19) has the following simple meaning. Suppose that no type takes a leadership behavior and that all the types \( (\alpha, \beta) \) with \( \beta > \beta^* \) follow \( CDD \) while all the types \( (\alpha, \beta) \) with \( \beta \leq \beta^* \) follow \( DDD \). Then, the belief \( \mu \) consistent with this Bayesian strategy \( s: T \rightarrow S \) is \( \mu_C = \phi(T_C(s)) = 0 \), \( \mu_{CDD} = \phi(T_{CDD}(s)) = \phi(\beta > \beta^*) \), and \( \mu_{DDD} = \phi(T_{DDD}(s)) = \phi(\beta \leq \beta^*) \). When the condition (19) is satisfied, a player of type \( (0, 0) \) prefers \( C \) to \( CDD \) and \( DDD \) given that belief \( \mu \) because

\[
U_{(0,0)}(C, \mu) - U_{(0,0)}(CDD, \mu) = [\phi(\beta > \beta^*)a + \phi(\beta \leq \beta^*)c] - [\phi(\beta > \beta^*)d + \phi(\beta \leq \beta^*)d] \\
= (a - c)\left[\phi(\beta > \beta^*) - \frac{d - c}{a - c}\right] > 0
\]

\[
U_{(0,0)}(C, \mu) - U_{(0,0)}(DDD, \mu) = (a - c)\left[\phi(\beta > \beta^*) - \frac{d - c}{a - c}\right] > 0.
\]

Thus, the condition (19) means the situation in which the type \( (0, 0) \) has an incentive to take a leadership behavior if no one takes a leadership, all the types with \( \beta > \beta^* \) follow \( CDD \), and all the others follow \( DDD \).

An intuition for why the condition (19) guarantees the existence of a three-mode equilibrium is as follows. A sequentially rational player chooses \( a_C(\alpha, \beta) = C \) if \( \beta > \beta^* \). If it were the case that \( \mu_C = 0 \), then all the types \( (\alpha, \beta) \) with \( \beta > \beta^* \) would choose \( \emptyset \) and plan to respond \( C \) with \( C \) at timing 2. As we explained above, this is precisely the situation that the condition (19) means. Then, the type \( (0, 0) \) and his neighbors would deviate to taking a leadership at timing 1. Hence, it can not be the case that no one takes a leadership under the condition (19). We have \( \mu_C > 0 \). When some type takes a leadership, there must be those types from whom the leader tries to induce \( C \). We must have \( \mu_{CDD} > 0 \)\(^{12}\). The types who follows \( CDD \) chooses \( \emptyset \) at timing 1 only because he fears the possibility of playing with types who follows \( DDD \). We must have \( \mu_{DDD} > 0 \). Thus we have a fixed point with \( \mu_C > 0, \mu_{CDD} > 0, \mu_{DDD} > 0 \).

### 4.4 A characterization of the existence of the three-mode equilibrium

Theorem 1 establishes a sufficient condition (19) for the existence of the three-mode equilibrium. Then, we proceed to develop a characterization of the existence of the three-mode equilibrium.

\(^{12}\)Suppose \( \mu_{CDD} = 0 \). Then, either \( \mu_C = 1 \) or \( \mu_{DDD} > 0 \). If \( \mu_C = 1 \), then the type \( (0, 0) \) and his neighbors deviate to \( DDD \). If \( \mu_{DDD} > 0 \), then the type who follows \( C \) deviates to \( CDD \) because he can avoid the risk of his choice \( C \) being matched with \( D \) by a type who follows \( DDD \) while he manages to respond with \( C \) to \( C \) by a type who follows \( C \).
For this purpose, let us reinterpret the condition (19) in the following way. Fix a population of players with a density \( f \) for type realization. Consider a set of prisoner’s dilemma \( PD(f) \equiv \{PD((a, b, c, d), f)|b > a > d > c\} \) that the players may play. One can normalize the set \( PD(f) \) by fixing \( b \) and \( c \) in the measurement of von Neumann–Morgenstern utilities over the outcomes. Then, a prisoner’s dilemma is identified by \((a, d)\). Now, Theorem 1 is restated as giving the following condition on \( d \) for each \( a \) under which a three-mode equilibrium exists when a pair of players selected according to \( f \) play the prisoner’s dilemma \((a, d)\).

**Corollary 1.** For any \( f \) and \( a, b, c \) with \( b > a > c \), there exists a three-mode equilibrium in \( PD((a, b, c, d), f) \) if

\[
c < d < \bar{d}(a, b, c)
\]

where we define the upper bound on \( d \) by \( \bar{d}(a, b, c) \equiv \phi(\beta > \beta^*)a + (1 - \phi(\beta > \beta^*))c \) and it is \( c < \bar{d}(a, b, c) < a, \frac{\partial}{\partial a}\bar{d}(a, b, c) > 0, \lim_{a \to c} \bar{d}(a, b, c) = c, \) and \( \lim_{a \to b} \bar{d}(a, b, c) = b. \)

The upper bound \( \bar{d}(a, b, c) \) in Corollary 1 gives the area of \((a, d)\) in the space of normalized prisoner’s dilemma for which a three-mode equilibrium is guaranteed to exist. It is shown in Figure 2.\(^{13}\)

![Figure 2: Sufficient condition for the existence of the three-mode equilibrium](image)

Now, in light of Corollary 1, we formulate our problem of characterizing the existence of three-mode equilibrium as follows. Fix \( a \) with \( b > a > c \) and consider the interval \((c, a)\) of \( d \) for prisoner’s dilemma games. Consider the subset \( D(a, b, c) \) of \( d \) for which there exists a three-mode equilibrium in the prisoner’s dilemma \( PD((a, b, c, d), f) \). Corollary 1 establishes the nonemptiness of the set \( D(a, b, c) \) by showing \( D(a, b, c) \supseteq (c, \bar{d}(a, b, c)) \neq \emptyset \). We will characterize the set \( D(a, b, c) \).

First, we can show the following fact that the set \( D(a, b, c) \) is an interval.

\(^{13}\)Figure 2 shows the area of the sufficiency condition (20) for the case of \( b = 3, c = 0, \) and \( f(\alpha, \beta) = 6/3 \). We discuss this example for the purpose of studying leadership patterns in a later section.
Theorem 2. For any $f$ and $a, b, c$ with $b > a > c$, there exists a threshold $\bar{d}(a, b, c)$ such that there exists a three-mode equilibrium in $PD((a, b, c, d), f)$ if $c < d < \bar{d}(a, b, c)$ and there is no three-mode equilibrium in $PD((a, b, c, d), f)$ if $\bar{d}(a, b, c) < d < a$, that is, $\mathcal{D}(a, b, c) = (c, \bar{d}(a, b, c))$.

This connectedness property of the set of prisoner’s dilemma games with a three-mode equilibrium follows from the following fact in player’s incentive to lead. Compare two prisoner’s dilemma games with $d$ and $d’$. Suppose $d’ < d$. Consider any belief $\mu$ over $C$, $CDD$, and $DDD$. Then, for any type, the expected utility from $C$ given $\mu$ is independent of $d$ and $d’$ while both the expected utility from $CDD$ and the expected utility from $DDD$ are strictly lower under $d’$ than under $d$. Therefore, under the belief $\mu$, the types that optimally follow $C$ under $d$ also follow $C$ optimally under $d’$ and some types that follow $CDD$ or $DDD$ under $d$ switch to follow $C$ under $d’$. Thus, the prisoner’s dilemma with $d’$ finds more leaders more easily than the one with $d$. When the prisoner’s dilemma with $d$ successfully admits a three-mode equilibrium, the one with $d’$ must also be able to admit a three-mode equilibrium.

Secondly, we can show that the upper bound $\bar{d}(a, b, c)$ of the interval $\mathcal{D}(a, b, c) = (c, \bar{d}(a, b, c))$ 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 the successful leadership. The leadership will not emerge if the gain from the Pareto improvement is limited.

We can also show that under some condition on $f$, the threshold $\bar{d}(a, b, c)$ is identical to the sufficiency bound of Corollary 1 itself, that is, $\bar{d}(a, b, c) = \bar{d}(a, b, c)$. This means that the leadership incentive for the type $(0, 0)$ given no leadership from others provides the minimum size of the gain from the Pareto improvement that is necessary for the leadership to emerge endogenously.

Formally, we denote by $\phi(\beta < \beta^* | \alpha^* \leq \alpha)$ the conditional probability of $\beta < \beta^*$ given an event of $\alpha^* \leq \alpha$ and we set

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

for the case of $\alpha^* = \bar{\alpha}$. Consider $\gamma \equiv \min_{\alpha^* \in [0, \bar{\alpha}]} \phi(\beta < \beta^* | \alpha^* \leq \alpha)$ and define $\bar{d}(a, b, c) \equiv a - \gamma(a - c)$. Then, we can show that this $\bar{d}(a, b, c)$ bounds $\bar{d}(a, b, c)$ from above as follows.

Theorem 3. Fix any $f$ and $a, b, c$ with $b > a > c$. Then, the threshold $\bar{d}(a, b, c)$ for the existence of the three-mode equilibrium is located as $\bar{d}(a, b, c) \leq \bar{d}(a, b, c) \leq \bar{d}(a, b, c)$. The bound $\bar{d}(a, b, c)$ is located as $\bar{d}(a, b, c) < \bar{d}(a, b, c) < a$. Furthermore, when $f$ and $a, b, c$ satisfy the property that

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

(21)

for any $\alpha^* \in (0, \bar{\alpha}]$, then there exists a three-mode equilibrium in $PD((a, b, c, d), f)$ if and only if $c < d < \bar{d}(a, b, c)$, that is, $\mathcal{D}(a, b, c) = (c, \bar{d}(a, b, c))$.

The bound $\bar{d}(a, b, c)$ has the following simple meaning. Take any value $\alpha^* \in (0, \bar{\alpha}]$ and suppose that all the type $(\alpha, \beta)$ with $\alpha < \alpha^*$ follow $C$ while a type $(\alpha, \beta)$ with $\alpha > \alpha^*$ follows $CDD$ if $\beta < \beta^*$ and $DDD$ if $\beta < \beta^*$. Then, the conditional probability of $DDD$ given an event of $CDD$ or $DDD$ is given by $\phi(\beta < \beta^* | \alpha^* \leq \alpha)$. The type $(\alpha, \beta) = (0, 0)$ is indifferent between following $C$ and following $CDD$ if and only if

$$(1 - \phi(\beta < \beta^* | \alpha^* \leq \alpha))a + \phi(\beta < \beta^* | \alpha^* \leq \alpha)c = d.$$ (22)
The lower the conditional probability $\phi(\beta < \beta^*|\alpha^* \leq \alpha)$ is, the higher the value of $d$ that satisfies (22) is. An upper bound of $d$ that satisfies (22) is given by minimizing the weight $\phi(\beta < \beta^*|\alpha^* \leq \alpha)$. We call this bound $\bar{d}(a,b,c)$. It is bounded away from $a$ because $\min_{\alpha \in [0,d]} \phi(\beta < \beta^*|\alpha^* \leq \alpha) > 0$.

An intuition for why there is no three-mode equilibrium in a prisoner’s dilemma with $d > \bar{d}(a,b,c)$ is as follows. Suppose that there exists a three-mode equilibrium for some $d > \bar{d}(a,b,c)$. Then, there must be a type with $\alpha = \alpha^*(\mu)$ who is indifferent between $C$ and $CDD$ given the corresponding belief $\mu$ over $C$, $CDD$, and $DDD$. Note, however, that the type $(\alpha, \beta) = (0,0)$ prefers $CDD$ to $C$ given $\mu$ because $\mu_{DDD} \geq \phi(\alpha^*(\mu) \leq \alpha, \beta < \beta^*)$ and $\mu_{CDD} \leq \phi(\alpha^*(\mu) \leq \alpha, \beta > \beta^*)$ imply

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

Then, the critical type with $\alpha = \alpha^*(\mu) > 0$ must also prefer $CDD$ to $C$ because by following $CDD$ he can avoid a utility loss $\alpha^*(\mu)(b - c)$ that he may incur when he follows $C$ and his opponent follows $DDD$. This is a contradiction.

This argument that there is no three-mode equilibrium for $d > \bar{d}(a,b,c)$ can be applied directly to the type with $\alpha = \alpha^*(\mu)$. It must hold that

$$(1 - \phi(\beta < \beta^*|\alpha^*(\mu) \leq \alpha))a + \phi(\beta < \beta^*|\alpha^*(\mu) \leq \alpha)(c - \alpha^*(\mu)(b - c)) = d. \quad (23)$$

We can show that, when the condition (21) is satisfied, this can not hold for $d > \bar{d}(a,b,c)$. Therefore, under the condition (21), there is no three-mode equilibrium for $d > \bar{d}(a,b,c)$ and the sufficiency bound $\bar{d}(a,b,c)$ is tight.

In general, however, the threshold $\bar{d}(a,b,c)$ for the existence of the three-mode equilibrium may be strictly higher than the sufficiency bound $\bar{d}(a,b,c)$ in Corollary 1 for the following reason. The bound $\bar{d}(a,b,c)$ corresponds to the prisoner’s dilemma in which the type $(0,0)$ is indifferent between taking a leadership and following $DDD$ if no type takes leadership at all. A value $d < \bar{d}(a,b,c)$ makes the type $(0,0)$ inclined to take a leadership, which is enough to guarantees the existence of the three-mode equilibrium. However, this does not necessarily means that there can not be a three-mode equilibrium for $d > \bar{d}(a,b,c)$. It may be the case that although no type has an incentive to take a leadership by himself, some type are willing to take leadership if some other type in addition to him is expected to take a leadership too. Then, $\bar{d}(a,b,c) < \bar{d}(a,b,c)$. We will show examples of $\bar{d}(a,b,c) < \bar{d}(a,b,c)$ and $\bar{d}(a,b,c) = \bar{d}(a,b,c)$ in a later section.

5 Who takes the leadership?

5.1 Classification of patterns in the emergence of leadership in three-mode equilibrium

We will characterize who takes the leadership in three-mode equilibrium when the condition in Theorem 1 is satisfied and at least one three-mode equilibrium exists. The characterization of the best response types (16), (17), and (18) tells that there are three possible types of three-mode equilibrium according to the shape of best response types in the equilibrium. It is shown in Table 6.

There are two criteria for our classification. The first criterion is a comparison of $\beta^*$ with $\alpha^*(\mu)$. Consider a particular type $(\alpha, \beta) = (\beta^*, \beta^*)$. The property $\alpha = \beta$ of this type means that he is subject to the same amount of utility loss $\beta^*$ from disadvantageous inequality and from advantageous inequality. Furthermore, the property
Table 6: classification of three-mode equilibrium

|                  | $H(0|\mu) > 0$ | $H(0|\mu) < 0$ |
|------------------|----------------|----------------|
| $\alpha^*(\mu) > \beta^*$ | Type 1          |                |
| $\alpha^*(\mu) < \beta^*$ |                | Type 3          |

$\beta = \beta^*$ of this type means that he is indifferent between $CDD$ and $DDD$ under the belief $\mu$. In this respect, the condition $\beta^* < \alpha^*(\mu)$ means that $\alpha < \alpha^*(\mu)$ so that he prefers $C$ to $CDD$ and $DDD$. He will take a leadership in equilibrium.

The second criterion is a sign of $H(0|\mu)$. When $H(0|\mu)$ is positive, the type $(0, 0)$ prefers $DDD$ to $C$. Therefore, he will not take a leadership in equilibrium.

Now combine the two criteria. Type 1 equilibrium occurs when $\alpha^*(\mu) > \beta^*$ and $H(0|\mu) > 0$. In this case, the type $(\beta^*, \beta^*)$ is supposed to take a leadership while the type $(0, 0)$ will not take a leadership. In fact, the best response types of type 1 equilibrium is shown in Figure 1, which we used to illustrate best response types. The best response type $T_C(\mu)$ is the area above $\beta = H(\alpha|\mu)$ and left of $\alpha = \alpha^*(\mu)$. Furthermore, this best response types consist of $\alpha > 0$ and $\beta > 0$. In words, leadership is taken by those types with intermediate value of envy parameter and intermediate value of guilt parameter. In other words, a leadership is taken by those players who do care inequality but not too sensitively. In contrast, a player who does not care inequality will not take a leadership and behave according to $DDD$.

Type 2 equilibrium occurs when $\alpha^*(\mu) > \beta^*$ and $H(0|\mu) < 0$. Now the type $(0, 0)$ is supposed to take a leadership since $H(0|\mu) < 0$. The best response types of type 2 equilibrium is shown in Figure 2. The best response type $T_C(\mu)$ is also the area above $\beta = H(\alpha|\mu)$ and left of $\alpha = \alpha^*(\mu)$. However, in contrast with type 1 equilibrium, this best response types are extended to cover the lower end of $\alpha$ and $\beta$ including $(0, 0)$. Leadership is taken by those types whose envy parameter is low or intermediate and whose guilt parameter is high enough to match his envy parameter. In this equilibrium, in addition to those players who care inequality moderately, those players who do not care inequality also take a leadership.

![Figure 3: Type 2](image-url)
Type 3 equilibrium occurs when $\alpha^*(\mu) < \beta^*$ and $H(0|\mu) < 0$. The best response types of type 3 equilibrium is shown in Figure 3. The best response type $T^*_C(\mu)$ is exactly the area above $\beta = H(\alpha|\mu)$. This means that leadership is taken by only those types with low envy parameter and low guilt parameter, that is, those types who do not care inequality.

![Figure 4: Type 3](image-url)

### 5.2 A characterization of leadership patterns

We characterize who will become a leader in three-mode equilibrium. In a three-mode equilibrium $s$, the set of leaders’ types is given by the set $T^*_C(s)$. The analysis of the previous section shows that this set may have different shapes. We characterize which type belongs to the leaders’ types $T^*_C(s)$ depending on the parameters $((a,b,c,d),f)$ of the PD.

Fix a population of players with a density $f$ for type realization. Take $a$ arbitrarily. Then, by Theorem 2, we have an interval $(c;\hat{d}(a,b,c))$ of $d$ for which a three-mode equilibrium exists when a pair of players selected according to $f$ play the prisoner’s dilemma $(a,d)$. We ask which type of equilibrium among the three that we classified above prevails depending on the value of $d$.

We are particularly interested in whether type 1/type 2 equilibrium exists or type 3 equilibrium exists. Type 3 equilibrium is clearly different from type 1/ type 2 equilibrium in their leadership patterns in that those types who care inequality take leadership in type 1 and type 2 equilibrium while only those types who do not care inequality take leadership in type 3 equilibrium. Based on this prominent feature of the leadership pattern in type 3 equilibrium, we will derive a managerial implication of our leadership theory to organizational design in a later section.

First, we develop a bound on $d$ for type 1/type 2 equilibrium to exist and a bound on $d$ for type 3 equilibrium to exist.

**Theorem 4.** Let $\bar{d}_{1,2}(a,b,c) \equiv a - (b - c)\phi(\beta < \beta^*)$ and $\bar{d}_3(a,b,c) \equiv a - (b - c)\gamma$ where we defined $\gamma$ for Theorem 3. Then, $\bar{d}_{1,2}(a,b,c) < \bar{d}_3(a,b,c) < a$, and (i) only type 1 equilibrium or type 2 equilibrium may exist and no type 3 equilibrium exists in
PD((a, b, c, d), f) for any \(d < d_{1,2}(a, b, c)\) and (ii) only type 3 equilibrium may exist and neither type 1 equilibrium nor type 2 equilibrium exists in PD((a, b, c, d), f) for any \(d > d_3(a, b, c)\).

Whether type 1/type 2 equilibrium exists or type 3 equilibrium exists is answered by an incentive for the type \((\alpha, \beta) = (\beta^*, \beta^*)\) to lead. This type is the type with the least envy parameter among those types who would respond to C with C if he does not take a leadership and chooses at timing 2. When the type \((\beta^*, \beta^*)\) has an incentive to lead, type 1/type 2 equilibrium may exist. When the type \((\beta^*, \beta^*)\) has no incentive to lead, only type 3 equilibrium may exist.

In this respect, the bound \(d_{1,2}(a, b, c)\) has the following simple meaning parallel to the bound \(d(a, b, c)\) in (20) that corresponds to the sufficiency condition (19) for the existence of a three-mode equilibrium. Suppose that no type takes a leadership behavior and that all the types \((\alpha, \beta)\) with \(\beta \geq \beta^*\) follow CDD while all the types \((\alpha, \beta)\) with \(\beta < \beta^*\) follow DDD. Then, the belief \(\mu\) consistent with this Bayesian strategy \(s : T \rightarrow S\) is \(\mu_C = \phi(T_C(s)) = 0, \mu_{CDD} = \phi(T_{CDD}(s)) = \phi(\beta \geq \beta^*), \) and \(\mu_{DDD} = \phi(T_{DDD}(s)) = \phi(\beta < \beta^*).\) Consider a player of type \((\beta^*, \beta^*)\). He is indifferent between CDD and DDD since \(\beta = \beta^*.\) When \(d < d_{1,2}(a, b, c)\), he prefers C to CDD given that belief \(\mu\) because

\[
U_{(\beta^*, \beta^*)}(C, \mu) - U_{(\beta^*, \beta^*)}(CDD, \mu) = \Phi(\beta > \beta^*)a + \Phi(\beta < \beta^*)(c - \beta^*(b - c)) - \Phi(\beta > \beta^*)d + \Phi(\beta < \beta^*)d
\]

\[
= (b - c) \left[ \frac{a - d}{b - c} - \Phi(\beta < \beta^*) \right] > 0.
\]

Thus, the condition \(d < d_{1,2}(a, b, c)\) means the situation in which the type \((\beta^*, \beta^*)\) has an incentive to take a leadership behavior if no one takes a leadership, all the types with \(\beta \geq \beta^*\) follow CDD, and all the others follow DDD. In such a situation, the type \((\beta^*, \beta^*)\) takes a lead in any equilibrium. Therefore, no type 3 equilibrium is possible. Only type 1 equilibrium or type 2 equilibrium may exist.

On the other hand, the bound \(d_3(a, b, c)\) has the following simple meaning parallel to the bound \(d(a, b, c)\) in Theorem 3 that describes an upper bound for the existence of a three-mode equilibrium. Take any \(\alpha^* \geq \beta^*\) and suppose that all the types \((\alpha, \beta)\) with \(\alpha < \alpha^*\) follow C while a type \((\alpha, \beta)\) with \(\alpha > \alpha^*\) follows CDD if \(\beta > \beta^*\) and DDD if \(\beta < \beta^*\). Then, the belief \(\mu\) consistent with this Bayesian strategy \(s : T \rightarrow S\) is \(\mu_C = \phi(T_C(s)) = \phi(\alpha^* > \alpha), \mu_{CDD} = \phi(T_{CDD}(s)) = \phi(\alpha^* \geq \alpha, \beta \geq \beta^*), \) and \(\mu_{DDD} = \phi(T_{DDD}(s)) = \phi(\alpha^* \leq \alpha, \beta < \beta^*).\) Consider a player of type \((\beta^*, \beta^*)\). When \(d > d_3(a, b, c)\), he prefers CDD to C given that belief \(\mu\) because

\[
U_{(\beta^*, \beta^*)}(CDD, \mu) - U_{(\beta^*, \beta^*)}(C, \mu) = \Phi(\alpha^* \leq \alpha, \beta \geq \beta^*)d + \Phi(\alpha^* \leq \alpha, \beta < \beta^*)d
\]

\[
- \Phi(\alpha^* \leq \alpha, \beta > \beta^*)a + \Phi(\alpha^* \leq \alpha, \beta \leq \beta^*)(c - \beta^*(b - c))
\]

\[
= (b - c) \phi(\alpha^* \leq \alpha) \left[ \phi(\beta < \beta^*) - \alpha - d \right] > 0.
\]

Thus, the condition \(d > d_3(a, b, c)\) means the situation in which the type \((\beta^*, \beta^*)\) has an incentive not to take a leadership if all the types with \(\alpha < \alpha^*\) take a 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 \((\beta^*, \beta^*)\) never takes a leadership in any equilibrium that would mean \(\alpha^* = \alpha^*(\mu) > \beta^*\) for the corresponding belief \(\mu\). Therefore, neither type 1 equilibrium nor type 2 equilibrium exists. Only type 3 equilibrium may exist.
Second, if the bound $\bar{d}_{1,2}(a, b, c)$ is included in the interval $(c, \hat{d}(a, b, c))$, we have an interval of $d$ for which type 1/ type 2 equilibrium does exist and no type 3 equilibrium exists. Similarly, if the bound $\bar{d}_3(a, b, c)$ is included in the interval $(c, \hat{d}(a, b, c))$, we have an interval of $d$ for which type 3 equilibrium does exist and no type 1/ type 2 equilibrium exist. Under some additional conditions on $f$, these intervals in fact exist as follows.

**Theorem 5.** (i) Suppose

$$\phi(\beta < \beta^*) < \frac{a - c}{b - c}. \quad (24)$$

Then, $c < \bar{d}_{1,2}(a, b, c) < \hat{d}(a, b, c) (\leq \hat{d}(a, b, c))$, and if

$$c < d < \bar{d}_{1,2}(a, b, c), \quad (25)$$

then there exists a three-mode equilibrium of either type 1 or type 2 and there is no three-mode equilibrium of type 3 in $PD((a, b, c, d), f)$.

(ii) Suppose that

$$\frac{a - c}{b - c} \phi(\beta < \beta^*) < \phi(\beta < \beta^*|\alpha^* \leq \alpha) \quad (26)$$

for any $\alpha^* \in [\frac{b - a}{b - c}, \bar{\alpha}]$. Then, $\bar{d}_3(a, b, c) < \hat{d}(a, b, c) (\leq \hat{d}(a, b, c))$ and if

$$\max[c, \bar{d}_3(a, b, c)] < d < \hat{d}(a, b, c), \quad (27)$$

then there exists a three-mode equilibrium of type 3 and there is no three-mode equilibrium of type 1 or type 2 in $PD((a, b, c, d), f)$.

Theorem 5 together with Theorem 4 tells that under some condition on $f$, all the types of equilibrium exist in $PD((a, b, c, d), f)$ depending on the value $d$. Specifically, type 1 and type 2 equilibrium prevail for lower $d$ in $(c, \hat{d}(a, b, c))$ and type 3 equilibrium prevails for higher $d$ in $(c, \hat{d}(a, b, c))$.

Third, when $f$ and $a, b, c$ satisfy the condition (21) so that the bound $\bar{d}(a, b, c)$ is tight for the existence of three-mode equilibrium, we go one step further beyond the classification of equilibrium into three types and we can develop a comparative static result on the most envious leader. Formally, for a Bayesian strategy $s : T \to S$, let $\bar{\alpha}_L \equiv \sup\{\alpha| (\alpha, \beta) \in T_C(s)\}$ be the supremum of the set of envy parameters of the leaders. Then, we can bound it from above arbitrarily by choosing $d$ as follows.

**Theorem 6.** Fix any $f$ and $a, b, c$ with $b > a > c$. Suppose that $f$ and $a, b, c$ satisfy the condition (21) in Theorem 3. Then, for any $\varepsilon > 0$, there exists a bound $\bar{d}_{\alpha_L}(a, b, c|\varepsilon)$ of $d$ such that $\bar{d}_{\alpha_L}(a, b, c|\varepsilon) < \bar{d}(a, b, c)$ and, for any $d$ with $\bar{d}_{\alpha_L}(a, b, c|\varepsilon) < d < \hat{d}(a, b, c)$, the supremum $\bar{\alpha}_L$ of envy parameters of leaders in any thee-mode equilibrium in $PD((a, b, c, d), f)$ is $0 < \bar{\alpha}_L < \varepsilon$.

When the condition (21) is satisfied, $PD((a, b, c, d), f)$ admits a three-mode equilibrium if and only if $c < d < \hat{d}(a, b, c)$. Theorem 6 means that when we let $d$ increase to the upper bound $\hat{d}(a, b, c)$ within this interval, only type 3 equilibrium exists and all the types of leaders in the equilibrium are arbitrarily close to the type $(\alpha, \beta) = (0, 0)$.14

---

14Take $\varepsilon$ in Theorem 6 as small enough as $\varepsilon \leq \beta^*$. Then, any equilibrium must be of type 3.
5.3 Examples

In this section, we demonstrate and elaborate the results achieved in earlier sections by using a set of examples of the PD games. We examine five type distributions over a common type space \( T = \{(\alpha, \beta) | 0 \leq \alpha \leq 1, 0 \leq \beta \leq 1, \alpha \leq \beta\} \). We fix \( a = 2, b = 3, c = 0 \) in all the examples and compute all the three-mode equilibria that exist for various values of \( d \) under the distributions. Figure 5 through 9 show the loci of \( \alpha^*(\mu) \) of all the three-mode equilibrium computed for varying \( d \) under the distributions. The blue line shows the \( \alpha^*(\mu) \) of type 1 equilibrium, the red line shows type 2 equilibrium, and the green line shows type 3 equilibrium. Note that \( \beta^* = \frac{b+c}{b-c} = \frac{1}{3} \) in all the examples. Therefore, \( \alpha^*(\mu) > \frac{1}{3} \) means that the equilibrium is either type 1 or type 2 while \( \alpha^*(\mu) < \frac{1}{3} \) means that the equilibrium is type 3.

First, recall our results of Theorem 4 and 5 saying that there exist bounds for the existence of equilibrium of type 1/2 and the existence of equilibrium of type 3. Figure 5 through 9 show that each of type 1, type 2, and type 3 equilibrium in fact exists for some value of \( d \) under some type distribution.

Second, fix a type distribution and consider which type equilibrium exist for varying \( d \) under the distribution. In the example of Figure 5, we have all of the three types of equilibrium depending on \( d \). Type 1 equilibrium prevails for lower values of \( d \), type 2 equilibrium prevails for middle values of \( d \), and type 3 equilibrium prevails for higher values of \( d \). This means that the players in the same population (represented by the type distribution of Figure 5) exhibit different leadership patterns in different prisoner’s dilemmas (corresponding to different values of \( d \)).

On the other hand, there exists only type 3 equilibrium over the whole range of \( d \) in the example of Figure 7. This means that the population by the distribution of Figure 7 plays in a similar way in all the prisoner’s dilemma. When leadership is realized, the leadership behavior is always taken by a more or less self-interested player (type \((0,0)\) and his neighbors).

The example of Figure 6 provides an intermediate case. Two types of equilibrium appear depending on \( d \). Type 2 equilibrium prevails for lower values of \( d \) and type 3 equilibrium prevails for higher values of \( d \). The leadership pattern is common to all the prisoner’s dilemma in that the self-interested player (type \((0,0)\)) takes the leadership behavior. However, a player with middle value of envy parameter may or may not take the leadership behavior depending on the payoffs of prisoner’s dilemma.

Third, observe that although there exists an unique equilibrium for each value of \( d \) in the examples of Figure 5 through 7, it is not always the case. In the examples of Figure 8 and 9, there exist multiple equilibria for some value of \( d \). In Figure 8, there exist multiple equilibria of the same type and multiple types of equilibrium for a given value of \( d \); two equilibria of type 2 plus one equilibrium of type 3 for some value of \( d \) and two equilibria of type 3 plus one equilibrium of type 2 for another value of \( d \). The players in the same population by the distribution of Figure 8 play differently in the same prisoner’s dilemma depending on the equilibrium selection. Especially, the leadership patterns may differ. If type 2 equilibrium is selected, a type with relatively high envy parameter \( \alpha > \beta^* \) takes the leadership. However, if type 3 equilibrium is selected, only a type with low envy parameter \( \alpha < \beta^* \) takes the leadership behavior in the same prisoner’s dilemma. The same applies to Figure 9.

Fourth, recall the bound \( \bar{d} \) sufficient for the existence of three-mode equilibrium described in Corollary 1. The dotted line describes \( \bar{d} \) in each example. Its location differs depending on the type distribution. The examples of Figure 5, 6, and 7 satisfy...
the condition (21) that the bound $\bar{d}$ is tight for the existence of three-mode equilibrium. The figures show that, as Theorem 3 guarantees, there exists a three-mode equilibrium if and only if $c < d < \bar{d}$ in these examples. On the other hand, the examples of Figure 8 and 9 do not satisfy the tight bound condition (21). The example of Figure 9 shows that there does exist a three-mode equilibrium for some $d > \bar{d}$ when the tight bound condition (21) is not satisfied. On the other hand, the example of Figure 8 shows that there is a case in which the sufficient bound $\bar{d}$ is in fact the upper bound for the existence of three-mode equilibrium even if the tight bound condition (21) is not satisfied.

Finally, recall our result of Theorem 6 describing the asymptotic property of $\alpha^*(\mu)$ when $d$ is increased to the sufficient bound $\bar{d}$. The figures of 5, 6, and 7 show that, as Theorem 6 guarantees, $\alpha^*(\mu)$ converges to 0 as $d$ goes to $\bar{d}$ when the tight bound condition (21) is satisfied. The example of Figure 8 shows that there is a case in which the same convergence property holds even if the tight bound condition (21) is not satisfied. The example of Figure 9 shows a case in which the convergence property collapses. In this example, when $d$ converges to $\bar{d}$, only type 2 equilibrium survives so that the most envious leader is the type $\bar{\alpha}_L = \alpha^*(\mu)$ and it is bounded from below as $\alpha^*(\mu) > \beta^* = \frac{1}{3} > 0$.

Figure 5: Equilibrium locus under $f(\alpha, \beta) = 6\beta$

Figure 6: Equilibrium locus under $f(\alpha, \beta) = 6\beta$
Figure 6: Equilibrium locus under $f(\alpha, \beta) = 2$

Figure 7: Equilibrium locus under $f(\alpha, \beta) = 3\{(\alpha - 1/3)^2 + (\beta - 1)^2\}$

Figure 8: Equilibrium locus under $f(\alpha, \beta) = 60(1 - \alpha)\beta^2$
5.4 A managerial implication to organizational design: appointment of leader for vertical teams

Our theory of leadership in the PD games can be applied to a study of organizational design. Many works in organizations are comprised of two types of team works. One is horizontal team work, and the other is vertical team work. Horizontal team work means that members in the same layer of a hierarchy work together as a team. This type of team work is a typical team production problem as in Holmström (1982). On the other hand, vertical team work means that members between adjacent layers in a hierarchy work together as a team. In this team work, a member in the upper layer typically decides a working flow of the team, and the one in the lower layer works along the line of the working flow.

Both types of team work can be interpreted as a situation in which members play a prisoner’s dilemma. However, timings of choosing actions differ depending on the types of team works. Team members simultaneously choose effort levels in horizontal team works. In contrast, in vertical team works, a member in the upper layer first chooses an effort level for designing a work flow, and then one in the lower layer selects an effort level for the given works.

For example, architectural offices have the two kinds of team works in each project. A project manager designs a master plan, and then the other architects design its details. Shops such as supermarkets and apparel shops are another example. Members of the shops are usually entitled the roles such as shop manager and shop clerk. The relationship between shop clerks corresponds to horizontal team work, and the one between a shop manager and a clerk corresponds to vertical team work.\footnote{This view corresponds to Likert’s linking pin model in management literature (see Likert (1961)).}

Note that in vertical team work, the manager of the team is usually appointed from horizontal team members. Therefore, selecting the most appropriate manager is one of the most important issues in designing organizations. Because our theory of leadership is the theory in which the emergence of leadership in horizontal team work, we can provide an implication of who should be the appointed leader.

Let us describe a vertical team work in the following framework. A member in the lower layer of the team corresponds to a member in horizontal team work and his type is realized by a distribution $f$. A leader of the team is appointed from some horizontal team in the organization by some procedure. Let $g$ be the type distribution of the leader selected under the procedure.

Given this organization structure, consider the sequential PD game (SPD game) as follows. There are two roles: leader and follower. Before a play, a type of the leader is realized by $g$ and a type of the follower is realized by $f$. Then, under the incomplete information about their utility functions, the players play the prisoner’s dilemma of 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 observe the choice of the leader. Then, the follower chooses $C$ or $D$ at timing 2. This is the end of a play. A player receives the material payoff in Table 3 corresponding to the pair of their choices.

The leader has one information set at timing 1. His (pure) strategy is a plan of choice at this move. Formally, it is $s_L = a_1 \in \{C, D\}$. 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. Formally, it is a pair $s_F = (a_C, a_D)$ where $a_C \in \{C, D\}$ is a prescribed choice
when the leader’s choice is \( C \) and \( a_D \in \{C, D\} \) is a prescribed choice when the leader’s choice is \( D \). Call a strategy \( CD \) when \( a_C = C, a_D = D \) and \( DD \) when \( a_C = a_D = D \).

The Bayesian strategy and the sequential equilibrium are defined in the same manner as for the PD game.

The sequential equilibrium in a SPD is characterized as follows.

**Theorem 7.** Fix \( f, g \) and \( a, b, c, d \) with \( b > a > d > c \). Then, there exists a sequential equilibrium in SPD\((a, b, c, d), (g, f)\). The sequential equilibrium strategy of the leader is to choose \( C \) if his type \((\alpha, \beta)\) is \( \alpha < \alpha^{**} \) and \( D \) if it is \( \alpha > \alpha^{**} \) where we define the threshold

\[
\alpha^{**} = \frac{1}{(1 - \phi(\beta^{**}))(b - c)}[d(a, b, c) - d].
\]

The sequential equilibrium strategy of the follower is to follow \( CD \) if his type \((\alpha, \beta)\) is \( \beta > \beta^{*} \) and \( DD \) if it is \( \beta < \beta^{*} \).

For a SPD, we say that leadership emerges in the SPD when the leader chooses \( C \) and the follower responds with \( C \). From Theorem 7, we have the following characterization of the emergence of leadership in SPD games.

**Theorem 8.** Fix \( f, g \) and \( a, b, c, d \) with \( b > a > d > c \). Then, leadership emerges in SPD\((a, b, c, d), (f, g)\) with a positive probability if and only if \( c < d < d(a, b, c) \).

Next, we investigate what type of \( g \) is better for organizations, when members in organizations are engaged in team production by vertical teamwork. Vertical team works must cope with prisoner’s dilemma for any \( a, b, c, d \). From that point of view, we can define relative superiority or inferiority of a leader as follows.

**Definition 4.** Fix \( f \). Player \( i \) is said to be a better leader in SPD than player \( j \) if

1. for any \( a, b, c, d \), if player \( j \) chooses \( C \) in SPD\((a, b, c, d), (f, g_j)\), then player \( i \) also chooses \( C \) in SPD\((a, b, c, d), (f, g_i)\), and
2. there exists \( a', b', c', d' \) such that player \( i \) chooses \( C \) in SPD\((a', b', c', d'), (f, g_i)\) while player \( j \) chooses \( D \) in SPD\((a', b', c', d'), (f, g_j)\)),

where \( g_i \) and \( g_j \) are the type distributions of players \( i \) and \( j \) respectively.

By Theorem 7, we can tell who is better as a leader according to the size of \( \alpha \). That is, a member who has smaller \( \alpha \) is appropriate for a leader. However, we can not observe the value of \( \alpha \) directly, but can observe the histories of actions that each member took in horizontal team works. Then, the distribution \( g \) is the belief updated by the action histories.

For a given history of actions in PDs by the players in the organization, the authority would like to choose such a player \( i \) to the leader position that there is no player \( j \) who is a better leader in SPD than player \( i \). For this purpose, the authority can define the following partial ordering on player’s records of behaviors in PD games.

**Definition 5.** Fix \( f \). Player \( i \) is said to have a better record of leadership in PD than player \( j \) if there are two PDs, PD\((a, b, c, d), f\) and PD\((a', b', c', d'), f\) such that

1. \( \bar{\alpha}_L < \bar{\alpha}'_L \),
2. Player \( i \) chose \( C \) at timing 1 in PD\((a, b, c, d), f\), and
3. Player \( j \) chose \( D \) at timing 1 in PD\((a', b', c', d'), f\),
where we defined $\alpha_L$ as the supremum of envy parameters of leaders in $PD((a, b, c, d), f)$ for Theorem 6 and we here define $\alpha'_f \equiv \inf\{\alpha | (\alpha, \beta) \in T_{DDD}(s')\}$ as the infimum of envy parameters of followers in three-mode equilibrium $s'$ of $PD((a', b', c', d'), f)$.

Then, we have the following criterion for a choice of leader for SPDs.

**Theorem 9.** Fix $f$. If player $i$ has a better record of leadership in $PD$ than player $j$, then player $i$ is a better leader in SPD than player $j$.

Theorem 9 tells that the authority should appoint the players with the best records of leadership in $PD$ to the leader in SPD.

Note that the authority needs at least two different PDs for each pair of players in order to screen them by records of behaviors in $PD$. Especially, a single $PD$ is not enough even if one player chose $C$ at timing 1 in the $PD$ and the other player chose $\emptyset$ in the same $PD$. For example, suppose that a player of type $(\beta^*, \beta^*)$ and a player of type $(0, 0)$ is paired and they play an equilibrium of type 1 in a $PD((a, b, c, d), f)$. Then, the type $(\beta^*, \beta^*)$ follows $C$ and chooses $C$ at timing 1 while the type $(0, 0)$ follows $DDD$ and chooses $\emptyset$. If a player with the former type $(\beta^*, \beta^*)$ is appointed to the leader position in SPD, he will choose $D$ in a SPD with $a^{**}$ such that $0 < a^{**} < \beta^*$. However, a player with the latter type $(0, 0)$ will choose $C$ in the same SPD.

In fact, the type $(0, 0)$ is the best type for the leader position in SPD because this type is the only type who chooses $C$ in any $SPD((a, b, c, d), (f, g))$ with $c < d < \tilde{d}(a, b, c)$, which is all that the authority can expect from effective leadership assignment for SPDs. When the authority has rich enough records of behaviors by the players in $PD$, it may screen out this best type as follows.

**Theorem 10.** Fix $f$ and $a, b, c$ with $b > a > c$. Suppose that $f$ and $a, b, c$ satisfy the condition (21). Consider a sequence of $PD((a, b, c, d^a), f)$ with $c < d^a < \tilde{d}(a, b, c)$ for each $n$ and $\lim_{n \to \infty} d^a = \tilde{d}(a, b, c)$. Suppose a player who chose $C$ at timing 1 in the $PD((a, b, c, d^a), f)$ for any $n$. Then, if the player is appointed to the leader position in SPD, he will choose $C$ for any $SPD((a', b', c', d'), (f, g))$ with $c < d' < \tilde{d}(a', b', c')$.

### 6 Discussion

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

#### 6.1 Other leadership equilibrium

We studied the leadership in prisoner’s dilemma by three-mode equilibrium in which players follow one of three behavior modes $C$, $CDD$, and $DDD$. Some discussions on other possibilities of the emergence of leadership are 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 more complexed 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; $C$, $CDD$, $DDD$, and $D$. 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; $C$, $CDD$, $DDD$, $D$, and $CDC$ that prescribes $a_C = C$, $a_D = D$, and $a_0 = C$. 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 above-mentioned equilibria exhaust the list of equilibrium in which leadership is realized. We focused on our three-mode equilibrium in our analysis of leadership in prisoner’s dilemma because it is the simplest manner in which leadership is realized in prisoner’s dilemma.

6.2 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 $D$.

Particularly, consider a wait-and-defect equilibrium in which all the types postpone their choices to timing 2 and then choose $D$. By Lemma 1, 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 $\bar{d}(a,b,c) \leq d < a$. Hence, together with Theorems 1 through 3, we have a condition for a combination of the $(C,C)$ outcome by 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 $c \leq d < \bar{d}(a,b,c)$, 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 $\bar{d}(a,b,c) \leq d < \bar{d}(a,b,c)$, both a three-mode equilibrium and the wait-and-defect equilibrium exist. Then, the players need to coordinate on a three-mode equilibrium to achieve cooperation. For $\bar{d}(a,b,c) \leq d < a$, 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 $c \leq d < \bar{d}(a,b,c)$ in which a three-mode equilibrium exists and the wait-and-defect equilibrium does not exist, there still remains a possibility of the $(D,D)$ outcome by a wider class of equilibrium $s : T \to S$ that assigns to each type $(a, \beta)$ either $D$, $CDD$, or $DDD$. This strategy prescribes an early defection at timing 1 for those types who are assigned $D$. Lemma 1 tells that $CDD$ must be assigned only to those types with $\beta \geq \beta^*$ and $DDD$ must be assigned only to those types with $\beta < \beta^*$. This means that $0 \leq \phi(T_{CDD}(s)) \leq \phi(\beta > \beta^*)$. By the same argument as the discussion after Theorem 1, such a $s$ is a sequential equilibrium strategy if and only if the $(0,0)$ type has no incentive to deviate from $s$ to choosing $C$ at timing 1, that is, if and only if $\phi(T_{CDD}(s))a + (1 - \phi(T_{CDD}(s)))c \leq d$. This condition is rewritten as $0 \leq \phi(T_{CDD}(s)) \leq \hat{\mu}_{CDD}(d)$ by defining $\hat{\mu}_{CDD}(d) = \min\left(\frac{d}{a-c}, \phi(\beta > \beta^*)\right)$. When $c \leq d < \bar{d}(a,b,c)$, one can find such a $s$ that $0 \leq \phi(T_{CDD}(s)) \leq \hat{\mu}_{CDD}(d) < \phi(\beta > \beta^*)$. This means that if many types who would become conditional cooperators (that is,
those types with $\beta > \beta^*$ with a probability more than $\phi(\beta > \beta^*) - \mu_{CDD}(d) > 0$ do not wait until timing 2 and choose $D$ early at timing 1, the uncooperative outcome $(D, D)$ prevails in equilibrium. Then, in order to achieve cooperation through leadership, the players need to coordinate on a three-mode equilibrium.

### 6.3 Complete information

We studied the leadership in prisoner’s dilemma under the incomplete information about payer’s preferences. However, there are also some cases that are appropriate to be modeled as complete information games. We briefly discuss the leadership in prisoner’s dilemma under the complete information.\(^\text{16}\)

One can show that, under the complete information of preferences, $(C, C)$ is realized by leadership if at least one of the players is of high-guilt type. Namely, if a player is of high-guilt type and the other player is of 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 of 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 the complete information of preferences is much simpler than the mechanism of the current paper. A player becomes a leader because he knows that his opponent is of 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 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 the current paper, both an envy parameter and a guilt parameter play essential roles in both 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 preference is his private information.

### 6.4 Simultaneous move

We studied the possibility of cooperation in prisoner’s dilemma by the leadership in a model with endogenous timings for moves. We argue that the cooperation by the leadership differ from the cooperation without leadership.

Consider a simultaneous move prisoner’s dilemma as the same game as $PD$ except that there is a single timing for move at which each player must choose $C$ or $D$ simultaneously and independently. A (pure) Bayesian strategy $s$ assigns to each type $(\alpha, \beta) \in T$ a choice $s(\alpha, \beta) \in \{C, D\}$. We describe the (symmetric pure) Bayesian Nash equilibrium condition in terms of payoff parameters $a, b, c, d$ and a type density $f$.\(^\text{17}\)

Let $\mu_C = \phi(s(\alpha, \beta) = C)$ denote the consistent belief that a player chooses $C$. Then, a player of type $(\alpha, \beta)$ 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:

$$\beta > \frac{1 - \mu_C}{\mu_C} \alpha + \frac{1 - \mu_C}{\mu_C} \frac{d - c}{b - c} + \frac{b - a}{b - c} : = \tilde{H}(\alpha | \mu_C). \quad (28)$$

\(^{16}\)See Abe, Kobayashi and Suehiro (2012) for a complete analysis.

\(^{17}\)See also Duffy and Muñoz-García (2011).
From this, we can derive a fixed point characterization of symmetric equilibria. The belief \( \mu_C \) supports an equilibrium if and only if:

\[
\mu_C = \phi \left( \beta > \tilde{H}(\alpha|\mu_C) \right). \tag{29}
\]

From the linear relation between \( \alpha \) and \( \beta \) in (28), it is straightforward to verify that the right hand side of (29) takes zero for all \( \mu_C \leq \frac{b-c+d-c}{a+c+b-c+d-c} =: \mu_C^* \) and monotonically reaches \( \phi(\beta > \beta^*) \) as \( \mu_C \) goes to 1, where \( \beta^* = \frac{b-a}{b-c} \) as in Lemma 1. 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 \( \mu_C \) supports an equilibrium, then \( \mu_C > \mu_C^* \), and (3) if there exists a positive \( \mu_C \) that supports an equilibrium, then there generically exists at least one other positive belief \( \mu_C' \) that supports another equilibrium. Figure 10 and 11 below illustrate these results.

Figure 10 displays the best response types given by equation (28). The blue area describes the set of types who optimally chooses \( C \) under the belief \( \mu_C \). Combined with \( \phi \), a belief \( \mu_C \) induces a probability that a player chooses \( C \) given \( \mu_C \), that is, \( \phi(\beta > \tilde{H}(\alpha|\mu_C)) \). Figure 11 shows two kinds of graph of \( \phi(\beta > \tilde{H}(\alpha|\mu_C)) \) and hence illustrates the fixed point characterizations of (29). The left graph displays the case in which no positive \( \mu_C \) supports an equilibrium, while the right graph illustrates that we have two positive beliefs as fixed points. Therefore, the cooperation is possible in a simultaneous move prisoner’s dilemma depending on the parameters \((a, b, c, d)\) and \( f \).
Furthermore, we can discuss how the cooperation by the leadership differs from the cooperation without leadership. In a simultaneous move prisoner’s dilemma, if $\phi(\beta > \beta^*) \leq \mu_C$, we never have an equilibrium in which a player cooperates, which is illustrated by the left graph of Figure 11. On the other hand, Theorem 1 states that if $\phi(\beta > \beta^*) > \frac{d-c}{a-c}$, there exists a three-mode equilibrium in a (endogenous timing) PD game. These results imply that if game parameters satisfy $\frac{d-c}{a-c} < \phi(\beta > \beta^*) \leq \mu_C$, then the cooperation outcome $(C, C)$ is realized as an equilibrium outcome with a positive probability in the (endogenous timing) PD game with the parameters, whereas we never observe the cooperation in the simultaneous move prisoner’s dilemma with the same parameters. This provides a case in which a prisoner’s dilemma with a chance of voluntary move is superior in achieving the cooperation to a prisoner’s dilemma without a chance of voluntary moves. This case may suggest that working face to face is different from working isolatedly in light of advancing the cooperation among the members. We leave a comprehensive understanding of this issue for future research.

Appendix

Proof of Lemma 1

The condition (5) means that $a_C(\alpha, \beta) = C$ if

$$u(\alpha, \beta)x(C, C) = a > b - \beta(b - c) = u(\alpha, \beta)x(D, C).$$

This is the condition in Lemma 1. The condition (6) implies that $a_D(\alpha, \beta) = D$ because

$$u(\alpha, \beta)x(C, D) = c - \alpha(b - c) < d = u(\alpha, \beta)x(D, D)$$

for every $(\alpha, \beta)$. For the condition (7), the consistent belief $\phi_0$ must be $\phi_0(T_{CDD}(b) \cup T_{DDD}(b)) = 1$ under a three-mode strategy. Then, the condition (7) implies that $f(\alpha, \beta) \leq \frac{d-c}{a-c}$, 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.

---

To state it by treating payoff parameters and a type distribution separately, if $\frac{d-c}{a-c} < \mu_C$, 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.

---

Appendix

Proof of Lemma 1

The condition (5) means that $a_C(\alpha, \beta) = C$ if

$$u(\alpha, \beta)x(C, C) = a > b - \beta(b - c) = u(\alpha, \beta)x(D, C).$$

This is the condition in Lemma 1. The condition (6) implies that $a_D(\alpha, \beta) = D$ because

$$u(\alpha, \beta)x(C, D) = c - \alpha(b - c) < d = u(\alpha, \beta)x(D, D)$$

for every $(\alpha, \beta)$. For the condition (7), the consistent belief $\phi_0$ must be $\phi_0(T_{CDD}(b) \cup T_{DDD}(b)) = 1$ under a three-mode strategy. Then, the condition (7) implies that $f(\alpha, \beta) \leq \frac{d-c}{a-c}$, 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.

---

To state it by treating payoff parameters and a type distribution separately, if $\frac{d-c}{a-c} < \mu_C$, 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.
Proof of Theorem 1

By Lemma 4, it is enough to show that if $\phi(\beta > \beta^*) > \frac{d - \zeta}{c - \zeta}$, then there exists $\mu \in \Delta$ such that

1. $\mu \in \Psi(\mu)$
2. $\mu_C > 0, \mu_{CDD} > 0, \mu_{DDD} > 0$.

**[Step 1]** We characterize the best response types $T^*_C(\mu)$, $T^*_{CDD}(\mu)$, and $T^*_{DDD}(\mu)$. For $\mu$ with $0 < \mu_C < 1$, we showed in the text that the best response types are given by (16), (17), and (18). Consider $\mu$ with $\mu_C = 1$. Then, all the types are in different between $C$ and $CDD$ against this $\mu$ because $U(\alpha, \beta)(CDD, \mu) = U(\alpha, \beta)(CDD, \mu) = a$. Therefore, the best response types are determined by the threshold $\beta^*$ for preference over $C$ and $DDD$ and we have

- $T^*_C(\mu) = T \cap \{(\alpha, \beta) | \beta \geq \beta^*\}$
- $T^*_{CDD}(\mu) = T \cap \{(\alpha, \beta) | \beta \geq \beta^*\}$
- $T^*_{DDD}(\mu) = T \cap \{(\alpha, \beta) | \beta \leq \beta^*\}$

Consider $\mu$ with $\mu_C = 0$. Then, all the types are in different between $CDD$ and $DDD$ against this $\mu$ because $U(\alpha, \beta)(CDD, \mu) = U(\alpha, \beta)(DDD, \mu) = d$. Therefore, the best response types are determined by the threshold $\alpha^*(\mu)$ for preference over $C$ and $CDD$ and we have

- $T^*_C(\mu) = T \cap \{(\alpha, \beta) | \alpha \leq \alpha^*(\mu)\}$
- $T^*_{CDD}(\mu) = T \cap \{(\alpha, \beta) | \alpha \geq \alpha^*(\mu)\}$
- $T^*_{DDD}(\mu) = T \cap \{(\alpha, \beta) | \alpha \geq \alpha^*(\mu)\}$

**[Step 2]** With the characterization of the best response types, we develop a characterization of $\Psi(\mu)$. For each $\mu \in \Delta$ and $\mu' \in \Psi(\mu)$, there is an associated $s : T \to \{C, CDD, DDD\}$ that generates $\mu'$ given $\mu$. It is

$$s(\alpha, \beta) = \begin{cases} 
C & \text{for } (\alpha, \beta) \in T^*_C(\mu) \setminus T^*_{CDD}(\mu), T^*_{DDD}(\mu) \\
CDD & \text{for } (\alpha, \beta) \in T^*_{CDD}(\mu) \setminus T^*_C(\mu), T^*_{DDD}(\mu) \\
DDD & \text{for } (\alpha, \beta) \in T^*_{DDD}(\mu) \setminus T^*_C(\mu), T^*_{CDD}(\mu) \\
C \text{ or } CDD & \text{for } (\alpha, \beta) \in T^*_C(\mu) \cap T^*_{CDD}(\mu) \setminus T^*_{DDD}(\mu) \\
CDD \text{ or } DDD & \text{for } (\alpha, \beta) \in T^*_{CDD}(\mu) \cap T^*_{DDD}(\mu) \setminus T^*_C(\mu) \\
C \text{ or } DDD & \text{for } (\alpha, \beta) \in T^*_C(\mu) \cap T^*_{DDD}(\mu) \setminus T^*_{CDD}(\mu) \\
C, CDD \text{ or } DDD & \text{for } (\alpha, \beta) \in T^*_C(\mu) \cap T^*_{CDD}(\mu) \cap T^*_{DDD}(\mu).
\end{cases}$$

Then, we can characterize $\Psi(\mu)$ as follows.

1. First, consider $\mu$ with $0 < \mu_C < 1$. Then, the best response types are given by (16), (17), and (18) so that we have

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

35
This implies that \( \Psi(\mu) = \{ \mu' = (\mu'_C, \mu'_{CDD}, \mu'_{DDD}) = (\phi(T^*_C(\mu)), \phi(T^*_C(\mu)), \phi(T^*_D(\mu))) \} \).

(2) Second, consider \( \mu \) with \( \mu_C = 1 \). Then, the best response types are given by (27), (28), and (29) so that we have

\[
T^*_c(\mu) = T^*_CDD(\mu) = T \cap \{ (\alpha, \beta)|\beta \geq \beta^* \}
\]

\[
T^*_C(\mu) \cap T^*_D(\mu) = T^*_CDD(\mu) \cap T^*_DDD(\mu) = T \cap \{ (\alpha, \beta)|\alpha = \alpha^*(\mu) \}.
\]

This implies that \( \Psi(\mu) \) is a line such that

\[
\Psi(\mu) = \{ \mu' = (\mu'_C, \mu'_{CDD}, \mu'_{DDD}) \in \Delta | \mu'_C = \phi(\alpha < \alpha^*(\mu)), \mu'_{CDD} + \mu'_{DDD} = \phi(\alpha > \alpha^*(\mu)) \}.
\]

(3) Third, consider \( \mu \) with \( \mu_C = 0 \). Then, the best response types are given by (30), (31), and (32) so that we have

\[
T^*_CDD(\mu) = T^*_DDD(\mu) = T \cap \{ (\alpha, \beta)|\alpha \geq \alpha^*(\mu) \}
\]

\[
T^*_C(\mu) \cap T^*_D(\mu) = T^*_CDD(\mu) \cap T^*_DDD(\mu) = T \cap \{ (\alpha, \beta)|\alpha = \alpha^*(\mu) \}.
\]

This implies that \( \Psi(\mu) \) is a line (or a point when \( \phi(\alpha > \alpha^*(\mu)) = 0 \)) such that

\[
\Psi(\mu) = \{ \mu' = (\mu'_C, \mu'_{CDD}, \mu'_{DDD}) \in \Delta | \mu'_C = \phi(\alpha < \alpha^*(\mu)), \mu'_{CDD} + \mu'_{DDD} = \phi(\alpha > \alpha^*(\mu)) \}.
\]

**[Step 3]** Define a restriction \( \tilde{\Psi} : \Delta \to \Delta \) of \( \Psi \) by

\[
\tilde{\Psi}(\mu) = \Psi(\mu) \cap \{ \mu' = (\mu'_C, \mu'_{CDD}, \mu'_{DDD}) \in \Delta | 0 \leq \mu'_C \leq \phi(\beta > \beta^*), 0 \leq \mu'_D \leq \phi(\beta < \beta^*) \}.
\]

Then, by Step 1, \( \Psi(\mu) \) is nonempty for each \( \mu \in \Delta \) because, for \( \mu \) with \( 0 < \mu_C < 1 \), the best response types (17) and (18) imply

\[
\mu'_C = \phi(T^*_CDD(\mu)) \leq \phi(\beta > \beta^*)
\]

\[
\mu'_D = \phi(T^*_DDD(\mu)) \leq \phi(\beta < \beta^*).
\]

Step 2 also tells that \( \tilde{\Psi}(\mu) \) is convex and compact for each \( \mu \in \Delta \). Furthermore, the correspondence \( \tilde{\Psi} : \Delta \to \Delta \) is upper hemi continuous. (The prolonged proof is postponed to Step 5 below.) Then, by Kakutani’s fixed point theorem, there exists \( \bar{\mu} = (\bar{\mu}_C, \bar{\mu}_{CDD}, \bar{\mu}_{DDD}) \in \Delta \) such that \( \bar{\mu} \in \tilde{\Psi}(\bar{\mu}) \).

**[Step 4]** The fixed point \( \bar{\mu} \) of \( \tilde{\Psi} \) satisfies Condition 1 that \( \bar{\mu} \in \Psi(\bar{\mu}) \), because \( \Psi(\bar{\mu}) \subset \Psi(\bar{\mu}) \). We will show that under the assumption \( \phi(\beta > \beta^*) > \frac{d-c}{a-d} \), it satisfies Condition 2 that \( \bar{\mu}_C > 0, \bar{\mu}_{CDD} > 0 \), and \( \bar{\mu}_{DDD} > 0 \).

First, we will show that \( \bar{\mu}_C > 0 \). Suppose to the contrary that \( \bar{\mu}_C = 0 \). Then, \( \bar{\mu} = (0, \phi(\beta > \beta^*), \phi(\beta < \beta^*)) \) because \( \bar{\mu} = (\bar{\mu}_C, \bar{\mu}_{CDD}, \bar{\mu}_{DDD}) \in \Psi(\bar{\mu}) \) implies a restriction such that \( \bar{\mu}_{CDD} \leq \phi(\beta > \beta^*) \) and \( \bar{\mu}_{DDD} \leq \phi(\beta < \beta^*) \). Under the assumption \( \phi(\beta > \beta^*) > \frac{d-c}{a-d} \), we have

\[
\frac{d-c}{a-d} < \frac{\phi(\beta > \beta^*)}{\phi(\beta < \beta^*)} = \frac{\bar{\mu}_{CDD}}{\bar{\mu}_{DDD}}.
\]

This means that the left inequality in (14) holds. Therefore, \( \alpha^*(\bar{\mu}) > 0 \). The case (3) in Step 2 tells that \( \bar{\mu}_C = \phi(\alpha < \alpha^*(\mu)) \) when \( \bar{\mu} \in \Psi(\bar{\mu}) \subset \Psi(\bar{\mu}) = \Psi((0, \phi(\beta > \beta^*), \phi(\beta < \beta^*))) \). Hence, \( \bar{\mu}_C > 0 \). This is a contradiction.
Second, we will show that \( \bar{\mu}_C < 1 \). Suppose to the contrary that \( \bar{\mu}_C = 1 \). Then, by the case (2) of Step 2, \( \bar{\mu} = (\bar{\mu}_C, \bar{\mu}_{CDD}, \bar{\mu}_{DDD}) \in \bar{\Psi}(\bar{\mu}) \subseteq \bar{\Psi}(\bar{\mu}) = \bar{\Psi}((1,0,0)) \) means \( \bar{\mu}_{DDD} = \phi(\beta < \beta^*) > 0 \). Therefore, \((1,0,0) \notin \bar{\Psi}(\bar{\mu})\). This is a contradiction.

Third, we will show that \( \bar{\mu}_{DDD} > 0 \). Suppose to the contrary that \( \bar{\mu}_{DDD} = 0 \). Step 2 tells that when \( 0 < \bar{\mu}_C < 1 \), \( \bar{\mu} = (\bar{\mu}_C, \bar{\mu}_{CDD}, \bar{\mu}_{DDD}) \in \bar{\Psi}(\bar{\mu}) \subseteq \bar{\Psi}(\bar{\mu}) \) means \( \bar{\mu}_{CDD} = \phi(T_{CDD}(\bar{\mu})) \). When \( \bar{\mu}_{DDD} = 0 \), we have \( \alpha^*(\bar{\mu}) = \infty \). Therefore, the best response types (17) imply \( T_{CDD}(\bar{\mu}) = 0 \). Hence, \( \bar{\mu}_{CDD} = 0 \). However, when \( \bar{\mu}_{DDD} = 0 \), the second conclusion \( \bar{\mu}_C < 1 \) implies \( 0 < \bar{\mu}_{CDD} \). This is a contradiction.

Finally, we will show that \( \bar{\mu}_{CDD} > 0 \). Suppose to the contrary that \( \bar{\mu}_{CDD} = 0 \). Step 2 tells that when \( 0 < \bar{\mu}_C < 1 \), \( \bar{\mu} = (\bar{\mu}_C, \bar{\mu}_{CDD}, \bar{\mu}_{DDD}) \in \bar{\Psi}(\bar{\mu}) \subseteq \bar{\Psi}(\bar{\mu}) \) means \( \bar{\mu}_C = \phi(T_C(\bar{\mu})) \). It follows from \( \bar{\mu}_{CDD} = 0 \) and \( \bar{\mu}_{DDD} > 0 \) that \( \alpha^*(\bar{\mu}) < 0 \). Therefore, the best response types (16) imply \( T^n_C(\bar{\mu}) = \emptyset \). Hence, \( \bar{\mu}_C = 0 \). This contradicts to the first conclusion \( \bar{\mu}_C > 0 \).

**[Step 5]** To complete the proof, we will prove the claim in Step 3 that the correspondence \( \bar{\Psi} : \Delta \rightarrow \Delta \) is upper hemi continuous. Consider a sequence \( \{(\mu^n, \hat{\mu}^n)\} \) such that

1. \( \mu^n = (\mu^n_C, \mu^n_{CDD}, \mu^n_{DDD}) \in \Delta \)
2. \( \hat{\mu}^n = (\hat{\mu}_C^n, \hat{\mu}_{CDD}^n, \hat{\mu}_{DDD}^n) \in \bar{\Psi}(\mu) \)
3. \( \exists \mu = (\mu_C, \mu_{CDD}, \mu_{DDD}) = \lim_{n \rightarrow \infty} \mu^n \in \Delta \)
4. \( \exists \hat{\mu} = (\hat{\mu}_C, \hat{\mu}_{CDD}, \hat{\mu}_{DDD}) = \lim_{n \rightarrow \infty} \hat{\mu}^n \in \Delta. \)

For this sequence, we will show that \( \hat{\mu} \in \bar{\Psi}(\mu) \).

First, consider a case in which \( 0 < \mu_C < 1 \). Then, there exists \( N \) such that \( 0 < \mu^n_C < 1 \) for any \( n \geq N \). From Step 2, when \( 0 < \mu^n_C < 1 \), \( \hat{\mu}^n \in \bar{\Psi}(\mu^n) \subseteq \bar{\Psi}(\mu^n) \) means that

\[
\hat{\mu}^n = (\hat{\mu}_C^n, \hat{\mu}_{CDD}^n, \hat{\mu}_{DDD}^n) = (\phi(T^n_C(\mu^n)), \phi(T^n_{CDD}(\mu^n)), \phi(T^n_{DDD}(\mu^n))).
\]

The correspondences \( T^n_C(\mu), T^n_{CDD}(\mu), \) and \( T^n_{DDD}(\mu) \) are continuous in \( \mu \) and the probability measure \( \phi(B) \) is continuous in \( B \) where the topology of Borel sets in \( \Delta \) is the Housdorff topology. Therefore, the composite functions \( \phi(T^n_C(\mu)), \phi(T^n_{CDD}(\mu)), \) and \( \phi(T^n_{DDD}(\mu)) \) are continuous in \( \mu \). Hence, we have

\[
\hat{\mu} = \lim_{n \rightarrow \infty} \hat{\mu}^n = \lim_{n \rightarrow \infty} (\phi(T^n_C(\mu^n)), \phi(T^n_{CDD}(\mu^n)), \phi(T^n_{DDD}(\mu^n))) = (\phi(T^n_C(\lim_{n \rightarrow \infty} \mu^n)), \phi(T^n_{CDD}(\lim_{n \rightarrow \infty} \mu^n)), \phi(T^n_{DDD}(\lim_{n \rightarrow \infty} \mu^n))) = (\phi(T^n_C(\mu)), \phi(T^n_{CDD}(\mu)), \phi(T^n_{DDD}(\mu))).
\]

From Step 2, this means that \( \hat{\mu} \in \bar{\Psi}(\mu) \) because \( 0 < \mu_C < 1 \). Furthermore, \( \hat{\mu}^n \in \bar{\Psi}(\mu^n) \) implies that \( \hat{\mu}^n_{DDD} \leq \phi(\beta > \beta^*) \) and \( \hat{\mu}^n_{CDD} \leq \phi(\beta > \beta^*) \). Therefore, it holds that \( \hat{\mu}_{CDD} = \lim_{n \rightarrow \infty} \hat{\mu}^n_{CDD} \leq \phi(\beta > \beta^*) \) and \( \hat{\mu}_{DDD} = \lim_{n \rightarrow \infty} \hat{\mu}^n_{DDD} \leq \phi(\beta < \beta^*). \) Hence, we have \( \hat{\mu} \in \bar{\Psi}(\mu) \).

Second, consider a case in which \( \mu_C = 1 \). Then, there exists \( N \) such that \( \mu^n_C > 0 \) for any \( n \geq N \). Take \( n \geq N \). Consider \((\alpha, \beta) \in T\) such that \( \beta < \frac{b - a}{\beta - \mu^n_C} - \frac{\mu^n_{CDD} a - d}{\beta - \mu^n_C} \). Then \( \beta < \beta^* \) and \( \beta < H(\alpha|\mu^n) \). By the best response types (18) or (29), this means that \( (\alpha, \beta) \in T^n_{DDD}(\mu^n) \). Hence, \( \phi(\beta < \frac{b - a}{\beta - \mu^n_C} - \frac{\mu^n_{CDD} a - d}{\beta - \mu^n_C}) \leq \phi(T^n_{DDD}(\mu^n)) < \phi(\beta < \beta^*). \) By Step 2, \( \hat{\mu} \in \bar{\Psi}(\mu) \subseteq \bar{\Psi}(\mu) \) means that either \( \hat{\mu}^n_{DDD} = \phi(T^n_{DDD}(\mu^n)) \) or \( \hat{\mu}^n_{DDD} = \phi(T^n_{DDD}(\mu^n)) \).
\( \phi(\beta < \beta') \) holds when \( \mu_C^n > 0 \). Hence, \( \phi(\beta < \frac{b-a}{b-c} - \frac{\mu^n_D}{\mu^n_C} \frac{a-d}{b-c}) \leq \mu^n_DDD \leq \phi(\beta < \beta') \).

When \( \mu_C = 1 \), \( \lim_{n \to \infty} \mu^n_{CD} = 0 \). Therefore, \( \mu_DDD = \lim_{n \to \infty} \mu^n_DDD = \phi(\beta < \beta') \). By Step 2, this means that \( \hat{\mu} \in \Psi(\mu) \). Therefore, as for the case of \( 0 < \mu_C < 1 \), we conclude that \( \hat{\mu} \in \Psi(\mu) \).

Third, consider a case in which \( \mu_C = 0 \). In light of the condition (14), let us divide this case into the following subcases.

1. \( \frac{\mu_{CDD}}{\mu_{DDD}} < \frac{d-c}{a-d} \) (therefore, \( \alpha^*(\mu) < 0 \))

2. \( \frac{d-c}{a-d} \leq \frac{\mu_{CDD}}{\mu_{DDD}} \leq \frac{d-c}{a-d} + \frac{b-c}{a-d} \hat{\alpha} \) (therefore, \( 0 \leq \alpha^*(\mu) \leq \hat{\alpha} \))

3. \( \frac{d-c}{a-d} + \frac{b-c}{a-d} \hat{\alpha} < \frac{\mu_{CDD}}{\mu_{DDD}} \) (therefore, \( \hat{\alpha} < \alpha^*(\mu) \))

Consider Subcase 1. Define

\[ \epsilon = \left[ |\mu_{CDD}d + \mu_{DDD}| - |\mu_{CDD}a + \mu_{DDD}| \right] = \mu_{DDD}(a-d) \left( \frac{d-c}{a-d} - \frac{\mu_{CDD}}{\mu_{DDD}} \right) > 0. \]

When \( \lim_{n \to \infty} \mu^n_C = 0 \), \( \lim_{n \to \infty} \mu^n_{CDD} = \mu_{CDD} \), and \( \lim_{n \to \infty} \mu^n_DDD = \mu_{DDD} \), there exists \( N \) such that \( \mu^n_C < \frac{\mu^n_{CDD}}{\mu^n_{DDD}} \) and \( (\mu^n_{CDD} - \mu_{CDD})(a-d) - (\mu^n_{DDD} - \mu_{DDD})(d-c) < \frac{\epsilon}{2} \) for any \( n \geq N \). Then,

\[
U_{(\alpha, \beta)}(D, \mu^n) - U_{(\alpha, \beta)}(C, \mu^n) = \\
= \mu^n_C[b - \beta(b-c) - a] + [\mu^n_{CDD}(a-d) + \mu^n_{DDD}(d-c)] + \mu^n_{DDD}a(b-c) \\
\geq \mu^n_C(c-a) + \left\{ -\mu^n_{CDD}(a-d) + \mu^n_{DDD}(d-c) \right\} - \left\{ -\mu_{CDD}(a-d) + \mu_{DDD}(d-c) \right\} \\
+ \left\{ -\mu_{CDD}(a-d) + \mu_{DDD}(d-c) \right\} \\
> \frac{\epsilon}{2} - \frac{\epsilon}{2} + \epsilon \\
= 0
\]

for any \( n \geq N \) where the first inequality follows from \( b - \beta(b-c) \geq c \) (0 \( \leq \beta \leq 1 \)) and \( \alpha(b-c) \geq 0 \). This means that \( T^n_C(\mu^n) = 0 \) for any \( n \geq N \). Therefore, by Step 1, \( \hat{\mu} \in \Psi(\mu^n) \subseteq \Psi(\mu^n) \) means \( \hat{\mu} = 0 \) for any \( n \geq N \). Hence, \( \mu_C = \lim_{n \to \infty} \mu^n_C = 0 \). By Step 1, when \( \mu_C = 0 \) and \( \alpha^*(\mu) < 0 \), this means that \( \hat{\mu} \in \Psi(\mu) \). Therefore, as for the case of \( 0 < \mu_C < 1 \), we conclude that \( \hat{\mu} \in \Psi(\mu) \).

For Subcase 2, a similar argument establishes \( \hat{\mu}_C = \lim_{n \to \infty} \hat{\mu}_C = 1 \) and leads to a conclusion that \( \hat{\mu} \in \Psi(\mu) \).

Finally, consider Subcase 2. Fix \( \epsilon > 0 \) arbitrarily small. Then, there exists \( N' \) such that \( \alpha^*(\mu) - \frac{\epsilon}{2} < \alpha^*(\mu^n) \), \( |\mu_{DDD} - \mu^n_{DDD}| < \frac{\epsilon}{2} \mu_{DDD} \), and \( \mu_C < \mu_C(1 - \frac{\epsilon}{2}) \frac{b-a}{b-c} \) for any \( n \geq N' \) where \( \mu_{DDD} > 0 \) in Subcase 2 allows the second and third requirements.

For \( n \geq N' \), we have

\[
U_{(\alpha^*(\mu)-\epsilon, \beta)}(C, \mu^n) - U_{(\alpha^*(\mu)-\epsilon, \beta)}(CDD, \mu^n) = \\
= \left[ U_{(\alpha^*(\mu)-\epsilon, \beta)}(C, \mu^n) - U_{(\alpha^*(\mu)-\epsilon, \beta)}(CDD, \mu^n) \right] + \left[ U_{(\alpha^*(\mu)-\epsilon, \beta)}(CDD, \mu^n) - U_{(\alpha^*(\mu)-\epsilon, \beta)}(CDD, \mu^n) \right] \\
= \mu^n_{DDD}(b-c)[\alpha^*(\mu^n) - (\alpha^*(\mu) - \epsilon)] \\
> \mu_{DDD}(1 - \frac{\epsilon}{2}) (b-c) \frac{\epsilon}{2} \\
> 0
\]
where the second braces in the second line is zero by the definition of $\alpha^*(\mu^n)$. From this inequality, we also have

$$ U_{(\alpha^*(\mu)-\epsilon,\beta)}(C,\mu^n) - U_{(\alpha^*(\mu)-\epsilon,\beta)}(DDD,\mu^n) $$

$$ = \left[ U_{(\alpha^*(\mu)-\epsilon,\beta)}(C,\mu^n) - U_{(\alpha^*(\mu)-\epsilon,\beta)}(CDD,\mu^n) \right] + \left[ U_{(\alpha^*(\mu),\beta)}(CDD,\mu^n) - U_{(\alpha^*(\mu),\beta)}(DDD,\mu^n) \right] $$

$$ > \mu_{DDD}(1-\frac{\epsilon}{2})(b-c) + \mu_{DDD}^e(a - (b - \beta(b-c))) $$

$$ \geq \mu_{DDD}(1-\frac{\epsilon}{2})(b-c) - \mu_{DDD}^e(b-a) $$

$$ > 0. $$

This set of inequalities guarantee that

$$ U_{(\alpha,\beta)}(C,\mu^n) - U_{(\alpha,\beta)}(CDD,\mu^n) \geq U_{(\alpha^*(\mu)-\epsilon,\beta)}(C,\mu^n) - U_{(\alpha^*(\mu)-\epsilon,\beta)}(CDD,\mu^n) > 0 $$

$$ U_{(\alpha,\beta)}(C,\mu^n) - U_{(\alpha,\beta)}(DDD,\mu^n) \geq U_{(\alpha^*(\mu)-\epsilon,\beta)}(C,\mu^n) - U_{(\alpha^*(\mu)-\epsilon,\beta)}(DDD,\mu^n) > 0 $$

for any $\alpha < \alpha^*(\mu) - \epsilon$ and for any $n \geq N'$. This means that $\alpha^*(\mu)$ is a three-mode equilibrium in $\mathbb{T}^3_{C}(\mu^n)$ for all $(\alpha, \beta) \in T_{C}(\mu^n)$ with $\alpha < \alpha^*(\mu) - \epsilon$ and for any $n \geq N'$. 

For the same $\epsilon > 0$, there exists $N''$ such that $\alpha^*(\mu^n) < \alpha^*(\mu) + \frac{\xi}{2}$ and $|\mu_{DDD} - \mu_{DDD}^b| < \frac{\xi}{2} \mu_{DDD}$ for any $n \geq N''$. For $n \geq N''$, we have

$$ U_{(\alpha^*(\mu)+\epsilon,\beta)}(CDD,\mu^n) - U_{(\alpha^*(\mu)+\epsilon,\beta)}(C,\mu^n) $$

$$ = \mu_{DDD}^b(b-c)[(\alpha^*(\mu) + \epsilon) - \alpha^*(\mu^n)] $$

$$ > \mu_{DDD}(1-\frac{\epsilon}{2})(b-c) \frac{\epsilon}{2} $$

$$ > 0. $$

This inequality guarantees

$$ U_{(\alpha,\beta)}(CDD,\mu^n) - U_{(\alpha,\beta)}(C,\mu^n) \geq U_{(\alpha^*(\mu)+\epsilon,\beta)}(CDD,\mu^n) - U_{(\alpha^*(\mu)+\epsilon,\beta)}(C,\mu^n) > 0 $$

for any $\alpha > \alpha^*(\mu) + \epsilon$ and for any $n \geq N''$. This means that $(\alpha, \beta) \notin T_{C}(\mu^n)$ for all $(\alpha, \beta) \in T_{C}(\mu^n)$ with $\alpha > \alpha^*(\mu) + \epsilon$. Therefore, we have $\lim_{n \rightarrow \infty} \phi(T_{C}(\mu^n)) = \phi(\alpha < \alpha^*(\mu))$. Then, by Step 2, we have $\hat{\mu} \in \hat{\Psi}(\mu)$ and conclude that $\hat{\mu} \in \hat{\Psi}(\mu)$. This completes the proof. (Q.E.D.)

**Proof of Theorem 2**

Fix $f$ and $a, b, c$ with $b > a > c$. Take $d, d'$ such that $c < d' < d < a$. Suppose that there exists a three-mode equilibrium in $PD((a, b, c, d), f)$. Let $\mu^* = (\mu_C^*, \mu_{CDD}^*, \mu_{DDD}^*)$ denote the corresponding distribution of behavior modes. We show that there also exists a three-mode equilibrium in $PD((a, b, c, d'), f)$.

To distinguish the analysis of $PD((a, b, c, d), f)$ and the analysis of $PD((a, b, c, d'), f)$, let us put $d$ and $d'$ as superscripts to relevant notations. For example, when we consider the correspondence $\Psi : \Delta \rightarrow \Delta$ for the analysis of $PD((a, b, c, d), f)$, we write it as $\Psi^d$. As we noted in the text, for a $\mu = (\mu_C, \mu_{CDD}, \mu_{DDD}) \in \Delta$ with $0 < \mu_C < 1$, its image of $\Psi^d$ is a singleton $\Psi^d(\mu) = \{(\phi(T^a_C(\mu)), \phi(T^{a}_CDD(\mu)), \phi(T^a_{DDD}(\mu)))\}$. In this case, we denote $\psi^d_C(\mu) = \phi(T^a_C(\mu)), \psi^d_{CDD}(\mu) = \phi(T^{a}_{CDD}(\mu)), \psi^d_{DDD}(\mu) = \phi(T^a_{DDD}(\mu))$, and use $\psi^d_C(\mu), \psi^d_{CDD}(\mu), \psi^d_{DDD}(\mu)$ instead of $\Psi^d(\mu)$.

**Step 1** Consider $\mu = (\mu_C, \mu_{CDD}, \mu_{DDD}), \bar{\mu} = (\bar{\mu}_C, \bar{\mu}_{CDD}, \bar{\mu}_{DDD}) \in \Delta$. We show that if $\mu_C = \bar{\mu}_C, \mu_{CDD} < \bar{\mu}_{CDD}$, and $\mu_{DDD} > \bar{\mu}_{DDD}$, then $T^d_C(\mu) \subseteq T^d_C(\bar{\mu})$. Take $(\alpha, \beta) \in \Delta$. To...
Then, together with $U^d_{(a,\beta)}(C, \mu) = U^d_{(a,\beta)}(CDD, \mu)$ and $U^d_{(a,\beta)}(C, \mu) \geq U^d_{(a,\beta)}(CDD, \mu)$. Then, together with $\mu_C = \bar{\mu}_C$, $\mu_{CDD} < \bar{\mu}_{CDD}$, and $\mu_{DDD} > \bar{\mu}_{DDD}$, we have

$$U^d_{(a,\beta)}(C, \bar{\mu}) - U^d_{(a,\beta)}(CDD, \bar{\mu}) = \mu_C(a - a) + \mu_{CDD}(a - d) + \mu_{DDD}\{c - \alpha(b - c)\} - d$$

$$\geq \mu_C(a - a) + \mu_{CDD}(a - d) + \mu_{DDD}\{c - \alpha(b - c)\} - d$$

$$= U^d_{(a,\beta)}(C, \mu) - U^d_{(a,\beta)}(CDD, \mu) \geq 0$$

$$U^d_{(a,\beta)}(C, \bar{\mu}) - U^d_{(a,\beta)}(DDD, \bar{\mu}) = \mu_C(a - [b - \beta(b - c)]) + \mu_{CDD}(a - d) + \mu_{DDD}\{c - \alpha(b - c)\} - d$$

$$\geq \mu_C(a - [b - \beta(b - c)]) + \mu_{CDD}(a - d) + \mu_{DDD}\{c - \alpha(b - c)\} - d$$

$$= U^d_{(a,\beta)}(C, \mu) - U^d_{(a,\beta)}(DDD, \mu) \geq 0.$$

Therefore, $(\alpha, \beta) \in T^d_{C\delta}(\bar{\mu})$. Hence, $T^d_{C\delta}(\mu) \subseteq T^d_{C\delta}(\bar{\mu})$.

Suppose in addition that $\emptyset \neq T^d_{C\delta}(\mu) \subseteq \Delta$. Then either $T^d_{C\delta}(\mu) \cap T^d_{C\delta}(\bar{\mu}) = \emptyset$ or $T^d_{C\delta}(\mu) \cap T^d_{C\delta}(\bar{\mu}) = \emptyset$ because $T^d_{C\delta}(\mu)$, $T^d_{C\delta}(\bar{\mu})$, and $T^d_{C\delta}(\mu)$ are closed subsets in $\Delta$. If $(\alpha, \beta) \in T^d_{C\delta}(\mu) \cap T^d_{C\delta}(\bar{\mu})$, then it follows from the above argument that $(\alpha, \beta) \in T^d_{C\delta}(\bar{\mu}) \backslash T^d_{C\delta}(\mu)$ for any $(\alpha, \beta)$ in some neighborhood of $(\alpha, \beta)$. Therefore, $T^d_{C\delta}(\mu) \subseteq T^d_{C\delta}(\bar{\mu})$. If $(\alpha, \beta) \in T^d_{C\delta}(\mu) \cap T^d_{C\delta}(\mu)$, then it follows similarly that $T^d_{C\delta}(\mu) \subseteq T^d_{C\delta}(\mu)$.

Suppose in addition that $0 < \mu_C, \bar{\mu}_C < 1$ and replace the supposition $\emptyset \neq T^d_{C\delta}(\mu) \subseteq \Delta$ with an alternative supposition that $0 < \psi^d_{C}(\mu) < 1$. Then, $0 < \psi^d_{C}(\mu) < 1$ implies $\emptyset \neq T^d_{C\delta}(\mu) \subseteq \Delta$ and so, by the above argument, we have $T^d_{C\delta}(\mu) \subseteq T^d_{C\delta}(\bar{\mu})$. This implies that $\psi^d_{C}(\mu) < \psi^d_{C}(\bar{\mu})$ because $T^d_{C\delta}(\mu)$ and $T^d_{C\delta}(\bar{\mu})$ are closed subsets in $\Delta$.

**Step 2** Consider $\mu = (\mu_C, \mu_{CDD}, \mu_{DDD}), \mu' = (\mu'_C, \mu'_{CDD}, \mu'_{DDD}) \subseteq \Delta$ with $0 < \mu_C, \mu'_C < 1$. Suppose that $\mu_C = \mu'_C, \psi^d_{C}(\mu) = \mu_C$, and $\psi^d_{C}(\mu') = \mu'_C$. Then, we show that it must be the case that $\mu_{CDD} > \mu'_{CDD}$ and $\mu_{DDD} > \mu'_{DDD}$.

Suppose to the contrary that $\mu_{CDD} \leq \mu'_{CDD}$ and $\mu_{DDD} \leq \mu'_{DDD}$. Then, by setting $\bar{\mu} = \mu'$ in Step 1, we must have

$$\psi^d_{C}(\mu') \geq \psi^d_{C}(\mu) = \mu_C = \mu'_C. \quad (36)$$

On the other hand, by $d > d'$, we have

$$\psi^d_{C}(\mu') > \psi^d_{C}(\mu') \quad (37)$$

for the following reason. Take $(\alpha, \beta) \in T^d_{C\delta}(\mu')$. Then, $U^d_{(a,\beta)}(C, \mu') \geq U^d_{(a,\beta)}(CDD, \mu')$ and $U^d_{(a,\beta)}(C, \mu') \geq U^d_{(a,\beta)}(DDD, \mu')$. Then, together with $d > d'$ and $\mu_C < 1$, we have

$$U^d_{(a,\beta)}(C, \mu') - U^d_{(a,\beta)}(CDD, \mu') = U^d_{(a,\beta)}(C, \mu') - U^d_{(a,\beta)}(CDD, \mu') + \mu'_{CDD}(d - d') + \mu'_{DDD}(d - d')$$

$$> U^d_{(a,\beta)}(C, \mu') - U^d_{(a,\beta)}(CDD, \mu')$$

$$\geq 0$$

$$U^d_{(a,\beta)}(C, \mu') - U^d_{(a,\beta)}(DDD, \mu') = U^d_{(a,\beta)}(C, \mu') - U^d_{(a,\beta)}(DDD, \mu') + \mu'_{CDD}(d - d') + \mu'_{DDD}(d - d')$$

$$> U^d_{(a,\beta)}(C, \mu') - U^d_{(a,\beta)}(DDD, \mu')$$

$$\geq 0.$$

Therefore, $(\alpha, \beta) \in T^d_{C\delta}(\mu')$. Hence, $T^d_{C\delta}(\mu') \subseteq T^d_{C\delta}(\mu')$. Furthermore, by the similar argument to Step 1, it follows from $0 < \psi^d_{C}(\mu') = \mu'_C < 1$ that $T^d_{C\delta}(\mu') \subseteq T^d_{C\delta}(\mu')$ and so $\psi^d_{C}(\mu') > \psi^d_{C}(\mu')$. 40
By combining (36) and (37), we have

$$\psi^d_C(\mu') > \psi^d_C(\mu') \geq \mu'_C.$$

This contradicts the supposition that $$\psi^d_C(\mu') = \mu'_C$$.  

**Step 3** Consider $$\mu = (\mu_C, \mu_{CDD}, \mu_{DDD}) \in \Delta$$ with $$0 < \mu_C, \mu_{CDD}, \mu_{DDD} < 1$$. Suppose that $$\psi^d_C(\mu) = \mu_C$$. Then, by a similar argument to Step 2, we have $$\psi^d_C(\mu) > \psi^d_C(\mu) = \mu_C$$. Note from (14) that $$T^d_C(\tilde{\mu}) = \emptyset$$ for any $$\tilde{\mu} = (\tilde{\mu}_C, \tilde{\mu}_{CDD}, \tilde{\mu}_{DDD}) \in \Delta$$ such that $$\tilde{\mu}_{CDD} \leq \frac{d-\delta}{\delta} \tilde{\mu}_{DDD}$$. Particularly, consider $$\tilde{\mu} = (\tilde{\mu}_C, \tilde{\mu}_{CDD}, \tilde{\mu}_{DDD}) \in \Delta$$ such that $$\tilde{\mu}_C = \mu_C$$ and $$\tilde{\mu}_{CDD} \leq \frac{d-\delta}{\delta} \tilde{\mu}_{DDD}$$. Then, $$0 < \tilde{\mu}_C, \tilde{\mu}_{CDD}, \tilde{\mu}_{DDD} < 1$$ and $$\psi^d_C(\tilde{\mu}) = 0$$. Furthermore, $$\mu_{CDD} > \tilde{\mu}_{CDD}$$ and $$\mu_{DDD} < \tilde{\mu}_{DDD}$$. Hence, if we define a line $$L = [\mu, \tilde{\mu}]$$ in the contrary to the contrary.  

**Step 4** Consider a case in class 1. We must have $$\psi^d_C(\mu) > \psi^d_C(\mu) = \mu_C$$. This means that $$\psi^d_C(\mu) = \mu_C$$ and $$\psi^d_C(\mu) = \mu_C$$ are common to $$T^d_C(\mu)$$, $$T^d_{CDD}(\mu)$$, $$T^d_{DDD}(\mu)$$ and $$T^d_{CDD}(\mu)$$, $$T^d_{DDD}(\mu)$$, $$T^d_{CDD}(\mu)$$, $$T^d_{DDD}(\mu)$$.

<table>
<thead>
<tr>
<th>$$\alpha^d(\mu^*) / \alpha^d(\mu')$$</th>
<th>$$\alpha^d(\mu') &gt; \beta^*$$</th>
<th>$$\alpha^d(\mu') &gt; \beta^*$$</th>
<th>$$\alpha^d(\mu') &lt; \beta^*$$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$$\alpha^d(\mu') &gt; \beta^*$$</td>
<td>class 1</td>
<td>class 2</td>
<td>class 4</td>
</tr>
<tr>
<td>$$\alpha^d(\mu') &lt; \beta^*$$</td>
<td>class 1</td>
<td>class 2</td>
<td>class 4</td>
</tr>
</tbody>
</table>

Table 7: classifications of patterns in $$T^d_C(\mu^*)$$s and patterns in $$T^d_C(\mu')$$s

Consider a case in class 1. We must have $$\alpha^d(\mu^*) > \alpha^d(\mu')$$ in this case. Suppose to the contrary that $$\alpha^d(\mu^*) \leq \alpha^d(\mu')$$. Note that the slopes of (15) for $$PD((a, b, c, d), f)$$ and $$PD((a, b, c, d'), f)$$ are $$\frac{\mu_{CDD}}{\mu_C} < \frac{\mu_{DDD}}{\mu_C}$$ because $$\mu^* = \mu_C$$ and $$\mu_{DDD} < \mu_{DDD}$$. Then,

$$T^d_C(\mu^*) = T \cap \{(\alpha, \beta) | \alpha \leq \alpha^d(\mu^*), \beta \geq H^d(\alpha)|\mu^*\}$$

$$\subsetneq T \cap \{(\alpha, \beta) | \alpha \leq \alpha^d(\mu'), \beta \geq H^d(\alpha)|\mu'\}$$

$$= T^d_C(\mu').$$

This means that $$\mu^*_C = \phi(T^d_C(\mu^*)) < \phi(T^d_C(\mu')) = \mu'_C$$. This contradicts $$\mu^*_C = \mu'_C$$. 

Now, compare

$$\mu^*_{CCC} = \phi(T^d_C(\mu^*) \cap \{(\alpha, \beta) | \beta > \beta^*\}) + \phi(T^d_C(\mu') \cap \{(\alpha, \beta) | \beta < \beta^*\})$$

$$\mu'_{CCC} = \phi(T^d_C(\mu^*) \cap \{(\alpha, \beta) | \beta > \beta^*\}) + \phi(T^d_C(\mu') \cap \{(\alpha, \beta) | \beta < \beta^*\}).$$
Then, \(\alpha_{\delta^*}(\mu^*) > \alpha_{\delta^*}(\mu')\) implies that
\[
\phi(T_{C}^{\delta^*}(\mu^*) \cap \{(\alpha, \beta) | \beta > \beta^*\}) > \phi(T_{C}^{\delta^*}(\mu') \cap \{(\alpha, \beta) | \beta > \beta^*\}).
\]

Therefore, \(\mu_C^* = \mu_C^*\) requires that
\[
\phi(T_{C}^{\delta^*}(\mu^*) \cap \{(\alpha, \beta) | \beta < \beta^*\}) < \phi(T_{C}^{\delta^*}(\mu') \cap \{(\alpha, \beta) | \beta < \beta^*\}).
\]

Hence,
\[
\psi_{CDD}^d(\mu^*) = \phi\{(\alpha, \beta) | \beta < \beta^*\} - \phi\{(\alpha, \beta) | \beta < \beta^*\}
\]
\[
> \phi\{(\alpha, \beta) | \beta < \beta^*\} - \phi(T_{C}^{\delta^*}(\mu') \cap \{(\alpha, \beta) | \beta < \beta^*\})
\]
\[
= \psi_{CDD}^d(\mu')
\]

Together with \(\psi_{CDD}^d(\mu^*) = \mu_{CDD}^d\) and \(\mu_{CDD}^* < \mu_{CDD}^d\), we conclude that \(\psi_{CDD}^d(\mu') < \psi_{CDD}^d(\mu^*) = \mu_{CDD}^d\) and \(\mu_{CDD}^* < \mu_{CDD}^d\). Together with \(\psi_{C}^d(\mu') = \mu_{C}^*\), it follows from this that \(\psi_{CDD}^d(\mu^*) > \mu_{CDD}^d\).

Consider a case in class 2. We immediately conclude that \(\alpha_{\delta^*}(\mu^*) > \beta^* > \alpha_{\delta^*}(\mu')\) in this case. Therefore, by the latter half of the above argument for class 1, we conclude that \(\psi_{CDD}^d(\mu') < \mu_{CDD}^d\) and \(\psi_{CDD}^d(\mu^*) > \mu_{CDD}^d\).

Consider a case in class 3. By the former half argument for class 1, we know that this case does not occur.

Finally, consider a case in class 4. Note that
\[
\psi_{CDD}^d(\mu^*) = \phi(T_{CDD}^{\delta^*}(\mu^*)) = \phi(T \cap \{(\alpha, \beta) | \beta > \beta^*\}) = \phi(T_{CDD}^{\delta^*}(\mu')) = T_{CDD}^{\delta^*}(\mu')
\]
in this case. Together with \(\psi_{CDD}^d(\mu^*) = \mu_{CDD}^d\) and \(\mu_{CDD}^* > \mu_{CDD}^d\), we conclude that \(T_{CDD}^{\delta^*}(\mu') = \psi_{CDD}^d(\mu^*) = \mu_{CDD}^d\) and \(\mu_{CDD}^* > \mu_{CDD}^d\). Together with \(\psi_{C}^d(\mu') = \mu_{C}^*\), it follows from this that \(\psi_{CDD}^d(\mu^*) < \mu_{CDD}^d\).

**Step 5** Consider a function \(m_{\delta^*} : (0, 1) \to \Delta\) that assigns to each \(\mu_C \in (0, 1)\) a distribution \(m_{\delta^*}(\mu_C) = (m_{\delta^*}(\mu_C), m_{\delta^*}(\mu_C), m_{\delta^*}(\mu_C)) \in \Delta\) by \(m_{\delta^*}(\mu_C) = \psi_{CDD}^d(\mu_C, 1 - \mu_C, 0)\).\(\mu_{CDD}^* = \psi_{CDD}^d(\mu_C, 1 - \mu_C, 0)\).

Note that \(m_{\delta^*(\mu_C)}(\mu_C) = \psi_{CDD}(\mu_C, 1 - \mu_C, 0)\) = \(0\) because the threshold \((13)\) is \(\alpha_{\delta^*(\mu_C, 1 - \mu_C, 0)}\) is true. Furthermore, \((\alpha, \beta) \in T_{\delta^*}^d(\mu_C, 1 - \mu_C, 0)\) if and only if \(\mu_C(b - \beta(b - c)) + (1 - \mu_C)d > a\). Hence, \(m_{\delta^*(\mu_C)} = \psi_{CDD}^d(\mu_C, 1 - \mu_C, 0)\) = \(0\) for \(\mu_C \in (0, \frac{a - d}{b - c})\) while \(m_{\delta^*(\mu_C)} = \psi_{CDD}^d(\mu_C, 1 - \mu_C, 0)\) = \(0\) and it is strictly increasing for \(\mu_C \in (\frac{a - d}{b - c}, 1)\). Therefore, together with the continuity of \(m_{\delta^*(\mu_C)} = \psi_{CDD}^d(\mu_C, 1 - \mu_C, 0)\), there exists an unique \(\bar{\mu}_C \in (\frac{a - d}{b - c}, 1)\) such that \(m_{\delta^*(\mu_C)}(\bar{\mu}_C) = 1\) and \(m_{\delta^*(\mu_C)}(\mu_C) < 1 - \mu_C\) if and only if \(\mu_C \in (0, \bar{\mu}_C)\).

Take any \(\mu_C \in (0, \bar{\mu}_C)\). Then, \(m_{\delta^*(\mu_C)} < 1 - \mu_C\) means that \(\psi_{CDD}^d(\mu_C, 1 - \mu_C, 0)\) = \(1 - \psi_{CDD}^d(\mu_C, 1 - \mu_C, 0)\) = \(\bar{\mu}_C\) = \(\mu_C\). Therefore, by a similar argument to Step 3, there exists a \(m' = (\mu_C', \mu_{CDD}', \mu_{DDDD}')\) with \(0 < \mu_C', \mu_{CDD}', \mu_{DDDD}' < 1\) such that \(\psi_{CDD}^d(\mu') = \mu_C\). Denote this \(\mu'\) as \(\mu'(\mu_C) = (\mu_C', \mu_{CDD}', \mu_{DDDD}')\) to show its dependence on \(\mu_C\) explicitly.

Furthermore, \(m_{\delta^*(\mu_C)} = \mu_C\), that is, that \(\psi_{CDD}^d(\mu_C') = \mu_C = \mu_C\) holds for \(\mu' = (\mu_C, 1 - \mu_C, 0)\) at \(\mu_C = \mu_C\). Therefore, we can extend the function \(\mu' : (0, \bar{\mu}_C) \to \{\mu' = (\mu_C', \mu_{CDD}', \mu_{DDDD}')\} < 1\) introduced above to a function \(\mu' : (0, \bar{\mu}_C) \to \Delta\) by assigning \(\mu'(\bar{\mu}_C) = (\mu_C, 1 - \mu_C, 0)\) to \(\mu_C\).

**Step 6** Now define a function \(\lambda_{\delta^*} : (0, \bar{\mu}_C) \to R\) that assigns to each \(\mu_C \in (0, \bar{\mu}_C)\) a value \(\lambda_{\delta^*}(\mu_C) = \psi_{CDD}^d(\mu'(|\mu_C|)) - \mu_{DDDD}(\mu_C)\).
Recall the distribution $\mu^t$ that, we showed in Step 4, exists corresponding to $\mu^*$. It is defined as a distribution that satisfies the condition that $\psi_C^d(\mu^t) = \mu_C = \mu_C^*$. We claim that $\mu_C^t < \mu_C$. Suppose that $\mu_C^t > \mu_C$. Then, $m_{DDD}^d(\mu_C^t) < 1 - \mu_C$ by Step 5. Therefore, $\psi_C^d((\mu_C^t, 1 - \mu_C, 0)) < \mu_C^t$. Then, by Step 1, $\psi_C^d((\mu_C, \hat{\mu}_{CDD}, \hat{\mu}_{DDD})) < \mu_C^t$ for all $\hat{\mu} = (\hat{\mu}_C, \hat{\mu}_{CDD}, \hat{\mu}_{DDD})$ with $\hat{\mu}_C = \mu_C^t$. This contradicts to $\psi_C^d(\mu^t) = \mu_C^t$. Then, suppose that $\mu_C^t = \mu_C$. This means that $\mu^t = (\mu_C^t, 1 - \mu_C, 0)$. However, as is shown in Step 4, $\mu_{DDD} > \mu_{DDD}^t > 0$. This is a contradiction. Hence, $\mu_C < \mu_C^t$ must hold. Therefore, $\mu_C^t$ is located as $\mu_C^t = \mu_C^t < \mu_C$! and the distribution $\mu^t$ identified in Step 4 satisfies $\mu^t = (\mu_C^t, \mu_C, \mu_C^*, \mu_{CDD}^t, \mu_{DDD}^t)(\mu_C^*)$ for the function $\mu^t : (0, \mu_C] \to \Delta$ that we introduced in Step 5.

Consider the value $\lambda^t(\mu_C^t)$. According to the result of Step 4, we have $\psi_C^d(\mu_C^t) > \mu_{CDD}^t(\mu_C^t)$ and $\psi_C^d(\mu_C^t) < \mu_{DDD}^t(\mu_C^t)$. Therefore, $\lambda^d(\mu_C^t) < 0$.

Consider the value $\lambda^d(\mu_C)$. By Step 5, we have $\psi_C^d(\mu_C^t) = \psi_C^d((\mu_C^t, 1 - \mu_C, 0)) = 1 - \mu_C > 0$ and $\mu_{DDD}^t(\mu_C^t) = 0$. Therefore, $\lambda^d(\mu_C^t) > 0$.

By the continuity of $\lambda^d$, there exists $\mu_C^* \in (\mu_C^t, \mu_C)$ such that $\lambda^d(\mu_C^*) = 0$. Define $\mu^t = (\mu_C^t, \mu_{CDD}^t, \mu_{DDD}^t)$ by $\mu_{CDD}^t = m_{CDD}(\mu_C^t)$, and $\mu_{DDD}^t = m_{DDD}(\mu_C^t)$. Then, by the construction of $\mu^t : (0, \mu_C] \to \Delta$, we have $\psi_C^d(\mu^t(\mu^t)) = \mu_C^t$. Therefore, the distribution $\mu^t$ thus defined is in fact $\mu^t = (\mu_C^*, \mu_{CDD}^t, \mu_{DDD}^t) = (\mu_C^t, \mu_{CDD}^t, \mu_{DDD}^t)$, and it satisfies $\psi_C^d(\mu^t) = \mu_C^*$. Furthermore, $\lambda^d(\mu_C^*) = 0$ means that $\psi_C^d(\mu_C^*) = \mu_C^*$, that is, $\psi_C^d(\mu^t) = \mu_C^*$. Therefore, the remaining $\mu_{CDD}^*$ also satisfies that $\psi_C^d(\mu^t) = \psi_C^d(\mu^t)$. Hence, $\psi_C^d(\mu^t)$ is a fixed point of $\mu^t$. Furthermore, $0 < \mu_C^* < 1$ because $\mu_C^t < \mu_C^* < \mu_C$. By the construction of $\mu_C^*$, $\psi_C^d((\mu_C^*, 1 - \mu_C^*, 0)) > \mu_C$ for any $\mu_C < \mu_C^*$. Therefore, $(\mu_C^*, \mu_{CDD}^*, \mu_{DDD}^*), \mu_{DDD}^*$ with $\mu_{DDD} = 0$ can not be a fixed point of $\psi_C^d$. $(\mu_C^*, \mu_{CDD}^*, \mu_{DDD}^*)$ with $\mu_{DDD}^t = 0$ can not be a fixed point of $\psi_C^d$ either because $\psi_C^d((\mu_C^t, 0, \mu_{DDD}^t)) = 0$. Therefore, we have $0 < \mu_C^t, \mu_{CDD}^t, \mu_{DDD}^t < 1$. Thus, the fixed point $\mu^t$ corresponds to a three-mode equilibrium. (Q.E.D.)

**Proof of Theorem 3**

Suppose that there exists a three-mode equilibrium in $PD((a, b, c, d, f)$. Then, $0 < \alpha^*(\mu) < \alpha$, and a type $(\alpha, \beta)$ with $\alpha = \alpha^*(\mu)$ is indifferent between $C$ and $CDD$. This means that $\mu_Ca + \mu_{CDD}a + \mu_{DDD}[c - \alpha^*(\mu)(b - c)] = \mu_Ca + \mu_{CDD}a + \mu_{DDD}[c - \alpha^*(\mu)(b - c)]$, 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 $\mu_{DDD} \geq \phi(\alpha^*(\mu) \leq \alpha, \beta < \beta^*)$ and $\mu_{CDD} \leq \phi(\alpha^*(\mu) \leq \alpha, \beta > \beta^*)$. This means that

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

Hence,

$$1 - \phi(\beta < \beta^*|\alpha^*(\mu) \leq \alpha))a + \phi(\beta < \beta^*|\alpha^*(\mu) \leq \alpha)[c - \alpha^*(\mu)(b - c)] \geq d.$$ (38)

This inequality (38) can not be satisfied if $d > \tilde{d}(a, b, c)$ because

$$\tilde{d}(a, b, c) = (1 - \gamma)a + \gamma c$$

$$\geq (1 - \phi(\beta < \beta^*|\alpha^*(\mu) \leq \alpha))a + \phi(\beta < \beta^*|\alpha^*(\mu) \leq \alpha)c$$

$$> (1 - \phi(\beta < \beta^*|\alpha^*(\mu) \leq \alpha))a + \phi(\beta < \beta^*|\alpha^*(\mu) \leq \alpha)[c - \alpha^*(\mu)(b - c)].$$
When the condition (21) holds, the inequality (38) can not be satisfied if \( d > \tilde{d}(a, b, c) \). Suppose to the contrary that the inequality (38) holds. Then,
\[
1 - \phi(\beta < \beta^*|\alpha^*(\mu) \leq \alpha) + \phi(\beta < \beta^*|\alpha^*(\mu) \leq \alpha)[c - \alpha^*(\mu)(b - c)] \geq d > (1 - \phi(\beta < \beta^*))a + \phi(\beta < \beta^*)c,
\]
which is rearranged into
\[
\phi(\beta < \beta^*|\alpha^*(\mu) \leq \alpha)) < \frac{1}{1 + \frac{b - a}{a - c}\alpha^*(\mu)}\phi(\beta < \beta^*).
\]
This contradicts the condition (21) at \( \mu^* = \alpha^*(\mu) \). (Q.E.D)

**Proof of Theorem 4**

Consider \( d < \tilde{d}_1(a, b, c) \). Suppose that there exists a three-mode equilibrium \( s : T \to \{C, CDD, DDD\} \) and it is of type 3. Then, the probability distribution \( \mu = (\mu_C, \mu_{CDD}, \mu_{DDD}) = (\phi(T_C)(s), \phi(T_{CDD})(s), \phi(T_{DDD})(s)) \) over \( C, CDD, DDD \) induced by \( s \) is such that \( \mu_{CDD} = \phi(\beta > \beta^*) \) and \( \mu_{DDD} < \phi(\beta < \beta^*) \). Therefore, it holds for a player of type \((\alpha, \beta) = (\beta^*, \beta^*)\) that
\[
\begin{align*}
U(\beta^*, \beta^*) &- U(\beta^*) \leq \mu_{CDD} [\phi(\beta < \beta^*) - \phi(\beta < \beta^*)](d - c) + (b - a) \\
&= \frac{b - a}{b - c} \beta^* \leq \alpha^*(\mu) = \frac{\mu_{CDD} a - d}{\mu_{DDD} b - a} - \frac{d - c}{b - a}.
\end{align*}
\]
Hence, \( \frac{b - a}{b - c} \leq \frac{a - d}{a - c} \). However, the inequality (18) implies
\[
\frac{b - c}{a - c} > \frac{a - d}{a - c} \phi(\beta < \beta^*|\alpha \leq \alpha) \geq \frac{a - d}{a - c} \left(1 + \frac{\mu_{CDD}}{\mu_{DDD}}\right)
\]
where the second inequality holds for the reason that
\[
\frac{\mu_{CDD}}{\mu_{DDD}} \leq \frac{\phi(\beta > \beta^*|\alpha \leq \alpha)}{\phi(\beta < \beta^*|\alpha \leq \alpha)}
\]
in a three-mode equilibrium of both type 1 and type 2. This is a contradiction. Therefore, there should be no three-mode equilibrium of type 1 or type 2. (Q.E.D)

**Proof of Theorem 5**

Suppose that \( \phi(\beta < \beta^*) < \frac{a - d}{a - c} \). Then, \( c < \tilde{d}_2(a, b, c) \). It is also the case that \( \tilde{d}_2(a, b, c) < \tilde{d}(a, b, c) \) since \( a < b \). The interval \((c, \tilde{d}_2(a, b, c))\) is nonempty and
including in \((c, \bar{d}(a, b, c))\). Take \(d \in (c, \bar{d}_1(a, b, c))\). Then, there exists a three-mode equilibrium. By Theorem 4, it should not be of type 3. Hence, there must exist a three-mode equilibrium of either type 1 or type 2.

Next, suppose that \(\frac{a-c}{b-c} \phi(b < \beta^*) < \phi(b < \beta^* | \alpha^* \leq \alpha)\) for any \(\alpha^* \in [\frac{b-c}{a-c}, \bar{a}]\). Then,

\[
\phi(b < \beta^*) < \frac{b-c}{a-c} \min_{\alpha^* \in [\frac{b-c}{a-c}, \bar{a}]} \phi(b < \beta^* | \alpha^* \leq \alpha),
\]

because \(\phi(b < \beta^* | \alpha^* \leq \alpha)\) is continuous on \([\frac{b-c}{a-c}, \bar{a}]\). Then, \(\bar{d}_3(a, b, c) < \bar{d}(a, b, c)\). The interval \((\max[c, \bar{d}_3(a, b, c)], \bar{d}(a, b, c))\) is nonempty and included in \((c, \bar{d}(a, b, c))\). Take \(d \in (\bar{d}_3(a, b, c), \bar{d}(a, b, c))\). Then, there exists a three-mode equilibrium. By Theorem 4, it should not be of neither type 1 nor type 2. Hence, there must exist a three-mode equilibrium of type 3. (Q.E.D.)

**Proof of Theorem 6**

Fix \(\epsilon > 0\) arbitrarily. Then, under the condition (21),

\[
\phi(b < \beta^* | \alpha^* \leq \alpha)\left[1 + \frac{b-c}{a-c} \alpha^*\right] - \phi(b < \beta^*) > 0
\]

for any \(\alpha^* \in [\epsilon, \bar{a}]\). Hence,

\[
\left(\min_{\alpha^* \in [\epsilon, \bar{a}]} \phi(b < \beta^* | \alpha^* \leq \alpha)\left[1 + \frac{b-c}{a-c} \alpha^*\right]\right) - \phi(b < \beta^*) > 0.
\]

Define

\[
d_{\alpha L} \equiv \bar{d}(a, b, c) - (a-c)\left\{\left(\min_{\alpha^* \in [\epsilon, \bar{a}]} \phi(b < \beta^* | \alpha^* \leq \alpha)\left[1 + \frac{b-c}{a-c} \alpha^*\right]\right) - \phi(b < \beta^*)\right\}.
\]

Then, \(d_{\alpha L} < \bar{d}(a, b, c)\). Consider any \(d\) with \(d_{\alpha L} < d < \bar{d}(a, b, c)\) and let \(\bar{a}_L\) be the supremum of the set of envy parameters of leaders in any three-mode equilibrium under \(d\). Then, we show that \(\bar{a}_L < \epsilon\). Suppose to the contrary that \(\bar{a}_L \geq \epsilon\). Then, in any three-mode equilibrium under \(d\), \(\alpha^*(\mu) \geq \bar{a}_L \geq \epsilon\) for the corresponding belief \(\mu\) over \(C, CDD\), and \(DDD\) because \(\alpha^*(\mu) \geq \alpha\) for any type \((\alpha, \beta)\) which follows \(C\). The type \((\alpha, \beta)\) with \(\alpha = \alpha^*(\mu)\) must be indifferent between \(C\) and \(CDD\) so that

\[
(1 - \phi(b < \beta^* | \alpha^*(\mu) \leq \alpha))a + \phi(b < \beta^* | \alpha^*(\mu) \leq \alpha)[c - \alpha^*(\mu)(b-c)] = d.
\]

This can not hold because \(\alpha^*(\mu) \geq \epsilon\) means that

\[
(1 - \phi(b < \beta^* | \alpha^*(\mu) \leq \alpha))a + \phi(b < \beta^* | \alpha^*(\mu) \leq \alpha)[c - \alpha^*(\mu)(b-c)] \\
\leq a - (a-c) \min_{\alpha^* \in [\epsilon, \bar{a}]} \phi(b < \beta^* | \alpha^* \leq \alpha)\left[1 + \frac{b-c}{a-c} \alpha^*\right] \\
= d_{\alpha L} < d.
\]

Hence, it must be the case that \(\bar{a}_L < \epsilon\). (Q.E.D.)

**Proof of Theorem 7**

The sequential rationality of the follower’s strategy is equivalent to the sequential rationality of the behaviors \(aC((\alpha, \beta))\) and \(aD((\alpha, \beta))\) in a strategy for \(PD\) games. Then, the prescription in Theorem 7 follows from Lemma 1.
The follower’s strategy generates a distribution of follower’s behaviors. If the leader chooses $C$, then the follower chooses $C$ with a probability $\phi(\beta > \beta^*)$ and $D$ with a probability $1 - \phi(\beta > \beta^*)$. Therefore, the expected utility for the leader from choosing $C$ is $\phi(\beta > \beta^*)a + (1 - \phi(\beta > \beta^*))[c - \alpha(b - c)]$. 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 $\phi(\beta > \beta^*)a + (1 - \phi(\beta > \beta^*))[c - \alpha(b - c)] > d$. This condition is rearranged into $\alpha < \alpha^{**}$. (Q.E.D)

**Proof of Theorem 8**

By Theorem 7, leadership emerges in $SPD((a, b, c, d), (f, g))$ with a positive probability if and only if $\phi(\alpha < \alpha^{**}) > 0$. This condition is equivalent to $\alpha^{**} > 0$. This holds if and only if $c < d < \bar{d}(a, b, c)$. (Q.E.D)

**Proof of Theorem 9**

Suppose that player $i$ has a better record of leadership in $PD$ than player $j$. Then, there exist $PD((a, b, c, d), f)$ and $PD((a', b', c', d'), f)$ such that $\bar{\alpha}_L < \bar{\alpha}_F$, player $i$ chose $C$ at timing 1 in $PD((a, b, c, d), f)$, and player $j$ chose $D$ at timing 1 in $PD((a', b', c', d'), f)$. Player $i$'s type $(\bar{\alpha}_L, \bar{\beta}_i)$ must satisfy $\alpha_i \leq \bar{\alpha}_L$ because he chose $C$ at timing 1 in $PD((a, b, c, d), f)$. Player $j$'s type $(\bar{\alpha}_F, \bar{\beta}_j)$ must satisfy $\bar{\alpha}_F \leq \bar{\alpha}_j$ because he chose $D$ at timing 1 in $PD((a', b', c', d'), f)$. Hence, $\alpha_i < \bar{\alpha}_L < \bar{\alpha}_F < \bar{\alpha}_j$. Then, for any $\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}$, if player $j$ chooses $C$ as the leader in $SPD((\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}), (f, g_j))$, then $\alpha_j < \alpha^{**}((\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}), (f, g_j))$ where $\alpha^{**}((\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}), (f, g_j))$ is the threshold $\alpha^{**}$ for $SPD((\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}), (f, g_j))$ described in Theorem 7. Then, $\alpha_i < \alpha_j < \alpha^{**}((\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}), (f, g_j))$. Therefore, Theorem 7 guarantees that player $i$ chooses $C$ as the leader in $SPD((\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}), (f, g_j))$. Furthermore, for $\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}$ such that $\alpha_i < \alpha^{**}((\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}), (f, g_j)) < \alpha_j$, player $i$ chooses $C$ as the leader and player $j$ chooses $D$ as the leader in $SPD((\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}), (f, g_j))$. Hence, player $i$ is a better leader in $SPD$ than player $j$. (Q.E.D)

**Proof of Theorem 10**

Suppose a player described in Theorem 10. Then, his type $(\alpha, \beta)$ must be $\alpha = 0$. Suppose to the contrary that $\alpha > 0$. Then, by Theorem 6, there exists $N$ such that $0 < \bar{\alpha}_L^N < \alpha$ where $\bar{\alpha}_L^N$ is the supremum of envy parameters of leaders in any three-mode equilibrium in $PD((a, b, c, d^N), f)$. Then, this player should not choose $C$ at timing 1 in $PD((a, b, c, d^N), f)$. This is a contradiction.

The player with $\alpha = 0$ chooses $C$ for any $SPD((a', b', c', d'), (f, g))$ with $c < d' < \bar{d}(a', b', c')$ because $d' < \bar{d}(a', b', c')$ means $\alpha = 0 < \alpha^{**}$ and Theorem 7 applies. (Q.E.D)

**References**


