Inductive Game Theory: Discrimination and Prejudices

MAMORU KANEKO
University of Tsukuba

AKIHIKO MATSUI
University of Tokyo and University of Tsukuba

Abstract

This paper proposes a new theory, which we call *inductive game theory*. In this theory, the individual player does not have a priori knowledge of the structure of the game that he plays repeatedly. Instead, he accumulates experiences induced by occasional random trials in the repeated play. A stationary state is required to be stable against intentional deviations based on the player's experiences, and then it turns out to be a Nash equilibrium. The main part of the paper is the consideration of possible individual views of the society based on individual experiences. This view is defined to be a model of the society which the player builds from his experiences. Coherency with these experiences and a condition called rationalization are required for a model. As concrete objects of the theory, this paper analyzes the phenomena of discrimination and prejudice. The development of the new theory is undertaken by contrasting its observational and behavioral aspects with mental and judgmental aspects. The relationship between discrimination and prejudice will emerge in this dichotomous consideration.

Mamoru Kaneko, Institute of Policy and Planning Sciences, University of Tsukuba, Ibaraki, 305, Japan (kaneko@shako.sk.tsukuba.ac.jp). Akihiko Matsui, Faculty of Economics, University of Tokyo, Hongo, Bunkyo-ku, Tokyo 113, Japan, and the Institute of Policy and Planning Sciences, University of Tsukuba, Ibaraki, 305, Japan (amatsui@e.u-tokyo.ac.jp, amatsui@shako.sk.tsukuba.ac.jp).

The authors thank J. Conley, the associate editor, and referees of this journal for useful comments on an earlier version of this paper.

Received 10 February 1997; Accepted 11 August 1998.

Journal of Public Economic Theory, 1 (1), 1999, pp. 101–137.

1. Introduction

1.1 Motivation and Backgrounds

Societies consisting of several racial, religious, and cultural groups are called *multiethnic*. In these societies, the phenomena of discrimination and prejudices are typically observed. These phenomena raise not only practical societal issues but also offer some theoretical problems for economics and game theory. Among these problems is the treatment of interactions between behavioral and mental attitudes. The purpose of this paper is to present a theoretical framework that enables us to analyze the relationships between these two components of multiethnic societies. In this subsection, we look at the nature of discrimination and prejudices, and argue that it is not captured in the standard framework of economics and game theory.

Discrimination is an overt attitude toward some ethnic groups. It is a certain mode of behavior that includes, as an example, denial of a minority's access to political power and economic opportunities. On the other hand, prejudices, which can be defined as associations of a certain group of people or objects with some negative traits, are covert in nature; they are beliefs or preferences as opposed to behavior. Unlike the beliefs and preferences typically assumed in economics, prejudices have some notable characteristics. They are categorical and generalized thoughts. They are usually caused by the lack of sufficient knowledge about the targeted people or objects. If we carefully listen to a negative opinion against a certain group of people, we often find that the person who expresses such an opinion has not met so many people of that group as to make a logical claim. Generalization of limited knowledge to a categorical judgment is an important characteristic of prejudices. Another related characteristic of prejudices is that they contain fallacious elements to a significant degree.

In order to incorporate these characteristics in the scope of our research, we develop an analytical framework called *inductive game theory*. As its name suggests, induction is the key concept. In this theory, each player has little a priori knowledge of the structure of the society, but the lack of such knowledge is partially compensated for by the player's experiences in a recurrent situation. Here he uses induction to derive an image of or a view of the society from these experiences.[1] In this framework, we treat prejudices as a "fallacious" image against some ethnic groups. By focusing on the problem of discrimination and prejudices, we try to develop a theory of interactions between the thoughts in the mind of the player and his behavior in a social context. In the development, we do not discuss

[1] We use the term "induction" to mean the act of deriving a general law or a causal relationship from limited experiences (observations) rather than the meaning used in the literature of game theory such as backward induction.

the information processing of the mind of the player; instead, we focus on logically possible images formed by induction in his mind.

An attempt to analyze fallacious beliefs and preferences in the existing frameworks of game theory poses some difficulty. To see this, we look at some existing theories, starting with the classical game theory of rational players and followed by learning and evolutionary theories.[2]

In classical game theory since Nash (1951), it is often implicitly, and sometimes explicitly, assumed that players are rational in the sense of having high abilities of logical reasoning and knowledge of the structure of the game. Based on such abilities and a priori knowledge, the individual player makes a decision ex ante. We call this theory *deductive game theory* because deduction is the main process of reasoning.[3] In this light, deductive game theory is appropriate for the study of societies where players are well informed—for example, small games played by experts. Since the reasoning process of the rational player is always "correct" and is based on a priori knowledge, there is no room for the emergence of prejudices in deductive game theory. Moreover, the problem of prejudices could be addressed in this approach only if players are assumed to have false beliefs a priori.

Other theories in the literature are non-Bayesian learning and evolution. In the models of non-Bayesian learning, some prespecified learning rules are used to adjust players' beliefs and/or behavior. Players may learn some parameters of the game and strategies of others as well as the payoffs from their own behavior. In evolutionary game theory, the survival of the fittest is the main force in the selection of strategies.[4] These approaches focus on economic problems where adaptive behavior and behavioral interactions are of prime importance. Inductive decision making is often their main focus, and little attention is paid to the formation of images or thoughts about the society in the minds of the players.

[2] Discrimination and prejudice have been studied extensively in sociology literature (cf. Marger 1991), but the weak point of such studies is a lack of analytical frameworks. However, the concept of prejudice appears not to fit to the analytical tools built into the literature of economics and game theory. Consequently, the study of discrimination and prejudice in economics and game theory is quite limited; see Arrow (1972) and Becker (1957).

[3] Refinement literature (cf. van Damme (1987)) is typically considered from the deductive point of view. Bayesian game theory since Harsanyi (1967–68) is along this line. Bayesian learning such as that described by Kalai and Lehrer (1993) also falls into this category. A more explicit treatment of the sophisticated logical and mathematical ability of each player is found in the game logic approach of Kaneko and Nagashima (1996, 1997).

Many papers have followed this view in their formal developments and applications of equilibrium theory. Sometimes, however, interpretations from other views like evolutionary or inductive game theory have been mixed with deductive interpretations (cf. Binmore 1987).

[4] See Selten (1991) for some discussions on basic postulates of evolutionary game theory.

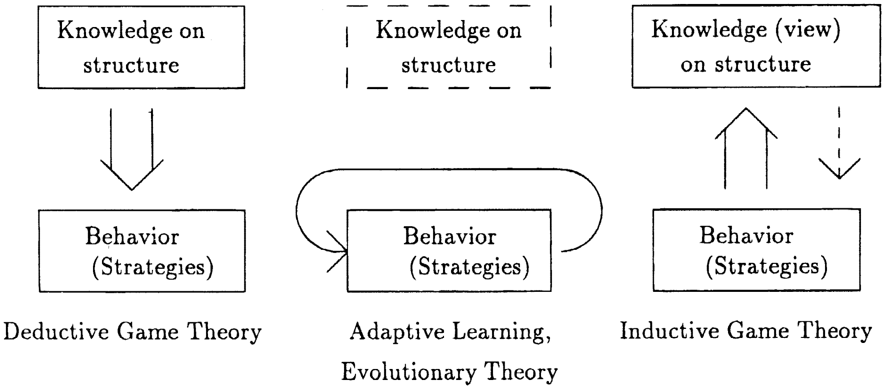

Figure 1: Three prototype theories.

Figure 1 summarizes the major differences between these three types of theories. Arrows indicate the causality flows between knowledge (view) of structure and behavior.

1.2 Development of Inductive Game Theory

Keeping the preceding discussions in mind, we describe our approach. We consider a specific game called the *festival game*, which is a variant of the game discussed in Kaneko and Kimura (1992). The festival game is a two-stage game in which the players are divided into several ethnic groups. These groups differ from each other only in their nominal ethnicities. In the first stage, all of the players simultaneously select a festival location. They observe which ethnic groups are present at their respective festival locations, and then they simultaneously decide to take either a friendly or unfriendly action. We consider the situation where the festival game is played repeatedly.

In the repeated situation of the festival game, we consider a stationary state subject to occasional random trials. The probabilities of such trials are assumed to be small so that each player does not take into account the events of simultaneous deviations of two or more players. In this environment, he accumulates his experiences from various unilateral deviations as well as from the stationary state. The experiences induced by the deviations of the individual player in question are called *active experiences*, and the ones induced by other players' unilateral deviations are called *passive experiences.*

In the absence of a priori knowledge, induction is taken as a general principle for the cognitive processes of the individual player. We consider two types of induction: (i) *inductive decision making*, and (ii) *inductive con-*

struction of an individual image of the society. The first is to choose a better strategy taught by experiences, and the second is to derive an interpretational view of the society based on experiences. The first type is categorized, more generally, into *inductive adjustments*, which could be found in classical equilibrium theory and learning theory in economics. The second type of induction is concerned not only with adjusting some parameters but also with building a new structure from experiences—and that is the main focus of this paper.

An individual player's image constructed inductively is formulated as a model of the society. Such a model is a partial description of the society including the individual's imaginary utility and observation functions. We give three coherency conditions on such a model, ones with the stationary, active, and passive experiences. These coherency conditions require that the utility and observation functions of the model generate the pieces of information corresponding to those experiences.

We consider another condition, called *rationalization*, for a model to satisfy. It follows from inductive decision making that the individual player makes a "rational" decision at every decision node that is reached. However, the action conceivable for him at such a decision node may lead to a social state never experienced. The rationalization condition requires that he rationalize his choice at such decision nodes. This goes beyond the coherency requirement, since it is a restriction over states never experienced.

There are many models that are coherent with experiences and satisfy the rationalization condition. One example is the *true-game model*, which is essentially the same as the game we consider from the objective point of view. Another obvious example is the *mere-enumeration model*, which enumerates one's experiences without giving any causal relationship. Although the second is important as a start from the inductive viewpoint, we do not consider it explicitly, since it needs a slight generalization of our definition of a model.

The active experiences impose few restrictions on models other than utility maximization. Indeed, each deviation by an individual player induces only a single pair of a utility value and an observation of ethnicities. Ignoring the observed ethnicities, a player can always construct a simplistic model coherent with the active experiences in which the utility function depends only on his own actions. Such a model is called a *naive hedonistic model*. However, this model can rarely explain the passive experiences in a satisfactory manner.

When passive experiences are taken into account, the inductive construction of an individual view may involve prejudices. Since passive experiences are induced by other players' deviations, they are associated with the effects of the presence of other ethnic groups. A *sophisticated hedonistic model* uses the observed ethnicities as explanatory variables of one's utility.

We show that this model explains the reality well in spite of its fallacy and exhibits preferential as well as perceptual prejudices.

In literature, there are many works treating inductive reasonings in social contexts. To help the reader differentiate our inductive game theory from existing works, we mention three recent related theories: the case-based decision theory of Gilboa and Schmeidler (1995), the theory of self-confirming equilibrium of Fudenberg and Levine (1993), and the theory of subjective equilibrium of Kalai and Lehrer (1995). Then we mention an ancient but more directly related work: *The allegory of the cave* in Book VII of Plato's *Republic* (1941).

The case-based decision theory emphasizes the information processing of a decision maker who evaluates alternative choices based on similarities between the present problem and past cases under the assumption that similar experiences lead to similar effects. Such evaluations get adjusted as more cases become available. This is an individual decision theory based on inductive adjustments of similarity evaluations.[5]

The self-confirming equilibrium of Fudenberg and Levine (1993) and the subjective equilibrium of Kalai and Lehrer (1995) describe a situation in which each player maximizes his expected payoff based on a belief consistent with what he observes in the course of play. Beliefs are expressed as subjective probabilities and are adjusted by new pieces of information obtained during the play of the game. These theories are also categorized into inductive adjustments, though they explicitly treat social aspects in contrast with the case-based decision theory.

The allegory of the cave in Book VII of Plato's *Republic* (1941) goes as follows. In the cave, prisoners have been from childhood chained by the leg and also by the neck, so that they cannot move and can see only the wall of the cave. On the wall, they see shadows of various things moving outside the cave, like the screen at a puppet-show. The only real things for them are the shadows. Each prisoner forms an individual view on the world from the shadows he has seen. Plato went on to discuss what might happen if one person were suddenly released to see the outside world, and how he would be treated after coming back to the other prisoners and telling them what he saw.

The framework of the present work, as well as its spirit, is similar to the story of the cave in that people with no a priori knowledge form a view of the society from experiences. The primary difference from the other works mentioned above is that the induction of our focus is one that builds a new structure out of limited experiences, while in the other works, parameters on the prespecified structures such as beliefs

[5] Matsui (1997) showed that the case-based decision theory of Gilboa and Schmeidler (1995) has virtually the same scope as that of expected utility theory; that is, the argument of the former can be captured in an appropriate setting of the latter, and vice versa.

about other players' strategies are adjusted so as to be consistent with experiences.[6]

The rest of the paper is organized as follows. Section 2 considers a recurrent situation in which a festival game is played repeatedly, and states the basic postulates for our analysis. Section 3 characterizes the set of Nash equilibria of the festival game. Section 4 provides the definitions of individual models, coherencies with experiences, and rationalization. Section 5 considers naive and sophisticated hedonistic models. Section 6 discusses implications of our analysis for the game theory, economics, and sociological literatures, considers interactions between individual views and behavior, and finally gives a remark on the traditional view of a game with common knowledge of the structure of the game.

2. Inductive Decision Making and Nash Equilibria

We consider a recurrent situation where a game called the *festival game* Γ has been and will be played many times:

$$\text{unilateral trials}$$

$$\text{past } \underline{\ldots\, \Gamma \,\ldots\, \Gamma \,\ldots\, \Gamma \,\ldots} \text{ future}$$

In section 2.1, we provide a description of the festival game and some concepts to be used in the subsequent analysis. In section 2.2, we describe the basic postulates for our analysis of the entire recurrent situation. Then we give the definitions of active and passive experiences for each individual player, and characterize Nash equilibrium from our point of view.

2.1 Festival Game Γ

The *festival game* Γ is a two-stage game.[7] The player set $N = \{1, \ldots, n\}$ is partitioned into ethnic groups $N_1, \ldots, N_{e_0}$ with $\#N_e \geq 2$ for $e = 1, \ldots, e_0$, where e_0 is the number of ethnic groups, and $\#N_e$ is the number of players in ethnic group N_e. Let $e(i)$ denote the ethnicity of player i, that is, $i \in$

[6] In this paper, we consider only possible logical structures of inductive construction of individual views on the society from experiences. We do not consider a dynamics of formations of views. This problem involves a lot of difficulties: First, there are (infinitely) many possible views coherent with experiences. Second, people have very different propensities to look for new views: some people are always looking for better views, and others have a tendency to get stuck in one view until they meet significantly contradictory evidence to that view. The development of a theory of a dynamic process of forming and revising a view seems to be a problem of the distant future.

[7] The festival game in normal form was discussed in Kaneko (1987) and Kaneko and Kimura (1992) in the context of stable conventions. The festival game of this paper is a modification of the festival game in extensive form that was discussed in Kaneko and Raychoudhuri (1993).

$N_{e(i)}$. All the players are identical except for their ethnicities. There are ℓ locations for festivals. We may call the festival at location k ($k = 1,\ldots,\ell$) *festival k.*

The game Γ has two stages: the stage of *choosing festival locations* and the stage of *acting in festivals* (see Figure 2). In the first stage of the game, each player simultaneously chooses a festival location. Player i's choice in this stage is denoted by $f_i \in \{1,\ldots,\ell\}$. We write $f = (f_1,\ldots,f_n)$.

After the choice of a festival, each player observes the ethnicity configuration in the festival he chose—which ethnic groups are present in his festival. Formally, given $f = (f_1,\ldots,f_n)$, player i observes the *ethnicity configuration* of festival f_i, which is defined to be the set:

$$E_i(f) = \{e(j) : f_j = f_i \quad \text{and} \quad j \neq i\}.$$

Each player can distinguish neither the identity of each participant nor the number of the participants of each ethnic group in the festival he chose. This is an assumption to simplify the subsequent analysis. Note that player i's own ethnicity is not counted if no other players of the same ethnicity are in the festival.

In the second stage, after observing the ethnicity configuration $E_i(f)$ of festival f_i, player i chooses his attitude, either *friendly* or *unfriendly,*

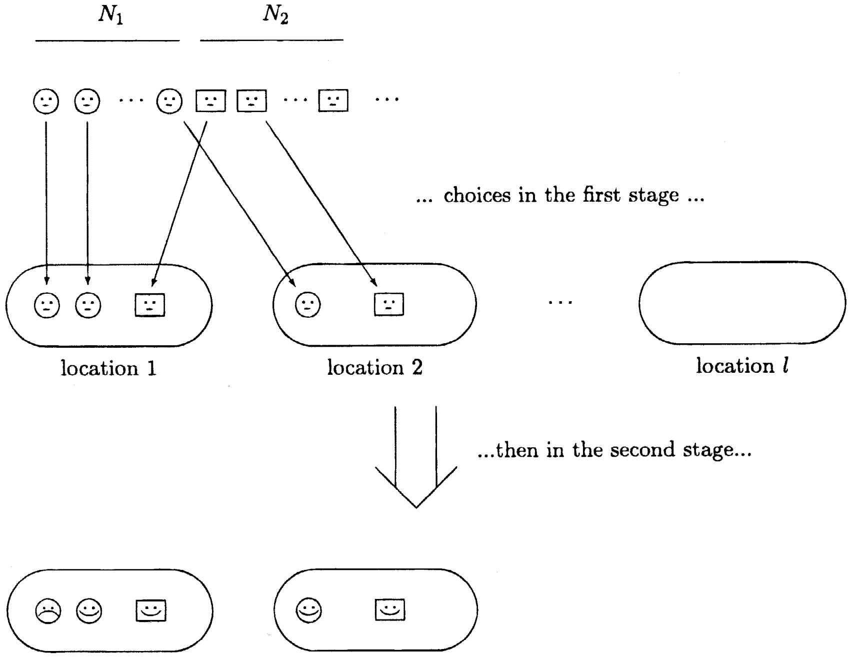
N_1
N_2
... choices in the first stage ...
location 1
location 2
location l
...then in the second stage...

Figure 2: Festival game.

denoted by 1 and 0, respectively. Following standard game theory, a choice in the second stage is expressed by a function $r_i : \{1, \ldots, \ell\} \times 2^{\{1,\ldots,e_0\}} \to \{0,1\}$. A value $r_i(k,\mathsf{E})$ is player i's attitude at festival k if he observes the ethnic configuration E.

A *strategy* of player i is a pair (f_i, r_i) of choices for the first and second stages. We write $r_i(f) = r_i(f_i, E_i(f))$ and $r(f) = (r_1(f), \ldots, r_n(f))$. Let Σ_i be the set of strategies of player i. For a strategy profile $\sigma = (f, r) \in \Sigma := \Sigma_1 \times \cdots \times \Sigma_n$, the *realization path* is given by a pair $(f, r(f))$.

Given a strategy profile $\sigma = (f, r)$, each player's payoff is determined by his attitude and the mood of the festival he chose. The *mood* of festival f_i for player i is given by the number of friendly people in festival f_i other than player i himself; that is,

$$\mu_i(f, r) = \sum_{f_j = f_i, j \neq i} r_j(f). \tag{1}$$

We define the *payoff* function of player i as

$$H_i(f, r) = r_i(f) \cdot (\mu_i(f, r) - m_0), \tag{2}$$

where m_0 is assumed to be a noninteger number greater than 2. Thus, the unfriendly action always induces the zero payoff, the payoff from the friendly action is increasing in the number of friendly people in the same location, and m_0 is the *threshold* beyond which the friendly action is preferred to the unfriendly action.

A strategy profile $\sigma^* = (f^*, r^*)$ is said to be a *Nash equilibrium* iff for all $i \in N$ and all $\sigma_i \in \Sigma_i$, $H_i(\sigma^*) \geq H_i(\sigma^*_{-i}, \sigma_i)$, where $(\sigma^*_{-i}, \sigma_i)$ denotes the strategy profile obtained from σ^* by replacing σ_i^* with σ_i. For the game Γ, we have the following equivalent definition of Nash equilibrium: For all $i \in N$, $H_i(\sigma^*) \geq H_i(\sigma^*_{-i}, (f_i, \delta_i))$ for all $(f_i, \delta_i) \in \{1, \ldots, \ell\} \times \{0,1\}$, where δ_i can be identified with a constant strategy taking value δ_i in the second stage.

We have formulated the festival game Γ and relevant game theoretic concepts in the standard manner. However, because we do not follow the standard ex ante view, we should be careful about the interpretation of each concept. For example, the payoff function $H_i(\cdot)$ is not known to player i himself as a function; instead, only each value is perceived by him. Also, we should be careful about the use of the standard definition of a strategy because, being a complete list of contingent actions, it appears to presuppose the knowledge of the extensive form of Γ. However, we can avoid this interpretation, and each player can "play" the game without being aware of the full-fledged concept of strategies.[8]

[8] A complete strategy can be replaced by a *partial* strategy defined on the domain of nodes experienced with the nonnegligible frequencies in the past. See Kaneko and Matsui (1999).

2.2 Stationary State, Individual Experiences, and Inductive Stability

In the recurrent situation of the game Γ, we consider a stationary state (strategy profile) $\sigma^* = (f^*, r^*)$, subject to unilateral deviations of individual players from the stationary state σ^*. Unilateral deviations give some knowledge about the society's responses, and under certain postulates, such knowledge enables each player to "maximize" his payoff against the stationary state and leads to a Nash equilibrium.

We first describe the basic postulates behind our mathematical formulation. Some postulates are often presupposed in standard game theoretical works. However, in order to emphasize what is different from these standard works and what is not, we make some of them explicit and write them in the form of postulates.

POSTULATE 1 (Knowledge structure):

(a) *After each play of game* Γ*, player i observes only his utility value,* $H_i(\sigma)$*, if the game is played according to* σ*, in addition to the information he obtained during the play of the game.*

(b) *Player i knows that there are festival locations* $1, \ldots, \ell$ *for his first choice, and that he has two options, friendly and unfriendly actions (0 and 1) in the festival he chose.*

Other than this knowledge, each player is entirely ignorant of the structure of the festival game including the player set N. In particular, though player i has the payoff function $H_i(\cdot)$, he does not know it as a function but receives a realized payoff value after each play of the game.

After each play of the festival game, player i has gained an *experience* that is a collection of the information he obtained through the course of the play. By Postulate 1, the collection is given by a quadruple:

$$[f_i, \delta_i, \mathsf{E}; h_i],$$

where f_i is the festival that player i went to, δ_i is his own attitude, E is the ethnicity configuration he observed, and h_i is the payoff received.

We assume that players behave following their behavior patterns σ^*, while making certain experiments and recording the information obtained from such experiments. The following postulate makes this assumption explicit.

POSTULATE 2 (Behavior patterns and experimentations):

(a) *Given a stationary state* σ^**, each player i behaves according to his behavior pattern* σ_i^**, subject to stochastic trial deviations with small probabilities once in a while, but after each trial, he returns to his own behavior pattern* σ_i^**.*

(b) *Events of trials simultaneously made by two or more players have negligible frequencies, and they are ignored by the players.*

When the experiences for player i tell that it might be better to deviate, he would intentionally change his behavior pattern, which will be stated by another postulate.

Under Postulate 2, the individual experiences in the past are categorized into the following three classes:

(s): *stationary* experience: that induced by the stationary state $\sigma^* = (f^*, r^*)$

(a): *active* experiences: those induced by his own deviations

(p): *passive* experiences: those induced by deviations of some other players.

Each of the active and passive experiences is attained by the strategy profile induced by a unilateral deviation of a single player from σ^* by Postulate 2b.

The focus of this paper is not on the structure described in Postulate 2, but is on the step next to Postulate 2; that is, the focus of the paper is on possible individual views about the society constructed from experiences. Hence a mathematical treatment of these three types of experiences is crucial in this paper.[9]

The *stationary experience for player i under* $\sigma^* = (f^*, r^*)$ is expressed as

$$\textbf{(S)}: \; s(i|\sigma^*) = [f_i^*, r_i^*(f^*), E_i(f^*); H_i(\sigma^*)].$$

This is the collection of information that player i has regularly observed. Here f_i^* and $r_i^*(f^*)$ are his own actions, $E_i(f^*)$ is received after the first

[9] We do not fully specify the time structure and timing of trials. Although such a specification is not used in this paper, it would help in understanding the above argument to specify such possible time structures.

One possible formulation is to have a discrete time structure $\{\ldots -2, -1, 0, 1, 2, \ldots\}$. Each player's behavior is subject to a stochastic disturbance, and if such a disturbance occurs then his behavior (f_i, δ_i) is randomly chosen. One possible assumption is that each disturbance occurs, with a small probability ϵ in each period, independently across the players. Then the probability of two or more players making simultaneous trials is at most of the second order. It means that the frequency of such trials is negligible relative to that of unilateral trials when ϵ is very small. Then player i collects the experiences of the first order.

Another model can be regarded as the limit of the above discrete time structure as the time interval tends to zero. The time structure is expressed as the real continuum $(-\infty, +\infty)$. The festival game is played at each point in time. All players behave according to their stationary state σ^* at every point in $(-\infty, +\infty)$, except for occasional disturbances, which make players try other actions. For each player, these disturbances follow a Poisson process. The Poisson processes are assumed to be independent across the players. Therefore, there is at most one trial made at each point in time with probability one, and negligible frequency of simultaneous trials is a consequence of this process.

stage, and $H_i(\sigma^*)$ is the payoff value received after the second stage. Note that player i is not aware of the expressions in the brackets; that is, only the values described by these (meta-)expressions are observed by player i.

With a small frequency, player i himself deviated from his own behavior pattern σ_i^* and learned an active experience. An *active experience under* σ^* *induced by a trial* (f_i, δ_i) of player i is given as

$$\textbf{(A):}\ [f_i, \delta_i, E_i(f_{-i}^*, f_i); H_i(\sigma_{-i}^*, (f_i, \delta_i))], \text{ where } (f_i, \delta_i) \neq (f_i^*, r_i(f^*)).$$

The third element, $E_i(f_{-i}^*, f_i)$, is the ethnicity configuration player i observed in the festival f_i, and the fourth element, $H_i(\sigma_{-i}^*, (f_i, \delta_i))$, is the utility value enjoyed by him when he chose attitude δ_i in festival f_i. Let $\mathcal{A}(i|\sigma^*)$ denote the set of all active experiences of player i. Note that the stationary information $[f_i^*, r_i^*(f^*), E_i(f^*); H_i(\sigma^*)]$ is not contained in $\mathcal{A}(i|\sigma^*)$.

The passive experiences for player i are classified into: (PO) those induced by some *outsider j* who goes regularly to some festival f_j^* different from f_i^*, and (PI) those induced by some *insider j* who regularly comes to the festival f_i^*. These two types of experiences are formulated as follows:

$$\textbf{(PO):}\ [f_i^*, r_i^*(f_{-j}^*, f_j), E_i(f_{-j}^*, f_j); H_i(\sigma_{-j}^*, (f_j, \delta_j))], \text{ where } f_j^* \neq f_i^* = f_j;$$

$$\textbf{(PI):}\ [f_i^*, r_i^*(f_{-j}^*, f_j), E_i(f_{-j}^*, f_j); H_i(\sigma_{-j}^*, (f_j, \delta_j))],$$
$$\text{where } f_j^* = f_i^* \text{ and } (f_j, \delta_j) \neq (f_j^*, r_j^*(f^*)).$$

We denote the set of all passive experiences of player i by $\mathcal{P}(i|\sigma^*)$.

Notice that there is the following asymmetry between the active and passive experiences. Player i notices that the differences between the stationary and active experiences were caused by his own deviations. However, by Postulate 1, he does not identify any other player to cause the difference between the stationary and passive experiences; he only receives sometimes different information.

We denote the union $\mathcal{A}(i|\sigma^*) \cup \mathcal{P}(i|\sigma^*)$ by $\mathcal{E}(i|\sigma^*)$. A generic element of $\mathcal{E}(i|\sigma^*)$ is denoted by $[\phi_i; h_i]$, where ϕ_i consists of f_i, δ_i, and E.

Each player does not know his own utility function. However, he has experienced various utility values in $\mathcal{E}(i|\sigma^*)$. If he has found a higher utility value that can be induced by his own trial, then he would have an incentive to increase the frequency of this deviation from his present stationary behavior σ_i^*. Therefore, we make a postulate on his behavior in such a case, which defines the stability of σ^*.

POSTULATE 3 (Inductive decision making):

(a) *If no active experience in* $\mathcal{A}(i|\sigma^*)$ *gives a higher payoff to player i than his stationary payoff* $H_i(\sigma^*)$, *then he continues behaving according to* σ_i^* (*still subject to his occasional trials*).

(b) *If some active experience* $[\phi_i; h_i]$ *in* $\mathcal{A}(i|\sigma^*)$ *gives a higher payoff to player i than his stationary payoff* $H_i(\sigma^*)$*, then he would increase intentionally (maybe slightly, or maybe drastically) the frequency of the deviation inducing* $[\phi_i; h_i]$.

The following definition is based on this postulate. We say that a player i has an *incentive for an intentional deviation* in σ^* iff there is an active experience $[\phi_i; h_i] \in \mathcal{A}(i|\sigma^*)$ with $h_i > H_i(\sigma^*)$. A strategy profile σ^* is an *inductively stable state* iff no player has an incentive for an intentional deviation.

PROPOSITION 1: *A strategy profile* σ^* *is inductively stable if and only if it is a Nash equilibrium in* Γ.

Proof: Inductive stability is equivalent to that for any player i, $H_i(\sigma^*) \geq h_i$ for all $[\phi_i; h_i] \in \mathcal{A}(i|\sigma^*)$. This is equivalent to $H_i(\sigma^*) \geq H_i(\sigma^*_{-i}, \sigma_i)$ for all $\sigma_i \in \Sigma_i$. ■

Inductive stability is simply a translation of the mathematical definition of Nash equilibrium. However, it is important to evaluate the claim of Proposition 1 from the viewpoint of inductive decision making.

The *if* part means that if player i has no experience with a utility value higher than that in the stationary state, then he continues playing his strategy, which is Postulate 3a. Hence if no player has actively experienced a higher utility value, then σ^* is stable in the sense that all the players continue playing σ^*. This part involves a weak form of induction: When he has experienced the same stationary information except for some occasional changes, he expects that if he does not change his action, nothing will change, either.

The *only-if* part is more substantive. If a player has an active experience with a higher utility value than that in the stationary state, then he intentionally changes his behavior, slightly or drastically—Postulate 3b. In this sense, σ^* is no longer stationary. Here he does not know well the possible consequences of his intentional deviations. He is making an inductive decision based on a generalization of his active experiences, and he expects to receive a higher utility more frequently by making that deviation more often than before.

Finally, we should give one comment on subgame perfection—the equilibrium requirement for the second stage (sequential rationality according to the literature of refinements since the game has no proper subgame). The essential part of subgame perfection in the festival game Γ is the equilibrium requirement for the reactions of the players to a deviation by a single player—for example, the reactions of the players in a festival when an outsider comes to it. The payoff maximization for such a reaction requires an insider to have trial responses to the deviation of the outsider. Thus, subgame perfection needs experiences induced by trials of two or more players. However, Postulate 2b assumes that those events are negligible for each player. Thus, we cannot assume subgame perfection in our context. In Section 6.1 we will consider the problem of subgame perfection again.

3. Segregation Patterns and Discriminatory Behavior in Nash Equilibria

Before going to the main part of inductive game theory, we consider the structure of Nash equilibria in the festival game Γ. There are three types of equilibria, one of which exhibits segregation of some ethnic groups and discriminatory behavior to support such segregation. The other two are degenerated ones. The following theorem characterizes the set of Nash equilibria, a fortiori, inductively stable profiles, whose proof is given in the Appendix.

THEOREM 1: *A strategy profile $\sigma^* = (\sigma_1^*, \ldots, \sigma_n^*) = ((f_1^*, r_1^*), \ldots, (f_n^*, r_n^*))$ is a Nash equilibrium if and only if for any $i \in N_e$ and $e = 1, \ldots, e_0$,*

(a) *if $\mu_i(\sigma^*) \geq m_0$, then $f_j^* = f_i^*$ for any j with $e(j) = e$ and $r_j^*(f^*) = 1$ for any j with $f_j^* = f_i^*$*

(b) *if $\mu_i(\sigma^*) \geq m_0$, then $\mu_i(\sigma^*) \geq \mu_i(\sigma_{-i}^*, (f_i, 1))$ for any $f_i \in \{1, \ldots, \ell\}$*

(c) *if $\mu_i(\sigma^*) < m_0$, then $\mu_i(\sigma^*) = 0$, i.e., $r_j^*(f^*) = 0$ for any j with $f_j^* = f_i^*$*

(d) *if $\mu_i(\sigma^*) < m_0$, then $m_0 > \mu_i(\sigma_{-i}^*, (f_i, 1))$ for any $f_i \in \{1, \ldots, \ell\}$.*

If the number of friendly people at f_i^* reaches the threshold m_0, then: (1) every player of the same ethnicity as player i goes to the same festival, and every player in this festival takes a friendly action, and (2) if player i chooses another location f_i, the number of friendly people at f_i becomes smaller than or equal to the number at f_i^*. Note that (1) allows more than one ethnic groups to go to the same festival (such as Figures 3 and 4). On the other hand, if the number of friendly people at f_i^* is less than the threshold m_0, then: (3) no player at f_i^* takes a friendly action (such as festivals 2 and 3 in Figure 4), and (4) wherever player i goes, the number of friendly people would not exceed the threshold m_0.

The above theorem enables us to classify the set of equilibria into the following three classes:

Amalgamation equilibria: $f_i^* = f_j^*$ and $r_i^*(f^*) = r_j^*(f^*) = 1$ for all $i, j \in N$: all players choose the same festival and behave in the friendly manner. The players enjoy the highest mood. See Figure 3.

Segregation equilibria: $f_i^* \neq f_j^*$ and $\mu_i(\sigma^*) \geq m_0$ for some $i, j \in N$: some players of different ethnicities go to different festivals and at

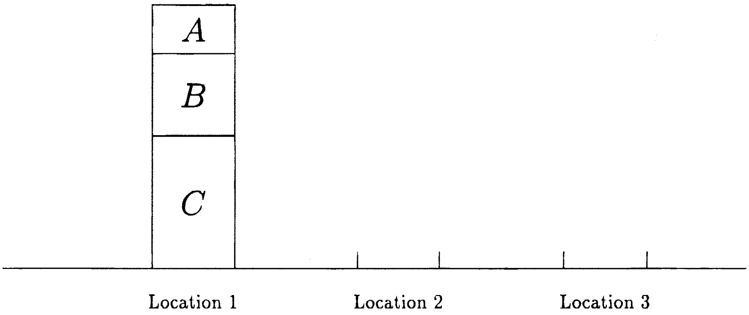

Figure 3: Amalgamation equilibrium.

least one festival is active. Segregation occurs in this equilibrium. See Figures 4 and 5.

No-festival equilibria: $\mu_i(\sigma^*) = 0$ for all $i \in N$: all players take unfriendly actions in their festivals. In this equilibrium, each player's choice of a location is arbitrary.

The first class consists of all equilibria in which everyone goes to the same festival and takes the friendly action. They attain the best possible payoffs.[10] With respect to realization, there are ℓ kinds of equilibria in this class, but they can be regarded as identical. The third class consists of degenerate equilibria in which nobody behaves in a friendly manner and the distribution of players is arbitrary. The second class is the one we will focus on in the subsequent sections.

In a typical segregation equilibrium, there are several active festivals, some of which are larger than others. A player who goes to a small festival

[10] These are efficient equilibria in the present formulation of the festival game. Nevertheless, the focus of this paper is on the segregation equilibria and the simple payoff functions are adopted for the consideration of segregation equilibria. Therefore it would not be very sensible to discuss the efficiency of equilibria in the present formulation.

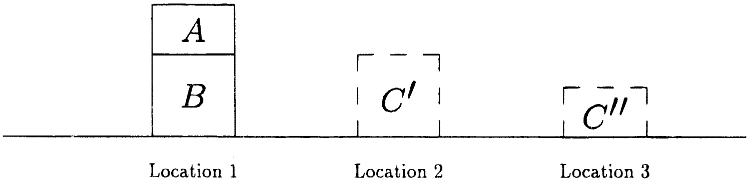

Figure 4: (Partially active) segregation equilibrium.

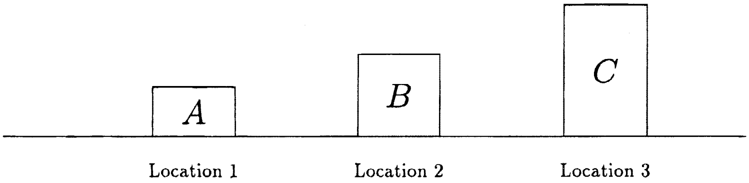

Figure 5: Fully active segregation equilibrium.

would have an incentive to visit a larger festival if the players at the larger festival took friendly action in response to his participation. However, if he goes to a larger festival and many players there take unfriendly action, the payoff to the newcomer as well as to others in the larger festival decreases considerably.

The following corollary states that in a segregation equilibrium, some players at festival f_i^* respond to j's participation in the unfriendly manner if j is from a smaller festival.

COROLLARY 1: *Let $\sigma^* = (f^*, r^*)$ be a segregation equilibrium. Suppose $f_i^* \neq f_j^*$ with $\mu_i(\sigma^*) > \mu_j(\sigma^*)$, and let $f_j = f_i^*$.*

(a) *If $\mu_j(\sigma^*) \geq m_0$, then $\mu_j(\sigma^*_{-j},(f_j,1)) \leq \mu_j(\sigma^*)$.*

(b) *If $\mu_j(\sigma^*) = 0$, then $\mu_j(\sigma^*_{-j},(f_j,1)) < m_0$.*

In case a, if player j of an active festival smaller than f_i^* comes to f_i^*, the induced mood by his presence is not better than the mood of festival f_j^* to which player j regularly goes. Thus, discrimination is necessarily incurred when a player of a smaller festival comes to a larger festival. The difference $\mu_j(\sigma^*) - \mu_j(\sigma^*_{-j},(f_j,1))$ is the number of players switching from friendly to unfriendly in response to the presence of j at festival f_i^*. Discrimination may or may not be incurred when a player in a larger festival visits a smaller festival. In case b, festival f_j^* is inactive, and then the induced mood must be worse than the threshold m_0.

To simplify the subsequent argument, we focus on the fully active equilibria $\sigma^* = (f^*, r^*)$ which satisfy:

FA: $r_i^*(f^*) = 1$ for any $i \in N$.

Condition FA allows some segregation equilibria, but eliminates some others such as the one in Figure 4, where some players take the unfriendly action on the equilibrium path.

When a Nash equilibrium σ^* satisfies subgame perfection, discriminators and nondiscriminators cannot coexist in one festival. In particular, all the players in a large festival have to discriminate against those who come from a smaller festival in such a case. Note that for each Nash equilibrium there is an equilibrium satisfying subgame perfection such that their realization paths are identical.

4. Inductive Construction of an Individual Image of the Society

In an inductively stable state $\sigma^* = (f^*, r^*)$, each player i has accumulated experiences $\mathcal{E}(i|\sigma^*) = \mathcal{A}(i|\sigma^*) \cup \mathcal{P}(i|\sigma^*)$ via occasional trials. He does not know the structure of the game Γ by Postulate 1, but may infer, from his experiences $\mathcal{E}(i|\sigma^*)$, what has been occurring in the society. Here we consider possible views about the society formed by player i from his experiences $\mathcal{E}(i|\sigma^*)$. Here we apply again an inductive principle to this process, which is considerably stronger than the induction used in Section 2: He generalizes his experiences into an explanatory causal relationship and builds a model of the society. In this section, we provide a general definition of such a model and two requirements for it: coherency with experiences and rationalization.[11]

4.1 Individual Models Built by a Player

An *individual model*, $\mathcal{M}_I$, of player i is given by a sextuple $(\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$, where

1. $\hat{N}$ is a finite set, the set of *imaginary players*
2. $\hat{Z}$ is a set, the set of *potential social states*
3. $\hat{o}_i$ is a function on $\hat{Z}$, the *observation function*
4. $\hat{u}_i$ is a real-valued function on $\hat{Z}$, the *utility function*
5. x^0 is an element of $\hat{Z}$, the *stationary social state*
6. X is a subset of $\hat{Z}$ containing x^0, the set of *relevant social states.*

These constituents are all imaginary in the sense that they are constructed in the mind of player i. The first four constituents, $\hat{N}, \hat{Z}, \hat{o}_i$, and $\hat{u}_i$, are intended to describe the basic structure of the society or game that player i imagines. That is, $\hat{N}, \hat{Z}, \hat{o}_i$, and $\hat{u}_i$ are an alternative description of the extensive form game Γ except that the other players' payoffs are omitted.[12] The last two constituents, x^0 and X, correspond to the realization path of the stationary state σ^* and the terminal nodes induced by indi-

[11] In this paper, we define individual models particularly for the festival game. A definition of models for general extensive form games is discussed in Kaneko and Matsui (1999).

[12] An individual model $\mathcal{M}_I$ describes only player i's observation and utility functions. It is natural to extend an individual model to include other players' observation and utility functions. Then the model becomes the *social model* imagined by player i which is given as $\mathcal{M}_S = (\hat{N}, \hat{Z}, (\hat{o}_j)_{j\in\hat{N}}, (\hat{u}_j)_{j\in\hat{N}}; x^0, \{X_j\}_{j\in\hat{N}})$. A social model raises quite different problems than an individual model. In this paper, we will consider only individual models; we will discuss social models in a separate paper.

vidual deviations. That is, x^0 and X describe the play of the game; in particular, X is the set of relevant states reachable from x^0 by unilateral deviations of player i himself and some other players.

The game Γ and an individual model $\mathcal{M}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ have a significant difference in their cognitive bases. The former is the objective description of the society, and the latter is its subjective description in the mind of player i. In particular, we emphasize the following difference: In the former, player i has the utility function $H_i(\cdot)$, which means that he receives each realized utility value, but not that he knows $H_i(\cdot)$ as a function. On the other hand, since he builds $\mathcal{M}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ in his mind, he perceives $\hat{u}_i$ as a function. This difference will be important in the rationalization requirement given in section 4.3.

In the following, we make the following assumptions on $\mathcal{M}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$:

Assumption M1: *$\hat{N}$ is a set expressed as the union of disjoint groups $\hat{N}_1, \ldots, \hat{N}_{e_0}$ with $i \in \hat{N}_{e(i)}$.*

Assumption M2: *$\hat{Z} = \{1, \ldots, \ell\}^{\hat{N}} \times \{0,1\}^{\hat{N}} \times Y$, where Y is some arbitrary set.*

Assumption M3: *$\hat{o}_i(g, \delta, y) = (g_i, \delta_i, \hat{E}_i(g))$ for all $(g, \delta, y) \in \hat{Z}$, where $\hat{E}_i(g) = \{e(j) : g_j = g_i, j \neq i \text{ and } j \in \hat{N}\}$ for any g.*

Since $x^0 \in \hat{Z}$, x^0 can be expressed as (g^0, δ^0, y^0). This notation will be used throughout the following.

Assumption M1 means that $\hat{N}$ is the set of imaginary players partitioned into the ethnic groups $\hat{N}_1, \ldots, \hat{N}_{e_0}$, and player i himself belongs to the group $\hat{N}_{e(i)}$. Assumption M2 expresses the idea that player i knows that every player in $\hat{N}$ has the same action space as that of player i. The additional Y is the set of hidden parameters player i imagines. When Y is singleton—that is, it is not used at all—M2 is essentially equivalent to $\hat{Z} = \{1, \ldots, \ell\}^{\hat{N}} \times \{0,1\}^{\hat{N}}$, which expression will be used instead of saying that Y is singleton.

Assumption M3 means that player i believes that he observes his own choice (g_i, δ_i) and the ethnicity configuration $\hat{E}_i(g)$. It is important to notice that M3 excludes the possibility that the observation function $\hat{o}_i$ depends on an additional variable y in Y (see Remark 1 below). On the other hand, we impose no further restriction on the utility function $\hat{u}_i$. Hence $\hat{u}_i$ may depend on the additional variable y.

The additional space Y is the domain of an exogenous *explanatory* variable. The introduction of this space gives some freedom to the possible models. Nevertheless, since a model with a large domain Y gives up a fine causal explanation, the less dependent a model is on y, the stronger is its explanatory power. In the models considered in Section 5, this additional variable y is used to explain some utility changes.

To illustrate the above definition of an individual model, we give one example of a model called the *true-game model* $\mathcal{TG}_I$. It is essentially a redescription of the festival game Γ itself in terms of the above language. It will be shown in Section 6.1 that this model hardly satisfies the second requirement, rationalization.

Let $\sigma^* = (f^*, r^*)$ be a strategy profile. The true-game model of player i is given as $\mathcal{TG}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$:

TG1: $\hat{N}_e = N_e$ for all $e = 1, \ldots, e_0$;

TG2: $\hat{Z} = \{1, \ldots, \ell\}^{\hat{N}} \times \{0,1\}^{\hat{N}}$; that is, it is the set of all terminal nodes in the game Γ;

TG3: $\hat{o}_i(x) = (f_i, \delta_i, E_i(f))$ for any $x = (f, \delta) \in \hat{Z}$;

TG4: $\hat{u}_i(x) = \delta_i \cdot (\mu_i(f, \delta) - m_0)$ for any $x = (f, \delta) \in \hat{Z}$;

TG5: $x^0 = (f^*, r^*(f^*))$;

TG6: $X = \{x^0\} \cup X_A \cup X_{PO} \cup X_{PI}$, where

$$X_A = \{((f^*_{-i}, f_i), (r^*_{-i}(f^*_{-i}, f_i), \delta_i)) : (f_i, \delta_i) \in \{1, \ldots, \ell\} \times \{0,1\}\},$$

$$X_{PO} = \bigcup_{j: f_j^* \neq f_i^*} \{((f^*_{-j}, f_j), (r^*_{-j}(f^*_{-j}, f_j), \delta_j)) : f_j = f_i^* \text{ and } \delta_j = 0,1\};$$

$$X_{PI} = \bigcup_{\substack{j: j \neq i \\ f_j^* = f_i^*}} \{((f^*_{-j}, f_j), (r^*_{-j}(f^*_{-j}, f_j), \delta_j)) : (f_j, \delta_j) \neq (f_j^*, r_j^*(f^*))\}.$$

The true game model, $\mathcal{TG}_I$, is described by focusing on the terminal nodes in the game Γ. The observation function $\hat{o}_i(x) = (f_i, \delta_i, E_i(f))$ gives the pieces of information obtained in the course of the play of Γ, and the utility function $\hat{u}_i(x) = \hat{u}_i(f, \delta)$ gives the value equal to the payoff $\delta_i \cdot (\mu_i(f, \delta) - m_0)$ assigned to the corresponding terminal node in Γ. The stationary state x^0 corresponds to the realization path $(f^*, r^*(f^*))$ of $\sigma^* = (f^*, r^*)$. The set X of relevant social states contains the three types of states: a state in X_A is induced by a deviation of player i himself, a state in X_{PO} is induced by a deviation of some outsider, and a state in X_{PI} is induced by a deviation of an insider. Thus, the first four constituents $(\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i)$ are an alternative description of the extensive form game Γ except for the absence of the other players' observation and utility functions. The last pair (x^0, X) corresponds to the stationary state σ^* and the terminal nodes induced by individual deviations. Consequently, the model $\mathcal{TG}_I$ satisfies Assumptions M1–M3. Here, the additional space Y is not used.

Remark 1: From the viewpoint of induction, there is another extreme and important example, which is an enumeration of the experiences without adding any additional structure. To have this example in our theory, we need to generalize Assumption M3 so that the observation function $\hat{o}_i$ depends on the additional variable y. Then all experiences, except the player's own actions, can be recorded in the space Y. We call this the *mere-enumeration model* $\mathcal{ME}_I$. We will refer to this example in the discussions in Section 6.2. (For a generalization of M3 and the definition of this model, see the original version of this paper: Kaneko and Matsui 1997.)

4.2 Coherencies of Models with Experiences

In this subsection, we formulate the requirements for a model to be coherent with the stationary, active, and passive experiences. For an understanding of these coherency requirements and of their uses, we give two theorems immediately after the formulation of these requirements, and we illustrate them by the true-game model.

Let $\sigma^* = (f^*, r^*)$ be a stationary state, and $\mathcal{M}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ be an individual model of player i satisfying M1–M3. Stationary state $\sigma^* = (f^*, r^*)$ defines the stationary experience $s(i|\sigma^*)$, active experiences $\mathcal{A}(i|\sigma^*)$, and passive experiences $\mathcal{P}(i|\sigma^*)$. The coherency requirement states that model $\mathcal{M}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ generates these experiences.

Formally, the coherency condition for a model $\mathcal{M}_I$ with the stationary experiences is given as follows:

Condition CS: $[\hat{o}_i(x^0); \hat{u}_i(x^0)] = s(i|\sigma^*)$.

Recall that $\hat{o}_i(x^0) = (g_i^0, \delta_i^0, \hat{E}(g^0))$ and $s(i|\sigma^*) = [f_i^*, r_i^*(f^*), E_i(f^*); H_i(\sigma^*)]$. This requires that the stationary state x^0 in $\mathcal{M}_I$ gives the stationary information $s(i|\sigma^*)$ that player i has obtained in σ^*.

Second, the coherency condition for $\mathcal{M}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ with $\mathcal{A}(i|\sigma^*)$ is formulated as follows:

Condition CA: For any $[\phi_i; h_i]$, $[\phi_i; h_i] \in \mathcal{A}(i|\sigma^*)$ if and only if there is a state $x = (g, \delta, y) \in X$ such that $(g_i, \delta_i) \neq (g_i^0, \delta_i^0)$, $g_{-i} = g_{-i}^0$, and $[\hat{o}_i(x); \hat{u}_i(x)] = [\phi_i; h_i]$.

This states that player i interprets each active experience $[\phi_i; h_i]$ by associating it with a state $x = (g, \delta, y)$ induced by his own deviation (g_i, δ_i). That is, he explains each active experience by his unilateral deviation.

Finally, the coherency condition for $\mathcal{M}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ with $\mathcal{P}(i|\sigma^*)$ is formulated as follows:

Condition CP: For any $[\phi_i; h_i]$, $[\phi_i; h_i] \in \mathcal{P}(i|\sigma^*)$ if and only if there is a state $x = (g, \delta, y) \in X$ such that $(g_j, \delta_j, y) \neq (g_j^0, \delta_j^0, y^0)$, $g_{-j} = g_{-j}^0$ for some $j \in \hat{N} - \{i\}$, and $[\hat{o}_i(x); \hat{u}_i(x)] = [\phi_i; h_i]$.

This states that player i interprets a passive experience $[\phi_i; h_i]$ by associating it with a state $x = (g, \delta, y)$ induced by a deviation (g_j, δ_j) of some other player j or by a change in y. That is, player i regards a passive experience as induced by some other (imaginary) player or caused by an exogenous variable y. We emphasize that because of the asymmetry between CA and CP player i cannot detect who (or what y) induced the passive experiences, but he is certain that he has induced his active experiences.

Since active and passive experiences are obtained through individual deviations from the stationary state, conditions CA and CP would be meaningful only when they are coupled with CS. Thus, we give the following definitions.

DEFINITION 1. *An individual model $\mathcal{M}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ is coherent with the active experiences $\mathcal{A}(i|\sigma^*)$ of player i iff CS and CA hold. We say that $\mathcal{M}_I$ is coherent with the experiences $\mathcal{E}(i|\sigma^*) = \mathcal{A}(i|\sigma^*) \cup \mathcal{P}(i|\sigma^*)$ iff CS, CA, and CP hold.*[13]

It is possible to consider the combination of CS and CP, but in this paper we will focus on coherency with the active experiences $\mathcal{A}(i|\sigma^*)$ and coherency with the entire experiences $\mathcal{E}(i|\sigma^*)$.

The true-game model $\mathcal{TG}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ is coherent with the experiences $\mathcal{E}(i|\sigma^*)$ with respect to $\sigma^* = (f^*, r^*)$ in the trivial sense that $\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i$ are constructed from our knowledge of the festival game Γ and x^0, X are constructed from $\sigma^* = (f^*, r^*)$. However, since player i has no knowledge of the game Γ except as described in Postulate 1, $\mathcal{TG}_I$ is no more than one of many candidates. In fact, it will be shown in Section 6.1 that the true-game model $\mathcal{TG}_I$ hardly satisfies the other criterion—rationalizability—under our postulates.

We give two theorems to illustrate the effects of the coherency conditions. The first theorem states that when an inductively stable stationary state σ^* is given, the player knows that he obtains the maximum utility at the stationary state over the states he can induce by his own deviations.

Recall that the proofs of the results are given in the Appendix.

THEOREM 2 (Utility maximization for player i): *Let $\sigma^* = (f^*, r^*)$ be an inductively stable stationary state. If an individual model $\mathcal{M}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ is coherent with the active experiences $\mathcal{A}(i|\sigma^*)$, then*

$$\hat{u}_i(x^0) \geq \hat{u}_i(x) \tag{3}$$

for all $x = (g, \delta, y) \in X$ with $g_{-i} = g^0_{-i}$ and $(g_i, \delta_i) \neq (g^0_i, \delta^0_i)$.

Thus, an individual model coherent with the active experiences should satisfy utility maximization for the player. This theorem has the same spirit as Proposition 1—that is, it is a manifestation of the postulate of inductive decision making in model $\mathcal{M}_I$.

[13] The reader may find some similarity between our consideration of models and model theory (semantics) in mathematical logic (cf. Mendelson 1987); nevertheless, model theory is a branch of deductive logic, while our theory is based on induction.

Conversely, if every player has developed a model that is coherent with the active experiences and satisfies utility maximization, the stationary state σ^* is inductively stable.

THEOREM 3: *Consider a stationary state σ^* where every player $i \in N$ has a model coherent with his active experiences $\mathcal{A}(i|\sigma^*)$. Then σ^* is inductively stable if and only if the model of each player satisfies utility maximization (3).*

For the same reason as that for Theorem 2, we can prove that player i can infer the ethnicity configurations of all festivals in the stationary state from his active experiences.

4.3 Rationalization

As remarked in Section 4.1, the cognitive bases for the festival game Γ and an individual model $\mathcal{M}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ are different in that in the former, player i does not know his own utility function $H_i(\cdot)$ as a function, but in the latter, he perceives $\hat{u}_i$ as a function. Knowing his own utility function $\hat{u}_i$ can be regarded as tantamount to maximizing $\hat{u}_i$. Theorem 2 states that the coherency requirement with the active experiences together with Postulate 3 implies utility maximization over the active experiences. Postulate 1 states that player i knows the availability of friendly and unfriendly action in addition to the choice of a festival location. Typically, he faces such a choice problem when an outsider comes to his festival. Then, he reacts always following his reaction function r_i^* and has never experienced the other attitude because only a deviation by a single player is perceivable by Postulate 2b. Since player i knows that he has an alternative choice for his attitude, we require an individual model $\mathcal{M}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ to satisfy utility maximization over these alternative choices.

We say that player i *rationalizes* his strategy in $\mathcal{M}_I$ at σ^* iff for any $x = (g, \delta, y) \in X$ with $[\hat{o}_i(x); \hat{u}_i(x)] \in \mathcal{P}(i|\sigma^*)$ and for any $x' = (g', \delta', y') \in \hat{Z}$ with $g' = g$ and $\delta'_{-i} = \delta_{-i}$,

$$\hat{u}_i(x) \geq \hat{u}_i(x'). \tag{4}$$

This means that the prescribed reaction of player i against a deviation of some player maximizes utility function $\hat{u}_i$ in the model $\mathcal{M}_I$. The determination of the utility value $\hat{u}_i(x') = \hat{u}_i(g, (\delta_{-i}, \delta'_i), y')$ is a speculation in the sense that player i has never experienced $\hat{u}_i(x')$ in the past. Hence, player i can manipulate his model so that it satisfies (4). In our context of the festival game, it suffices to consider only the above case in addition to the choice of a festival location, which Theorem 2 takes care of.

The term *rationalization* is motivated by the fact that player i adjusts his utility function $\hat{u}_i(x')$ so that his behavior satisfies utility maximization.[14] In the next section, we will see that rationalization plays a crucial role in deriving prejudices, and we will show in Section 6.1 that the true-game model $\mathcal{TG}_I$ is hardly rationalizable.

5. Hedonistic Models

This section introduces *hedonistic*[15] *models*, which explain the experiences in terms of the observables for an individual player. The class of hedonistic models is further divided into two subclasses, *naive hedonistic* and *sophisticated hedonistic models*. In a naive hedonistic model, one's utility is determined by his own actions f_i, δ_i and an exogenous variable y. A model of this type can fully explain the active experiences. To explain the passive experiences, however, the model relies heavily on the exogenous variable y; otherwise, it is not rationalizable. On the other hand, a sophisticated hedonistic model allows its utility function to depend on the observed ethnicities. A model of this type is coherent with all the experiences and is rationalizable with a slight use of y. These models exhibit perceptual prejudices, and the latter additionally exhibits preferential prejudices.

We call an individual model $\mathcal{NH}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ a *naive hedonistic model* iff its utility function $\hat{u}_i$ depends only on the player's actions and the exogenous variable y; that is, it is expressed as

NH4: $\hat{u}_i(x) = \hat{u}_i(g_i, \delta_i, y)$ for all $x = (g, \delta, y) \in \hat{Z}$.

This means that player i explains his observed utilities by his choices of a location and a friendly or unfriendly action together with the exogenous variable y. When the utility function $\hat{u}_i$ of an individual model $(\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ depends further on the observed ethnicities, that is, it is expressed as

SH4: $\hat{u}_i(x) = \hat{u}_i(g_i, \delta_i, \hat{E}_i(g), y)$ for all $x = (g, \delta, y) \in \hat{Z}$,

then $(\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ is called a *sophisticated hedonistic model*, which we denote by $\mathcal{SH}_I$.

A sophisticated hedonistic model $\mathcal{SH}_I$ differs from a naive one $\mathcal{NH}_I$ in that in $\mathcal{SH}_I$ the observations are fully used to define the imaginary utility function $\hat{u}_i$, but in $\mathcal{NH}_I$ only his own choices are used. Since an individual

[14] This notion should be distinguished from the rationalizability of Bernheim (1984) and Pearce (1984). It is to capture the notion of rationalization in sociology (cf. Marger 1991, pp. 98–102).

The concept may be well understood by recalling Aesop's *sour grapes*, in which the fox changed its belief so that the grapes must be sour. The problem may be interpreted as a distinction between conceivable and feasible choices. It is, however, the point relevant to our context that the fox adjusts his utility function so as to make an explanation consistent with utility maximization.

[15] The term "hedonistic" is borrowed from J. Bentham's hedonistic calculus.

player has, by Postulate 1, no knowledge about the structure of the society except these observables, it would be "natural" to construct a model based on these observables. This is in contrast with the true-game model $\mathcal{TG}_I$, in which the determination of the utility function needs more speculations on the structure of the society than in the hedonistic models.

When player i restricts his attentions to the active experiences $\mathcal{A}(i|\sigma^*)$, he succeeds in constructing a naive hedonistic model without using any exogenous variable y, which is stated by the following theorem. Its proof is omitted.

THEOREM 4: *Let $\sigma^* = (f^*, r^*)$ be an inductively stable stationary state. Then there is a naive hedonistic model $\mathcal{NH}_I = (\hat{N}, \hat{Z}, \hat{o}_i, \hat{u}_i; x^0, X)$ coherent with the active experiences $\mathcal{A}(i|\sigma^*)$ and such that $\hat{o}_i$ and $\hat{u}_i$ are independent of the exogenous variable y; that is, $\hat{Z} = \{1, \ldots, \ell\}^{\hat{N}} \times \{0,1\}^{\hat{N}}$.*

Thus, he succeeds in explaining his active experiences by ascribing his observed utilities to his actions. This explanation is fallacious from the objective point of view: player i finds an explanation of his observations based on an incorrect causal relationship, but it is still consistent with his observations. In this sense, the naive hedonistic model exhibits *perceptual* prejudices.

When player i also takes the passive experiences into account, a naive hedonistic model typically does not work well. In a segregation equilibrium, if the players in the smallest festival are all nondiscriminators, a naive hedonistic model does work with a slight use of the exogenous variable y. In a larger festival, however, a discriminator cannot rationalize his behavior, and a nondiscriminator needs to use heavily the exogenous variable to explain his passive experiences. The proof of the following theorem is given in the Appendix.

THEOREM 5: *Let $\sigma^* = (f^*, r^*)$ be an inductively stable stationary state satisfying condition FA* $[r_i^*(f^*) = 1$ for any $i \in N]$.

(a) *Failure of Rationalization: Let player i be a discriminator against some ethnicity. If a naive hedonistic model $\mathcal{NH}_I$ is coherent with his experiences $\mathcal{E}(i|\sigma^*)$, then it is not rationalizable at σ^*.*

(b) *Heavy Dependence on y: Let player i be a nondiscriminator toward any ethnicity. If a naive hedonistic model $\mathcal{NH}_I$ is coherent with the experiences $\mathcal{E}(i|\sigma^*)$ and is rationalizable at σ^*, then*

$$\hat{u}_i(g_i^0, 0, y^0) \le \hat{u}_i(g_i^0, \delta_i^0, y) \le \min_{j \in N} H_j(\sigma^*), \tag{5}$$

holds for some y.

In (a), it also follows from coherency with $\mathcal{E}(i|\sigma^*)$ that $\hat{u}_i(g_i^0, 0, y) \le \min_{j \in N} H_j(\sigma^*)$ for some y. The point of (b) is that the utility decrease in

(4) is caused by a change in y, but in (a), it may be regarded as caused by the change in δ_i. If i is in the smallest festival, (5) would not restrict $\hat{u}_i$ because the right-hand side, $\min_{j \in N} H_j(\sigma^*)$, is the stationary utility value of the smallest festival. However, if player i is in a larger festival, his imaginary utility function $\hat{u}_i$ takes a value smaller than or equal to $\min_{j \in N} H_j(\sigma^*)$. Thus, typically, $\hat{u}_i$ heavily depends on the exogenous variable y.

Let us look at why a naive hedonistic model for, say, player i in C (see Figure 5) in festival 3 is not rationalizable or has a heavy dependence on y. Suppose that player j in group A in festival 1 comes to festival 3. Then the mood of festival 3 should be less than or equal to the stationary mood of festival 1, for otherwise j would stay in festival 3. There are two cases to be considered: player i is a discriminator against ethnicity A or he is a nondiscriminator.

If player i is a discriminator against A, he takes unfriendly action in response to the presence of player j, though he regularly takes friendly action. In a naive hedonistic model, however, he cannot justify this switch in his behavior because his imaginary utility function $\hat{u}_i$ does not depend on ethnicities. That is, he fails to rationalize his own discriminatory behavior. On the other hand, if i is a nondiscriminator toward ethnicity A, he does not switch his behavior, but he remains friendly and observes a decrease in his utility when ethnicity A is present in festival 3. In this case, he could explain this change in utility only if he allows $\hat{u}_i$ to depend heavily on y.

When player i is a member in the smallest festival and all the members of it are nondiscriminators against any ethnicities, a naive hedonistic model works well with a slight use of the exogenous variable; that is, $\#Y = 3$ and

$$|\hat{u}_i(g_i^0, \delta_i^0, y^0) - \hat{u}_i(g_i^0, \delta_i^0, y)| = 1$$

for all $y \in Y - \{y^0\}$. The exogenous variable y is used to explain the changes induced when an insider goes out and when an outsider comes in.

From the preceding theorem and arguments, we know that a naive hedonistic model does not work well for a player i in a large festival who takes all the passive experiences into account. However, if player i notices that utility changes caused by passive experiences are almost always associated with the presence of a different ethnicity, then he could find that a sophisticated hedonistic model may be more suitable than a native hedonistic model. Indeed, whichever his festival is and whichever his attitude is, he could succeed in constructing a sophisticated hedonistic model that is rationalizable and has only slight dependence on the exogenous variable y.

THEOREM 6 (Hedonistic sophistication): *Let σ^* be an inductively stable stationary state satisfying condition* FA. *Then any player i has a coherent and* *rationalizable sophisticated hedonistic model* $\mathcal{SH}_I$ *with* $\#Y = 2$, *that is,* $Y = \{y^0, y'\}$, *and*

$$\hat{u}_i(g_i^0, \delta_i^0, E_i(g^0), y^0) - \hat{u}_i(g_i^0, \delta_i^0, E_i(g^0), y') = 1. \tag{6}$$

Perceptual prejudices are involved in a sophisticated hedonistic model $\mathcal{SH}_I$, as in a naive hedonistic model $\mathcal{NH}_I$, in that it involves a fallacious causal relationship. In $\mathcal{SH}_I$, the utility function $\hat{u}_i$ of player i further depends on the ethnicity configuration. In fact, we can regard this dependence as exhibiting that player i develops *preferential prejudices* against the ethnicities of outsiders. The greater the number of his fellow players who react as discriminators, the stronger these prejudices become. To look closely at this fact, let $\mathcal{SH}_I$ be a coherent and rationalizable hedonistic model satisfying (6).

Suppose that players i and j go regularly to different festivals, and that the festival f_j^* of player j is smaller than the festival $k = f_i^*$ of player i. In Figure 5, for example, i is in C and j is in A. Suppose that player j goes to festival k with the friendly action. Then j's induced utility is less than or equal to the original utility level enjoyed by j; that is,

$$H_j(\sigma_{-j}^*, (k,1)) \leq H_j(\sigma^*).$$

In response to the presence of player j, player i takes the nondiscriminatory action, $r_i^*(f_{-j}^*, k) = 1$, or the discriminatory action, $r_i^*(f_{-j}^*, k) = 0$. Consider each case:

Case N: If player i is a nondiscriminator his induced utility satisfies $H_i(\sigma_{-j}^*, (k,1)) = H_j(\sigma_{-j}^*, (k,1))$; that is, his utility decreases significantly, for example, if f_j^* is the smallest festival, $H_j(\sigma_{-j}^*, (k,1))$ is smaller than or equal to $\min_{j'} H_{j'}(\sigma^*)$. In this case, his imaginary utility function $\hat{u}_i$ satisfies $\hat{u}_i(g_i^0, 1, E_i(\sigma^*) \cup \{e(j)\}, y) = H_i(\sigma_{-j}^*, (k,1))$ for some y by CP. By rationalization, $\hat{u}_i(g_i^0, 1, E_i(\sigma^*) \cup \{e(j)\}, y) \geq \hat{u}_i(g_i^0, 0, E_i(\sigma^*) \cup \{e(j)\}, y')$ for any y'. In sum, we have

$$\begin{aligned} \hat{u}_i(g_i^0, 1, E_i(\sigma^*), y^0) = H_i(\sigma^*) &> H_i(\sigma_{-j}^*, (k,1)) \\ &= \hat{u}_i(g_i^0, 1, E_i(\sigma^*) \cup \{e(j)\}, y) \\ &\geq \hat{u}_i(g_i^0, 0, E_i(\sigma^*) \cup \{e(j)\}, y') \quad \text{for all } y'. \end{aligned}$$

Thus, $\hat{u}_i$ decreases with the presence of $e(j)$. Player i still behaves in the friendly manner to ethnicity $e(j)$, but he himself attaches a significant disutility to ethnicity $e(j)$.

Case D: If player i is a discriminator his induced payoff is zero; that is, $H_i(\sigma_{-j}^*, (k,1)) = 0$. In this case, $\hat{u}_i(g_i^0, 0, E_i(\sigma^*) \cup \{e(j)\}, y) = 0$ for some

y by CP, and $0 \geq \hat{u}_i(g_i^0, 1, E_i(\sigma^*) \cup \{e(j)\}, y')$ for all y' by rationalization. Thus, we have

$$\hat{u}_i(g_i^0, 1, E_i(\sigma^*), y^0) > 0 = \hat{u}_i(g_i^0, 0, E_i(\sigma^*) \cup \{e(j)\}, y)$$
$$\geq \hat{u}_i(g_i^0, 1, E_i(\sigma^*) \cup \{e(j)\}, y') \quad \text{for all } y'.$$

Here player i behaves in the unfriendly manner in response to ethnicity $e(j)$ and receives zero utility, but this is still better than taking the friendly action.

In either case, a coherent and rationalizable $\mathcal{SH}_I$ exhibits prejudices against ethnicity $e(j)$. It may be slightly paradoxical to have the possibility that both discriminators and nondiscriminators have preferential prejudices against some ethnicities, and the implications of this are discussed in Section 6.2. Here we note the following difference between the above two cases: If player i is a discriminator, he attaches zero or less payoff to the presence of ethnicity $e(j)$, which may be constant over any outside ethnicities he discriminates against. On the other hand, if player i is a nondiscriminator, then, typically, the smaller festival f_j^* is, the smaller the induced utility $\hat{u}_i(g_i^0, \delta_i, E_i(\sigma^*) \cup \{e(j)\}, y)$ is; for example, if f_j^* is the smallest, the induced utility is smaller than or equal to the smallest stationary utility, $\min_{j'} H_{j'}(\sigma^*)$.

6. Discussion

This paper presents a new theory, called inductive game theory, which targets the formation and emergence of individual views about society from experiences. Instead of general situations, we treat a specific game—the festival game—and specific object phenomena—discrimination and prejudices. This research strategy of considering specific problems was deliberate because the sound development of a theory needs simultaneous considerations of theory and applications. Although we can now generalize the basic principles of inductive game theory to a more general class of games and apply them to different phenomena, such general developments will be given in separate papers. Here we give additional discussions on implications of our theory on the existing game theory, look at the interactions of individual behavior and models, and consider the implications on relevant sociological and economics literatures. Finally, we comment, from the viewpoint of inductive game theory, on the common knowledge assumption on the structure of a game in deductive game theory.

6.1 Subgame Perfection and Rationalization

Proposition 1 characterizes inductive stability to be Nash equilibrium, and Theorems 2 and 3 relate Nash equilibrium to coherency with the active experiences. Game theory literature provides another well-known concept: subgame perfection. As remarked at the end of Section 2, subgame perfection is not derived from experiences, in contrast with Nash equilibrium, because the player has no experiences on the alternative attitude in the second stage of the festival. On the other hand, we have required rationalization for the second choice in an individual model. In the following, we consider the relationship between these concepts through the true-game model $\mathcal{TG}_I$.

In the festival game Γ, subgame perfection on a strategy profile $\sigma^* = (f^*, r^*)$ is relevant only to the nodes in the second stage reachable by a unilateral deviation from σ^*. In the terminology of this paper, the relevant region for subgame perfection is $\bigcup_i \mathcal{P}(i|\sigma^*)$; that is, each of these nodes is reached by a unilateral deviation of a player. We require subgame perfection over this region: $\sigma^* = (f^*, r^*)$ satisfies *subgame perfection over* $\bigcup_i \mathcal{P}(i|\sigma^*)$ iff for all $i, j \in N$ with $f_j^* \neq f_i^* = k$ and for $\delta_j = 0,1$,

$$H_i(\sigma^*_{-j},(k,\delta_j)) \geq H_i(\sigma^*_{-\{i,j\}},(f_i^*,\delta_i),(k,\delta_j)) \quad \text{for } \delta_i = 0,1,$$

where $(\sigma^*_{-\{i,j\}},(f_i^*,\delta_i),(k,\delta_j))$ is obtained from σ^* by replacing σ_i^* and σ_j^* with (f_i^*,δ_i) and (k,δ_j). This means that if an outsider j comes to $k = f_i^*$, the prescribed action $r_i^*(f_{-j}^*, k)$ of player i maximizes his payoff.

Note that we ignore the deviations by an insider. For example, if the festival $k = f_i^*$ has only one player of some ethnicity and if this player goes out from k, player i would observe a change in the ethnicity configuration of k. The above definition does not take this case into account. However, it was proved in Section 3 that we do not have this case in equilibrium with condition FA.

Though there always exists a coherent and rationalizable sophisticated hedonistic model, this may not be the case for the true-game model. Indeed, a striking result is that the true-game model $\mathcal{TG}_I$ typically fails to be rationalizable, as shown in the following theorem, whose proof is given in the Appendix.

THEOREM 7 (Rationalization for the true-game model): *Let σ^* be an inductively stable stationary state satisfying condition* FA. *Then the true-game model $\mathcal{TG}_I$ of player i in σ^* is rationalizable for all $i \in N$ if and only if σ^* satisfies subgame perfection over $\bigcup_i \mathcal{P}(i|\sigma^*)$.*

Theorem 7 states that rationalization corresponds to subgame perfection on σ^* in Γ. As discussed in Section 2, subgame perfection cannot be assumed on σ^* under our postulates, and also, Theorem 1 implies that there are many Nash equilibria that do not satisfy subgame perfection. Therefore, rationalization typically rejects the true-game model.

If we take the position to not necessarily regard subgame perfection as a requirement for the original objective game but to apply it to the individual models, then subgame perfection and rationalization could be regarded as conceptually equivalent. From this point of view, when an individual player thinks about the society as a model, he inclines to assume subgame perfection (rationalization), but unless he has enough experiences on his own payoff function, he may also incline to adjust (rationalize) his payoff function so as to satisfy subgame perfection. This is in sharp contrast with the derivation of Nash equilibrium in the sense of Proposition 1 and Theorem 3: Nash equilibrium for the objective game Γ as well as for the individual model is derived from the active experiences $\mathcal{A}(i|\sigma^*)$.

6.2 Interactions between Models and Behavior

It may be already clear from our discussions that coherency and rationalization express different reasoning processes. Coherency for a model is attained through the process of induction based on experiences; rationalization is obtained through the introspection of consistency between the model and behavior. When the model fails to be rationalizable, the player either alters the model (interpretation) or changes his behavior. Here, models and behavior interact with each other. This subsection discusses their evolution in a dynamic context, considering some possible scenarios.

As the reference point of induction, we should recall the mere-enumeration model $\mathcal{ME}_I$ of Remark 1, though its formulation needs a slight generalization of the definitions of a model. The mere-enumeration model does not go much beyond the state of collecting the experiences $\mathcal{E}(i|\sigma^*)$ since it gives no causal relationship between the player's observations and satisfaction. In this sense, it represents the state of mind of the player who has had experiences, is conscious of them, but has not further deliberated on them. In contrast, the true-game model could be regarded as the ultimate goal from the objective point of view. The problem associated with it is whether an individual player needs to or is able to consider the true-game model after the full deliberation of experiences.

Suppose that player i has experiences $\mathcal{E}(i|\sigma^*)$ and is conscious of them. If he wants to have a better explanation of his observations including utility values, he may start thinking about hedonistic models.

First, consider naive hedonistic models. It follows from Theorem 4 that if player i cares only about active experiences or if he is in the smallest festival where every player is a nondiscriminator, a naive hedonistic model could work and he need not think about the society any further. Let player i be in a larger festival, and suppose that he takes passive experiences as well as active ones into account and introspects his explanation of experiences. Theorem 5a states that if he is a discriminator against some ethnicities, he cannot rationalize his behavior in a naive hedonistic model. Hence a discriminator would sophisticate his explanation, and may reach a sophisticated hedonistic model. Theorem 5b states that if he is a nondiscriminator toward any ethnicities, he can succeed in explaining his behavior in a naive hedonistic model by using an exogenous variable. However, if he wants a better explanation to avoid a heavy use of an exogenous variable, he may sophisticate his model. In either case, a natural candidate for a modification would be a sophisticated hedonistic model.

When the players reach sophisticated hedonistic models that are coherent with experiences and rationalizable, no further change will be induced. Then the inductively stationary state is truly stable. In this process, models have deviated from naive hedonistic models to become models whose utility functions involve ethnicities as a fallacious explanatory variable. This is the emergence of preferential prejudices against ethnicities.

Yet, there is another logical possibility: that the discriminators change their actions to the friendly ones at the same time, though it could be rather accidental and could hardly ever occur. However, if this happens, their utility values would not decrease even when an outsider from a smaller festival visits their festival. Since such an outsider receives a higher utility value, he would stay in the larger festival. The dissolution of segregation would then follow. But this possibility would be very accidental.

Finally, let us look at what happens if player i starts seeking the true-game model $\mathcal{TG}_I$ in the process of deviating from $\mathcal{NH}_I$ or finding a better explanation of his experiences than $\mathcal{SH}_I$. Suppose that he thinks about $\mathcal{TG}_I$. He must be uncertain about the correctness of his true-game model because he has no evidence other than his experiences in our context (therefore, it must be very difficult for him to think about $\mathcal{TG}_I$) . From Theorem 1, a segregated equilibrium typically does not satisfy subgame perfection, and Theorem 7 then implies that the true-game models of players in the larger festival are not rationalizable. Since the player is uncertain about his model, he modifies it so as to rationalize his behavior. One possibility is to go to or return to a sophisticated hedonistic model. Thus, a sophisticated hedonistic model is also regarded as stable in this sense.

When players have reached coherent and rationalizable sophisticated hedonistic models, they cannot reject such prejudicial models unless the players have new experiences, for example, by going to another society with a different stationary state. In a separate paper, we will consider effects of such new experiences on individual models.

6.3 Comparisons with Merton's Classification

Finally, we mention some implications to the studies of discrimination and prejudices in sociology and economics.

Merton (1949) suggested four ideal types by combining the prejudicial attitudes with the propensity either to engage in discriminatory actions or to refrain from them.

	Unprejudiced	Prejudiced
Nondiscriminators	All-weather liberals	Timid bigots
Discriminators	Fair-weather liberals	Active bigots

These types have counterparts in our theory. First, we interpret "all-weather liberals" as the nondiscriminators whose utility functions $\hat{u}_i$ in their models are independent of ethnicity configurations such as naive hedonistic models. Second, we interpret "active bigots" as the discriminators whose utility functions in their models depend on ethnicities such as player i with a sophisticated hedonistic model $\mathcal{SH}_I$ in Case D of Section 5. Third, "timid bigots" are the nondiscriminators, but their utility functions in their models are based on negative images of other ethnic groups, such as player i in Case N of Section 5. Fourth, "fair-weather liberals" are the discriminators but explain their utilities without referring to ethnicities. In our context, those are interpreted as players either with the mere-enumeration models or with the true-game models.

Merton (1949) introduced those four types to examine the causal relationship between prejudices and discrimination. If prejudices induce discrimination, then people could simply be categorized into either all-weather liberals or active bigots. However, Merton argued (also see Marger (1991), Chap. 3, for recent assessments of this view) that because all four types of people are observed in our society, the causal relationship from prejudices to discriminatory behavior is questionable. In our theory, as discussed above, prejudices may emerge in evolutions of the behavior of players together with their views on the society, and those four types of people seem to appear. In this sense, our theory supports Merton's view, though our theory goes beyond his.

In neoclassical economics (e.g., Becker 1957), it has been assumed that behavioral attitudes are determined by mental attitudes. Thus, the view described in this paper looks contradictory to neoclassical economics. It should be noticed, however, that our theory is about a long-run situation, and that if we take a snapshot of this long-run situation, the causal relation would become one-directional; that is, prejudices induce discriminatory behavior.

6.4 Large Societal Games versus Small Micro Games

If the source of the individual knowledge of the structure of a game is the inductive construction of a view about the game from experiences, then deductive game theory could be regarded as part (a result) of inductive game theory. In deductive game theory, it is a traditional view that the structure of the game is common knowledge. On the other hand, we have considered an individual model that has only an imaginary utility function of the player. This assumption can be easily extended to a social model including the utility functions of other imaginary players. Nevertheless, it would be a different problem whether these utility functions are assumed to be common knowledge. It is important to notice that common knowledge requires each player to look into the other players' minds. In our context, since the individual player does not know even the identities of the other players, this requirement could be hardly met. Deductive game theory with the common knowledge assumption may appear to have no room in inductive game theory.

We take, however, a view to regard deductive game theory with the common knowledge assumption as an ideal (limit) case in a direction, different from that taken in this paper, in inductive game theory. The basic criterion is whether or not an individual player looks into the other players' minds and thinks about the other players' thinking. This criterion may be regarded as being met in a small micro game where the players are playing the game face-to-face. This is another direction of inductive game theory, and Kaneko (1998) considers that the common knowledge assumption is regarded as a limit case of the evolution of the knowledge structure obtained by induction and deduction in such a small micro game.

In a large society, the assumption of looking into other people's minds is simply inadequate because even knowing the identities of other people is difficult. After all, what we have discussed is a problem of a large societal game. An individual model is an external description of the society, and should be interpreted as a partial description of the society, though our definition may still be too detailed.

Appendix

Proof of Theorem 1: For the *only-if* part: Let $\sigma^* = (r^*, f^*)$ be a Nash equilibrium. Suppose $\mu_i(\sigma^*) \geq m_0$. It is better for each player in festival f_i^* to behave in a friendly manner; hence $r_j^*(f^*) = 1$ for any j with $f_j^* = f_i^*$, which is the second conclusion of assertion (a). To prove the first conclusion, we suppose, on the contrary, that some player j of ethnicity e chooses $f_j^* \neq f_i^*$. Note that neither a move of i to f_j^* nor that of j to f_i^* affects the ethnicity configuration of f_j^* or of f_i^*. This means that neither move induces a new response. There are the two cases $\mu_i(\sigma^*) \geq \mu_j(\sigma^*)$ and $\mu_i(\sigma^*) < \mu_j(\sigma^*)$ to be considered. In the former case, player j would be better off by coming to f_i^* and taking a friendly action than being at f_j^* by the definition of $H_i(\cdot)$, since the mood at f_i^* relevant to j is $\mu_i(\sigma^*) + 1$. In the latter case, player i would be better off by going to festival f_j^*. In either case, we have a contradiction. Thus we have the first conclusion of assertion (a). Assertion (b) follows the definition of Nash equilibrium.

Suppose $\mu_i(\sigma^*) < m_0$. Then it is better for any player in festival f_i^* to take an unfriendly action by the definition of $H_i(\cdot)$. Thus we have assertion (c). When player i moves to another festival f_i, then his payoff must be smaller than or equal to 0 at festival f_i since σ^* is a Nash equilibrium. Hence the induced mood should not exceed the critical level m_0. Thus we have assertion (d).

For the *if* part: Consider a strategy configuration $\sigma^* = (f^*, r^*)$ that satisfies (a)–(d). First, suppose $\mu_i(\sigma^*) \geq m_0$. Then he does not have an incentive to change his attitude to 0 at festival f_i^* since from (a) he now obtains a positive payoff. Also it follows from (b) that he does not have an incentive to move to any other festival with $\delta_i = 1$. Second, suppose $\mu_i(\sigma^*) < m_0$. Then it follows from (c) and (d) that there is no incentive for player i to change his attitude at f_i^* as well as to move to any other festival. ■

Before we prove Theorem 2, we present the following lemma.

LEMMA 1: *Suppose that* $\mathcal{M}_I$ *is coherent with* $\mathcal{A}(i|\sigma^*)$. *Let* $x = (g,\delta,y)$ *and* $x' = (g',\delta',y')$ *in X satisfy* $g = g' = (g^0_{-i}, g_i)$, $(g_i,\delta_i) \neq (g_i^0,\delta_i^0)$, *and* $\delta_i = \delta_i'$. *Then* $[\hat{o}_i(x);\hat{u}_i(x)] = [\hat{o}_i(x');\hat{u}_i(x')]$.

Proof: By the *if* part of CA, both $[\hat{o}_i(x);\hat{u}_i(x)]$ and $[\hat{o}_i(x');\hat{u}_i(x')]$ belong to $\mathcal{A}(i|\sigma^*)$. If two experiences $[\phi_i;h_i]$ and $[\phi_i';h_i']$ in $\mathcal{A}(i|\sigma^*)$ are induced by the same deviation (g_i,δ_i) from $\sigma^* = (f^*,r^*)$, they coincide, since they are expressed as $[g_i,\delta_i,E_i(f^*_{-i},g_i);\ H_i(\sigma^*_{-i},(f_i,\delta_i))]$. Since the deviation part of player i in $x = (g,\delta,y)$ and $x' = (g',\delta',y')$ are the same, so are $[\hat{o}_i(x);\hat{u}_i(x)]$ and $[\hat{o}_i(x');\hat{u}_i(x')]$. ■

Proof of Theorem 2: First, $\hat{u}_i(x^0) = H_i(\sigma^*)$ by CS. Consider $(\sigma^*_{-i},(g_i,\delta_i))$. This gives an active experience $\phi_i = (g_i,\delta_i,E_i(\sigma^*_{-i},(g_i,\delta_i)))$ and $h_i = H_i(\sigma^*_{-i},(g_i,\delta_i))$. By CA, there is a state $x' = ((g^0_{-i},g_i),(\delta'_{-i},\delta_i),y') \in X$ such that $\hat{o}_i(x') = \phi_i$ and $u_i(x') = h_i$. Now we take an arbitrary $x = ((g^0_{-i},g_i),(\delta_{-i},\delta_i),y) \in X$. Lemma 1 implies $h_i = \hat{u}_i(x') = \hat{u}_i(x)$. Since σ^* is inductively stable, we have $H_i(\sigma^*) \geq h_i$ by Proposition 1. Hence $\hat{u}_i(x^0) = H_i(\sigma^*) \geq h_i = \hat{u}_i(x') = \hat{u}_i(x)$. ■

Proof of Theorem 3: The only-if part is Theorem 2. Thus it suffices to show that if the model $\mathcal{M}_I = (\hat{N},\hat{Z},\hat{o}_i,\hat{u}_i;x^0,X)$ of player i is coherent with $\mathcal{A}(i|\sigma^*)$ and satisfies inequality (3), then his stationary actions maximize his objective payoff function. Consider an arbitrary $(f_i,\delta_i) \in \{1,\ldots,\ell\} \times \{0,1\}$. This induces an experience $[\phi_i;h_i] = [f_i,\delta_i,E_i(f^*_{-i},f_i);H_i(\sigma^*_{-i},(f_i,\delta_i))] \in \mathcal{A}(i|\sigma^*)$. Since $\mathcal{M}_I$ is coherent with active experiences $\mathcal{A}(i|\sigma^*)$, there is a state $x = (g,\delta,y) \in X$ such that $g_{-i} = g^0_{-i}$, $\hat{o}_i(x) = (f_i,\delta_i,E_i(f^*_{-i},f_i))$ and $\hat{u}_i(x) = H_i(\sigma^*_{-i},(f_i,\delta_i))$. Also, $\hat{u}_i(x^0) = H_i(\sigma^*)$ by CS. Then $H_i(\sigma^*) = \hat{u}_i(x^0) \geq \hat{u}_i(x) = H_i(\sigma^*_{-i},(f_i,\delta_i))$. ■

Proof of Theorem 5: For Part a, let j come to the festival k of player i with the friendly action. Suppose that player i takes the unfriendly action to j's presence. Then $H_i(\sigma^*_{-j},(k,1)) = 0 < H_i(\sigma^*)$. By CP, $\hat{u}_i(g,\delta,y) = \hat{u}_i(g_i^0,0,y) = 0$ for some (g,δ,y). However, since $\hat{u}_i(g_i^0,1,y^0) =$

$H_i(\sigma^*) > 0$ by CS, it holds that $\hat{u}_i(g,\delta,y) = \hat{u}_i(g_i^0,0,y) = 0 < \hat{u}_i(g_i^0,1,y^0) = \hat{u}_i(g,(\delta_{-i},1),y^0)$. This violates rationalization.

For Part b, let j be a player in the smallest festival. Then his payoff $H_j(\sigma^*)$ is the lowest. When j comes taking the friendly action to the festival k of player i, it holds that $H_j(\sigma^*_{-j},(k,1)) \leq H_j(\sigma^*)$, since otherwise j would stay in k. By CP, there is (g,δ,y) such that $g_i = g_i^0$, $\delta_i = r_i^*(f^*_{-i},k) = 1 = \delta_i^0$ and $\hat{u}_i(g_i^0,1,y) = H_i(\sigma^*_{-j},(k,1)) = H_j(\sigma^*_{-j},(k,1))$. Thus $\hat{u}_i(g_i^0,\delta_i^0,y) \leq H_j(\sigma^*) = \min_{j'} H_{j'}(\sigma^*)$. By rationalization, $\hat{u}_i(g_i^0,0,y^0) \leq \hat{u}_i(g_i^0,\delta_i^0,y) \leq H_j(\sigma^*) = \min_{j'} H_{j'}(\sigma^*)$. ■

Proof of Theorem 6: Let $\sigma^* = (f^*,r^*)$ be an inductively stable stationary state satisfying FA. Let $k = f_i^*$. We define the constituents of a sophisticated hedonistic model other than utility function $\hat{u}_i$ as follows:

SH1: $\hat{N}$ is an arbitrary imaginary player set partitioned into nonempty disjoint ethnic groups $\hat{N}_1,\ldots,\hat{N}_{e_0}$ with $i \in \hat{N}_{e(i)}$ and $\#\hat{N}_{e(i)} \geq 2$;

SH2: $\hat{Z} = \{1,\ldots,\ell\}^{\hat{N}} \times \{0,1\}^{\hat{N}} \times \{-1,0\}$;

SH3: $\hat{o}_i(x) = (g_i,\delta_i,\hat{E}_i(g))$ for all $x = (g,\delta,y) \in \hat{Z}$;

SH5: $x^0 = (g^0,\delta^0,0)$, where $\delta^0 = \mathbf{1}^{\hat{N}}$ and $g^0 = (g_j^0)_{j\in\hat{N}}$ is defined by: for each $j \in \hat{N}_e$ $(e = 1,\ldots,e_0)$, $g_j^0 = f_{j'}^*$ for some $j' \in N_e$;

SH6: $X = \{x^0\} \cup X_A \cup X_{PO} \cup X_{PI}$, where

$$X_A = \{((g^0_{-i},g_i),(\delta^0_{-i},\delta_i),0) : (g_i,\delta_i) \in \{1,\ldots,\ell\} \times \{0,1\}$$

$$\text{and } (g_i,\delta_i) \neq (g_i^0,1)\};$$

$$X_{PO} = \{((g^0_{-j},k),(\delta^0_{-i},r_i^*(k,\hat{E}_i(g^0_{-j},k))),0) : j \in \hat{N} \quad \text{and} \quad g_j^0 \neq k\};$$

$$X_{PI} = \{(g^0,\delta^0,-1)\}.$$

We define a utility function $\hat{u}_i : \{1,\ldots,\ell\} \times \{0,1\} \times 2^{\{1,\ldots,e_0\}} \times \{-1,0\} \rightarrow \mathsf{R}$ by

$$\hat{u}_i(g_i,\delta_i,\mathrm{E},y) = \begin{cases} \delta_i \cdot (\mu_i(\sigma^*_{-i},(g_i,\delta_i)) + y) \\ \quad \text{if } \mathrm{E} = E_i(f^*_{-i},g_i) \text{ for some } g_i \in \{1,\ldots,\ell\} \\ \delta_i \cdot (\mu_i(\sigma^*_{-j},(k,\delta_i)) + y) \\ \quad \text{if } \mathrm{E} = E_i(f^*_{-j},k) \text{ for some } j \text{ with } f_j^* \neq k \text{ and } \delta_i = r_i^*(f^*_{-j},k) \\ h^-(g_i,\delta_i,\mathrm{E},y) \\ \quad \text{if } \mathrm{E} = E_i(f^*_{-j},k) \text{ for some } j \text{ with } f_j^* \neq k \text{ and } \delta_i \neq r_i^*(f^*_{-j},k), \\ \text{arbitrary} \\ \quad \text{otherwise,} \end{cases}$$

where $h^-(g_i,\delta_i,\mathrm{E},y)$ is a real number not greater than $H_i(\sigma^*_{-j},(k,0))$ in the third case. The well-definedness of each case is guaranteed by FA and Theorem 1. The first case gives utility values to i's own deviations, which together with $\hat{o}_i$ implies coherency with active experiences. This covers also the case where an insider takes an unfriendly action or moves out of festival k. The second case gives utility values to the cases where outsiders come to festival k, which together with $\hat{o}_i$ implies coherency with passive experiences. The third case gives utility values to the unexperienced cases where an outsider come to festival k but he took the action not prescribed by r_i^*. In fact, we set $h^-(g_i,\delta_i,\mathrm{E},y)$ so that the sophisticated hedonistic model is rationalizable. ■

Proof of Theorem 7: The *if* part is straightforward. We show the *only-if* part. By FA, σ^* is a fully active equilibrium. If all players in each festival are all nondiscriminators or discriminators toward each ethnicity of an outsider, then σ^* enjoys subgame perfection over $\bigcup_i \mathcal{P}(i|\sigma^*)$. Now we show the contrapositive of the *only-if* part. Suppose that σ^* does not enjoy subgame perfection over $\bigcup_i \mathcal{P}(i|\sigma^*)$ Then some festival has nondiscriminators as well as discriminators toward some ethnicity.

Let k be a festival where some are discriminators and some are nondiscriminators when an outsider j comes to k. Let i and i' be a discriminator and a nondiscriminator, respectively, in k toward the ethnicity of j. There are two cases to consider: (a) $\mu_i(\sigma^*_{-j},(k,1)) \geq m_0$ and (b) $\mu_i(\sigma^*_{-j},(k,1)) < m_0$.

In case (a), let x be the path determined by $(\sigma^*_{-j},(k,1))$. Then $\hat{u}_i(x) = H_i(\sigma^*_{-j},(k,1)) = 0$. Let $\sigma'_i = \sigma'_j = (k,1)$, $\sigma'_{-i,j} = \sigma^*_{-i,j}$, and x' be the path determined by σ'. Then $0 < H_i(\sigma') = \hat{u}_i(x')$. Hence $\mathcal{TG}_I$ of player i is not rationalizable.

Consider case (b) in which $\mu_{i'}(\sigma^*_{-j},(k,1)) < \mu_i(\sigma^*_{-j},(k,1)) < m_0$. Define σ', the strategy profile by $\sigma'_{i'} = (k,0), \sigma'_j = (k,1)$ and $\sigma'_{-i',j} = \sigma^*_{-i',j}$, and let x' be the path determined by σ'. Then $\hat{u}_{i'}(x) = H_{i'}(\sigma^*_{-j},(k,1)) < 0 = H_{i'}(\sigma') = \hat{u}_{i'}(x')$. Thus $\mathcal{TG}_I$ of player i' is not rationalizable. ■

References

ARROW, K. J. (1972) Some models of racial discrimination in labor markets; in *Racial Discrimination in Economic Life,* 187–203, A. H. PASCAL, ed. Lexington: Lexington Books.

BECKER, G. S. (1957) *The Economics of Discrimination.* Chicago: University of Chicago Press.

BERNHEIM, B. D. (1984) Rationalizable strategic behavior, *Econometrica* **52**, 1007–1028.

BINMORE, K. (1987) Modeling rational players I, *Economics and Philosophy* **3**, 179–214.

FUDENBERG, D., and D. LEVINE (1993) Self-confirming equilibrium, *Econometrica* **61**, 523–545.

GILBOA, I., and D. SCHMEIDLER (1995) Case-based decision theory, *Quarterly Journal of Economics* **110**, 605–639.

HARSANYI, J. (1967–68) Games with incomplete information played by 'Bayesian' players, *Management Sciences* **14**, 159–182, 320–334, 486–502.

KALAI, E., and E. LEHRER (1993) Rational learning leads to Nash equilibrium, *Econometrica* **61**, 1019–1045.

KALAI, E., and E. LEHRER (1995) Subjective games and equilibria, *Games and Economic Behavior* **8**, 123–163.

KANEKO, M. (1987) The conventionally stable sets in noncooperative games with limited observations: Definitions and introductory arguments, *Mathematical Social Sciences* **13**, 93–128.

KANEKO, M. (1998) Evolution of thoughts: Deductive game theories in the inductive game situation, IPPS. DP. 781 and 782, University of Tsukuba.

KANEKO, M., and T. KIMURA (1992) Conventions, social prejudices and discrimination: A festival game with merrymakers, *Games and Economic Behavior* **4**, 511–527.

KANEKO, M., and A. MATSUI (1997) Inductive game theory: Discrimination and prejudices, IPPS. DP. 711, University of Tsukuba, and DP. 97-F-39, University of Tokyo (the original of the present paper).

KANEKO, M., and A. MATSUI (1999) Inductive Game Theory: Individual Perceptions of Game Structures. Forthcoming.

KANEKO, M., and T. NAGASHIMA (1996) Game logic and its applications I, *Studia Logica* **57**, 325–354.

KANEKO, M., and T. NAGASHIMA (1997) Game logic and its applications II, *Studia Logica* **58**, 273–303.

KANEKO, M., and S. RAYCHOUDHURI (1993) Segregation, discriminatory behavior, and fallacious utility functions in the festival game with merrymakers, ISEP–DP. No. 535.

MARGER, M. N. (1991) *Race & Ethnic Relations*, 2nd ed. Belmont, CA: Wadsworth Publishing.

MATSUI, A. (1997) Expected utility and case-based reasoning, mimeo.

MENDELSON, E. (1987) *Introduction to Mathematical Logic*. Monterey, CA: Wadsworth & Brooks.

MERTON, R. K. (1949) Discrimination and the American Creed; in *Discrimination and National Welfare*, H. MACLVER, ed. New York: Harper & Row.

NASH, J. F. (1951) Noncooperative games, *Annals of Mathematics* **54**, 286–295.

PEARCE, D. G. (1984) Rationalizable strategic behavior and the problem of perfection, *Econometrica* **52**, 1029–1050.

PLATO (1941) *The Republic of PLATO*, Translated by F. M. Cornford. London: Oxford University Press.

SELTEN, R. (1975) Reexamination of the perfectness concept for equilibrium points in extensive games, *International Journal of Game Theory* **4**, 25–55.

SELTEN, R. (1991) Evolution, learning, and economic behavior, *Games and Economic Behavior* **3**, 3–24.

VAN DAMME, E. (1987) *Stability and perfection of NASH equilibria.* Berlin: Springer Verlag.